

# Curtin PhD Thesis Cover Page

Thin K. Nguyen

January 2012

Department of Computing

A Sentiment Based Approach to Pattern Discovery  
and Classification in Social Media

Thin K. Nguyen

This thesis is presented for the Degree of  
Doctor of Philosophy  
of

Curtin University

January 2012

## Declaration

To the best of my knowledge and belief, this thesis contains no material previously published by any other person, except where due acknowledgement has been made.

This thesis contains no material that has been accepted for the award of any other degree or diploma in any university.

Signature: .....

Date: .....

# Contents

<b>Abstract</b>	<b>xix</b>
<b>Acknowledgements</b>	<b>xxi</b>
<b>Relevant Publications</b>	<b>xxiii</b>
<b>Notation</b>	<b>xxv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Aims and Approach . . . . .	2
1.2 Contribution and Significance . . . . .	4
1.3 Outline of the Thesis . . . . .	6
<b>2 Background</b>	<b>8</b>
2.1 Social Media: An Overview . . . . .	8
2.2 Weblogs . . . . .	13
2.3 Sentiment Analysis in Social Media . . . . .	15

2.3.1	Subjectivity detection . . . . .	16
2.3.2	Sentiment classification . . . . .	16
2.3.3	Mood analysis . . . . .	17
2.3.4	Joint topic-sentiment analysis . . . . .	20
2.3.5	Opinion summarisation . . . . .	20
2.4	Techniques for Mining of Social Media . . . . .	21
2.4.1	Classification . . . . .	21
2.4.1.1	Naive Bayes classifier . . . . .	22
2.4.1.2	Support vector machine classifier . . . . .	23
2.4.1.3	Logistic classification . . . . .	25
2.4.1.4	Classification evaluation . . . . .	26
2.4.2	Clustering with affinity propagation . . . . .	27
2.4.2.1	Similarity measures . . . . .	28
2.4.2.2	Evaluation of clustering . . . . .	29
2.5	Features for Sentiment Analysis . . . . .	30
2.5.1	Parts of speech . . . . .	30
2.5.2	Sentiment bearing lexicon features . . . . .	31
2.5.3	Other sentiment related features . . . . .	34
2.5.4	Psycholinguistic features . . . . .	34
2.5.5	Topics and topic modelling . . . . .	36

2.5.6	Feature selection methods . . . . .	37
2.6	Sentiment-based Detection of Events and Bursts . . . . .	38
2.6.1	Event detection . . . . .	38
2.6.2	Burst detection . . . . .	39
2.7	Identity in Social Media . . . . .	41
2.8	Networking Aspect of Social Media . . . . .	42
2.9	Conclusion . . . . .	43
<b>3</b>	<b>Feature Selection and Mood Classification</b>	<b>45</b>
3.1	Mood Classification . . . . .	45
3.2	Feature Selection for Mood Classification . . . . .	48
3.2.1	Term-based selection . . . . .	48
3.2.2	Term-class interaction-based selection . . . . .	49
3.3	Datasets . . . . .	51
3.3.1	The IR05 dataset . . . . .	51
3.3.2	The CHI06 dataset . . . . .	53
3.3.3	The WSM09 dataset . . . . .	54
3.4	Classification Results . . . . .	54
3.4.1	Effect of feature selection schemes and linguistic components.	55
3.4.2	Performance of LIWC . . . . .	56

3.4.3	Performance of ANEW . . . . .	59
3.4.4	Binary versus counting features . . . . .	61
3.5	Conclusion . . . . .	61
<b>4</b>	<b>Mood Patterns and Affective Lexicon Access in Weblogs</b>	<b>63</b>
4.1	Data-driven Discovery of Mood Patterns . . . . .	64
4.1.1	Basic emotions . . . . .	64
4.1.2	Proposed framework . . . . .	65
4.1.3	Self-organised mood patterns . . . . .	68
4.1.4	Affinity propagation mood patterns . . . . .	68
4.1.5	Mood distance . . . . .	72
4.2	Moods and Affective Lexicon Access . . . . .	74
4.2.1	Mood and ANEW usage association . . . . .	74
4.2.2	Results . . . . .	75
4.3	Conclusion . . . . .	77
<b>5</b>	<b>Event Extraction Using Behaviours of Sentiment Signals and Burst Structure in Social Media</b>	<b>78</b>
5.1	Sentiment Reactions in Social Media . . . . .	79
5.2	Dataset and Evaluation . . . . .	81
5.3	Sentiment Index and Event Extraction . . . . .	85

5.3.1	Mood-based sentiment index extraction . . . . .	86
5.3.1.1	Daily and monthly valence . . . . .	88
5.3.1.2	Abnormal valence . . . . .	89
5.3.2	Content-based sentiment index extraction . . . . .	91
5.3.3	Happy events versus sad events . . . . .	93
5.4	Mood-based Burst and Event Extraction . . . . .	94
5.4.1	Burst-based event extraction . . . . .	94
5.4.1.1	Incremental burst detection . . . . .	97
5.4.1.2	Bursty event detection . . . . .	98
5.4.2	Experimental results . . . . .	99
5.4.2.1	Inferring impact of mood labels for event extraction .	99
5.4.2.2	Top events and burst detection algorithm comparison . . . . .	101
5.4.2.3	Extreme moods and burst detection algorithm comparison . . . . .	105
5.5	Conclusion . . . . .	106
<b>6</b>	<b>Egocentric Aspect in Social Media</b>	<b>107</b>
6.1	Digital Presence in Social Media . . . . .	108
6.2	Hypotheses . . . . .	110
6.3	Experimental Setup . . . . .	110



6.3.1	Dataset . . . . .	110
6.3.1.1	Old and young blogger sub-corpus . . . . .	112
6.3.1.2	Social connectivity corpora . . . . .	112
6.3.2	Feature sets and classifiers . . . . .	114
6.4	Prediction Results . . . . .	115
6.5	Demographic and Personality Impact . . . . .	116
6.5.1	Difference of topic across age and social connectivity . . . . .	118
6.5.2	Difference of linguistic style across age and social connectivity . . . . .	120
6.5.2.1	Old versus young bloggers . . . . .	120
6.5.2.2	Social versus solo bloggers . . . . .	120
6.5.3	Difference of mood across age and social connectivity . . . . .	123
6.6	Conclusion and Future Work . . . . .	125
<b>7</b>	<b>Networking Aspect of Social Media</b>	<b>126</b>
7.1	Networking in Social Media: Communities . . . . .	127
7.2	Hyper-community Detection Framework . . . . .	128
7.2.1	Community representation . . . . .	131
7.2.1.1	Topic-based representation . . . . .	132
7.2.1.2	Sentiment-based representation . . . . .	132
7.2.1.3	Psycholinguistic-based representation . . . . .	135

7.2.2	Community clustering . . . . .	135
7.3	Hyper-community Detection Results . . . . .	137
7.3.1	Topic-based hyper-groups . . . . .	137
7.3.2	LIWC-based hyper-groups . . . . .	143
7.3.3	Mood-based hyper-groups . . . . .	144
7.3.4	ANEW-based hyper-groups . . . . .	145
7.3.5	Discussion . . . . .	145
7.4	Community Prediction . . . . .	147
7.4.1	Features . . . . .	148
7.4.2	Community prediction . . . . .	150
7.5	Conclusion . . . . .	151
<b>8</b>	<b>Conclusion</b>	<b>153</b>
8.1	Summary . . . . .	153
8.2	Future Directions . . . . .	155
<b>A</b>	<b>Appendix</b>	<b>157</b>
A.1	Mood Hierarchical Tree . . . . .	157
A.2	Sentiment Burst Detection and Retrieval System . . . . .	157
	<b>Bibliography</b>	<b>161</b>

# List of Figures

2.1	Examples of moods suggested by Livejournal. . . . .	18
2.2	Core affect model expressing emotion structure on a 2D plane of valence and arousal (from Russell [2003]). . . . .	19
2.3	Valence and arousal values of 1,034 ANEW words on the affect circle.	31
2.4	Cloud visualisation of ANEW words used in the content of the CHI06 database. . . . .	32
2.5	Examples of ANEW words (with valence in the parentheses) in favour of happy (above) and sad (below) blog posts. The magnitude shows the difference between the percentage of corresponding ANEW words in the content of happy and sad blog posts. . . . .	33
2.6	Latent Dirichlet Allocation [Blei et al., 2003]. . . . .	36
3.1	An example of blog posts in Livejournal with the ‘current mood’ tag ( <i>hopeful</i> in this case). . . . .	52
3.2	Cloud visualisation of 132 predefined mood labels tagged in the CHI06 dataset. . . . .	52
3.3	Performances of binary versus counting features for mood classification in the top best results. The performance is measured in F-score and sorted in increasing order of performance. . . . .	58

3.4	The LIWC features above the zero line are in favour of <i>happy</i> ; otherwise, they are in favour of <i>sad</i> ( $ps < 0.001$ ). . . . .	59
3.5	Examples of the use of ANEW words in happy and sad blog posts. The colour reflects the valence and arousal values an ANEW word conveys, as shown in Figure 2.3. . . . .	60
4.1	The ANEW word usage in the blog posts tagged with different moods.	66
4.2	The ANEW word proportion in two groups of moods: $\{angry, p^*ssed\}$ and $\{happy, cheerful\}$ . The valence of ANEW words is increasing on the horizontal axis. . . . .	67
4.3	Discovered mood structure map. Each cluster is annotated with the top six mood categories. . . . .	69
4.4	Discovered emotion patterns on the affect circle. . . . .	70
4.5	Projection of moods onto a two-dimensional mesh using multidimensional scaling. . . . .	72
4.6	Moods in a dendrogram using a hierarchical clustering. . . . .	73
5.1	Daily number of blog posts over the corpus period. . . . .	82
5.2	Examples of querying Google for top five words for two topics for the date of 11 September 2001. By general search, we cannot determine which topics mention an event since both return millions of results. By restraining the search results to September 2001, ‘school’ receives five results, whereas ‘WTC’ returns 17,500 results. Within the time-frame of 10 September–12 September 2001, the search engine returns zero for ‘school’ and 6,830 results for ‘WTC’. . . . .	83
5.3	Sentiment indices computed using the valence value of the tagged moods. . . . .	85

5.4	Weekly and monthly patterns of mood-based SI. . . . .	88
5.5	Weekly and monthly number of posts tagged with the moods. . . . .	88
5.6	Sentiment indices using negative and positive words in blog posts and the corresponding events. . . . .	90
5.7	Happy versus sad events: the moods above the zero line are used more often for happy events and those below the zero line are for traumatic or sad events ( $ps<0.005$ ). . . . .	92
5.8	Happy versus sad events: visualisation of corresponding mood labels in the core effect model [Russell, 1980, 2009]. Those in blue are used more often for happy events, whereas moods in white are used more often for sobering or sad events. The size of the label reflects the extent of the difference ( $ps<0.005$ ). . . . .	92
5.9	The set of bursty moods (that is, moods with at least one burst detected) and significant moods that are found to be suitable for event extraction (cloud visualisation size is proportional to the number of bursts detected for the respective mood). . . . .	96
5.10	Bursty moods (sorted by their valence increasingly from top to bottom) and their bursty tagged intervals in 2004. . . . .	100
5.11	Highest points (from the normalised signal of the count of blogs tagged with mood <i>sad</i> ) and burst detection results for mood <i>sad</i> from three burst detection algorithms. For THD, we manually adjust the threshold so that 10 bursts are detected. . . . .	104
6.1	An example of a user's profile related to his networking. This user has 79 friends, 2,140 followers and belongs to four communities. . . . .	111
6.2	Top 10 countries Livejournal bloggers live in. . . . .	111
6.3	Age distribution of bloggers. . . . .	112

6.4	Distributions of the number of communities a blogger joins and the number of followers and friends a blogger has. . . . .	113
6.5	Prediction performance (F-measure). . . . .	117
6.6	The LIWC features above the zero line are favoured by the <i>old</i> ; otherwise, they are favoured by the <i>young</i> ( $ps < 0.001$ ). . . . .	121
6.7	The LIWC features above the zero line are in favour of the <i>social</i> ; otherwise, they are in favour of the <i>solo</i> ( $ps < 0.001$ ). . . . .	122
6.8	Moods above the zero line are in favour of the <i>old</i> ; otherwise, they are in favour of the <i>young</i> . . . . .	123
6.9	Moods above the zero line are in favour of the <i>social</i> ; otherwise, they are in favour of the <i>solo</i> ( $ps < 0.001$ ). . . . .	124
7.1	Illustration of profiles for six Livejournal communities: <i>Cooking</i> and <i>Bentolunch</i> , consider similar topics; as do <i>Bookish</i> and <i>50bookchallenge</i> ; and <i>Pokemon</i> and <i>Pkmcollectors</i> . . . . .	129
7.2	An example of blog posts posted in <i>obama_2008</i> community. It is tagged with the ‘current mood’ <i>good</i> . ANEW words used in the blog content are highlighted, including <i>mother</i> , <i>hell</i> , <i>proud</i> , <i>mind</i> , <i>lost</i> and <i>month</i> . . . . .	129
7.3	Topic proportions of 100 communities. . . . .	133
7.4	Above: topic proportions of 10 communities. Below: example topics and most likely words sized by $p(\text{word} \mid \text{topic})$ . . . . .	134
7.5	Communities and mood usage. . . . .	136
7.6	Topic-based hyper-communities with Livejournal category; multi-coloured clusters are less pure. . . . .	142

7.7	Aggregated mood distribution for two users, plotted by valence (pleasure) x-axis, and arousal (activation) y-axis. Larger mood labels indicate more frequent occurrence. . . . .	146
7.8	Interests and tags as features. . . . .	148
7.9	Community classification results. . . . .	149
A.1	Happy, sad and angry and related moods in the hierarchical layout. . . . .	158
A.2	The input Web user interface for querying bursty moods and related events. . . . .	159
A.3	Partial result of the bursty moods, related events, topics and named entities returned for the query shown in Figure A.2. . . . .	160

# List of Tables

2.1	The contingency table for category $c_i$ . . . . .	26
2.2	The overall contingency table for classification. . . . .	26
2.3	LIWC language groups of the CHI06 corpus. . . . .	35
3.1	Top 10 moods in the datasets. . . . .	53
3.2	Mood classification results for different feature selection schemes and for different feature subsets. Different combinations of selection methods and feature spaces are run, but we report only the top results sorted in ascending order of F-score. . . . .	57
4.1	Livejournal moods clustered by similarity to ANEW word use. . . . .	71
4.2	Mood and ANEW correlation. . . . .	76
5.1	The events detected for the dates of lowest sentiment (based on the topics highest ranked by Google Timeline) and their annual ranking by CNN. . . . .	91
5.2	Top 10 bursty events, according to the Google Timeline results, accompanied by CNN's annual ranking, detected from the bursty time of moods. . . . .	97



5.3	The set of moods found bursty in the 9/11 event and the related topics found in the blog posts tagged with the bursty moods in the bursty periods. . . . .	98
5.4	The capacity of KLB and iKLB to detect the top 70 events. . . . .	102
5.5	The capacity of KLB and iKLB to detect the top 70 events (continued from Table 5.4). . . . .	103
5.6	The bursty periods associated with the mood <i>sad</i> detected by KLB and the corresponding events, accompanied by the Google Timeline results and annual ranking by CNN. Only the top-ranked topics (by the Google Timeline results) are shown. . . . .	104
6.1	The mean and standard deviation (in brackets) of connectedness variables between <i>social</i> and <i>solo</i> bloggers categorised by another connectedness variable. . . . .	114
6.2	Differences in topic consideration by old and young bloggers ( $ps < 0.001$ ). . . . .	118
6.3	Differences in topic consideration between the <i>social</i> and the <i>solo</i> ( $ps < 0.001$ ). . . . .	119
7.1	Communities from 10 Livejournal directories used in experiments. . . . .	130
7.2	20 topic-based hyper-communities (exemplar listed first for each). . . . .	138
7.3	Nine mood-based hyper-communities. . . . .	139
7.4	15 ANEW-based hyper-communities. . . . .	140
7.5	12 LIWC-based hyper-communities, showing pie charts of favourite LIWC features; Livejournal categories; and grouped communities. . . . .	141
7.6	CP and NMI of the clusterings based on different community representations. . . . .	142

7.7	The intra-hypergroup and inter-hypergroup JS-based distances on the topic distribution. . . . .	143
7.8	Examples of users joining more than one community and their valence. . . . .	147
7.9	Communities used in the community membership prediction: the number of members and posts. . . . .	147

# List of Algorithms

5.1	Two-state automaton to detect bursts incrementally (iKLB). . . . .	96
-----	--	----

# Abstract

Social media allows people to participate, express opinions, mediate their own content and interact with other users. As such, sentiment information has become an integral part of social media. This thesis presents a sentiment-based approach to analyse content and social relationships in social media.

First, this thesis aims to construct building blocks for sentiment analysis in social media, using sentiment in the form of *mood*. To that end, the problem of supervised mood classification is investigated. This line of work provides insights into what features in a generic document classification problem can be transferred to a mood classification problem in social media. As data in social media is normally large scale, novel scalable feature sets are introduced for this task. In particular, a novel set of psycholinguistic features is proposed and validated, which does not require a supervised feature selection phase and can therefore be applied for mood analysis at a large scale. Next, under an unsupervised setting, this thesis explores the new problem of pattern discovery in social media using sentiment information. The result is the discovery of intrinsic patterns of moods, each of which can be considered as a group of moods similar to a basic emotion studied in psychology, and therefore providing valuable empirical evidence about the structure of human emotion in the social media domain in a data-driven approach.

The second major contribution of this thesis explores the use of sentiment information conveyed in on-line social diaries for detection of real-world events in a large-scale setting. In particular, this thesis introduces the novel concept of ‘sentiment burst’ and employs a stochastic model for detection, and subsequent extraction, of events in social media. The resultant model is a powerful bursty detection algorithm

suitable for on-line deployment on ever-growing datasets such as social media. An additional contribution in this line of work is an effective method for evaluating and ranking events using Google Timeline. This offers an objective measure by which to evaluate event detection—a topic that is largely under explored in the current literature due to a general lack of human groundtruth.

Next, under an egocentric analysis, sentiment information is used to study the impact of the demographics and personalities of users on the messages they create. In particular, we examine how the age and social connectivity of on-line users correlate with the affective, topical and psycholinguistic features of the texts they author. Using a large, ground-truthed dataset of millions of users and on-line diaries, we investigate various important questions posed in social media analysis, psychology and sociology. For example, is there a difference with regard to topic, psycholinguistic features and mood in the messages written by old versus young users? What features are predictive of a user’s personality? Of extraversion and introversion? Are there features that are predictive of influence? The results obtained by our sentiment-based approach are encouraging, do not require an expensive feature selection phase and thus suggest a new and promising approach for egocentric analysis in the social media domain.

Finally, the sentiment information conveyed in media content is investigated with respect to the networking and interaction aspects of a social media system. Sentiment information is studied in parallel with two other common aspects of social media content: topics and linguistic styles. Sentiment information is proved in this thesis to provide additional insights into the process of community formation. It is also shown to be a powerful predictor of community membership for a message or a user at a lighter computational cost.

# Acknowledgements

This thesis would not have come into existence without the help of many people, to whom I would like to express my thanks.

First, I am deeply grateful to my principal supervisor, Dr Dinh Phung. Had this supervision not been provided, things would have been much more difficult. In particular, I will never forget the time that he talked to me about graphical models at a humble coffee shop three years ago.

It is my pleasure to acknowledge Prof. Svetha Venkatesh for giving me the chance to do research here. This thesis would have been impossible without her guidance, inspiration and patience.

I wish to thank my co-supervisor, Dr Brett Adams. I appreciate that he has slowed down the pace of speech to allow me to keep up with him. Sometimes, he surprises me with my own papers!

I am indebted to Leshed and Kaye at Cornell University and to Mishne at Yahoo! Applied Research for kindly providing data.

I would like to thank my friends and colleagues at IMPCA—Sunil, Truyen, Sonny, Saha and Santu—for their generous support, both technical and personal. Thanks to Kurt for colouring emotions. Thanks to Mary for her administrative assistance. Thanks to Jeff for constantly reminding me not to develop another ‘astronaut pen’!

This thesis is dedicated to my Mother for her continuous encouragement and to the memory of my late Father, who taught me to keep hope alive.

Last, but surely not least, I owe special thanks to my wife and son for their love and tolerance.

# Relevant Publications

Part of this thesis has been published or documented elsewhere. The details of these publications are as follows:

Chapter 3:

- Nguyen, T., Phung, D., Adams, B., Tran, T., and Venkatesh, S. (2010). Classification and pattern discovery of mood in weblogs. In *Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Hyderabad, India, June 2010.

Chapter 4:

- Nguyen, T. (2010). Mood patterns and affective lexicon access in weblogs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop (ACL)*, Uppsala, Sweden, July 2010.

Chapter 5:

- Nguyen, T., Phung, D., Adams, B., and Venkatesh, S. (2011). Event extraction using behaviors of sentiment signals and burst structure in social media. Accepted for publication in the *Journal of Knowledge and Information Systems (KAIS)*, Springer, *Special Issue on Behavior Computing: Modeling, Analysis, Mining and Applications*.



- Nguyen, T., Phung, D., Adams, B., and Venkatesh, S. (2012). Emotional reactions to real-world events in social networks. In *New Frontiers in Applied Data Mining*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7104, Springer.

#### Chapter 6:

- Nguyen, T., Phung, D., Adams, B., and Venkatesh, S. (2011). Towards discovery of influence and personality traits through social link prediction. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Spain, July 2011.
- Nguyen, T., Phung, D., Adams, B., and Venkatesh, S. (2011). Prediction of age, sentiment, and connectivity from social media text. In *Proceedings of the 12th International Conference on Web Information System Engineering (WISE)*, Sydney, Australia, October 2011.

#### Chapter 7:

- Nguyen, T., Phung, D., Adams, B., Tran, T., and Venkatesh, S. (2010). Hypercommunity detection in the blogosphere. In *Proceedings of the ACM SIGMM Workshop on Social Media (WSM 2010)*, Firenze, Italy, October 2010.
- Nguyen, T., Phung, D., Adams, B., and Venkatesh, S. (2012). A sentiment-aware approach to community formation in social media. To appear in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Dublin, Ireland, June 2012.

## Abbreviations

ANEW	Affective Norms for English Words
AP	Affinity Propagation
BC	Bhattacharyya coefficient
CHI	$\chi^2$ statistic
CNN	Cable News Network
CP	cluster purity
DF	document frequency
IDF	inverse document frequency
IG	information gain
iKLB	an incremental implementation of KLB
JS	Jensen-Shannon divergence
KL	Kullback-Leibler divergence
KLB	Kleinberg's burst detection algorithm
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
MI	mutual information
NB	Naive Bayes
NBC	Naive Bayes classifier
NMI	normalised mutual information
PLSA	Probabilistic Latent Semantic Analysis
S&P 500	Standard and Poor's 500 Index
SI	sentiment index
SMO	sequential minimal optimization
SOM	self-organising maps
SVM	Support Vector Machine
TDT	Topic Detection and Tracking
TF	term frequency
TF-IDF	term frequency-inverse document frequency
THD	threshold
Weka	Waikato Environment for Knowledge Analysis

# Chapter 1

## Introduction

People are no longer merely readers on the Internet. Since the development of a new, many-to-many media distribution model termed Web 2.0, on-line users have become news creators, contributing to an explosion of user-generated content, by sharing and interacting with others on the Web. This new media, termed ‘social media’, was recently added to the Collegiate Dictionary of Merriam-Webster in August 2011<sup>1</sup>, and is defined as ‘forms of electronic communication (as Web sites for social networking and microblogging) through which users create on-line communities to share information, ideas, personal messages, and other content’.

On the Web, people are comfortable to express their feeling, attitude and ideas towards events in real life. As a result, the content in social media tends to be more subjective than in other genres such as news. As such, sentiment analysis of social media is a growing and timely area of research [Pang and Lee, 2008]. Existing approaches to analysing content, relationships and interactions in social media have typically relied on topical and/or linguistic analysis of text. Sentiment information, on the other hand, has received less attention as a core component of social media analysis to date.

In this thesis we propose a sentiment-based approach to analyse social media from a local-to-holistic view. First we construct a foundation for sentiment analysis in social media through mood classification and clustering in the new media. Next

---

<sup>1</sup><http://www.merriam-webster.com/>, retrieved September 2011.

we consider the temporal dimension of the sentiment information conveyed in social media for the problem of event detection. Last we utilise sentiment information when analysing the egocentric and networking aspects of social media.

## 1.1 Aims and Approach

The goal of this thesis is to propose a sentiment-based approach to pattern discovery and classification in social media, with the following aims:

- The demonstration of the feasibility and usefulness of an approach based on sentiment information in social media analysis.
- The identification of the advantages and limitations of using the sentiment factor in social media analysis in comparison with other conventional facets.

Given the aims of the thesis, our approach focuses on the sentiment aspect of social media text and makes comparisons where relevant with other factors. These ideas are then implemented as follows:

- Constructing building blocks for sentiment analysis in social media. This includes determining which features are good for mood classification and introducing feature sets scalable to large datasets. Further, patterns of moods are extracted, producing valuable empirical evidence about human emotion structures in social media.
- Investigating whether the collective mood of a set of on-line users can be utilised to provide a sensitive index for detecting social reaction towards events taking place in the real world, and to show how sentiment time series can be utilised to detect real-world events and bursts.
- Determining the differences in expressed sentiment among users of different personalities (for example, extraverts or introverts) and demography (for example, age). The impact of demographics and personality on sentiment is

compared with two other facets of text; that is, its topical and psycholinguistic features. This is followed by an examination of how well these features perform for user profile prediction.

- Exploring latent hyper-groups in social communities through their sharing of sentiment information. This is analysed with reference to topics and language styles shared by hyper-communities. The differences in these properties across communities are explored to determine which are useful—sentiment information, topics, or language styles—for the prediction of the community membership of a user or a blog post.

Data crawled from Livejournal, a popular blogging platform, was selected for experiments in this thesis. The most important reason for this selection is that Livejournal allows users to annotate their posts with their moods at the time of writing. This annotation enables supervised learning, for example, for mood classification. Secondly, in Livejournal, users can publish their digital presence in several ways, such as in their profile, by revealing their gender, age, preferences and memberships. This kind of data is useful for egocentric studies. Further, the site provides space for people of common interests to join ‘communities’. This property makes the data suitable for social networking research.

There exist numerous representations for topics, language styles and sentiment features of textual content. In this thesis, for language styles and sentiment information properties, we use psychology-inspired representations. In particular, for sentiment, the content is represented as bags of sentiment words, for which terms are sourced from sentiment bearing lexicons, such as Affective Norms for English Words (ANEW) [Bradley and Lang, 1999] (see Section 2.5.2). For linguistic styles, the content is characterised by linguistic groups extracted using the Linguistic Inquiry and Word Count (LIWC) package [Pennebaker et al., 2007b] (see Section 2.5.4) in which English words are grouped into one of four high-level categories: linguistic processes, psychological processes, personal concerns and spoken categories. Probabilistic topic models, such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003], are used to capture the topical and conceptual aspects of content.

Finally, state-of-the-art methods in machine learning and data mining are employed for the tasks of classification, clustering and regression.

## 1.2 Contribution and Significance

The main contributions of this thesis are:

- Methods for mood classification and the formulation of a novel use of psychology-inspired sets of features that do not require supervision, and thus suitable for large-scale mood classification in social media.
- Methods for emotion-based pattern discovery using a novel feature representation method and non-parametric clustering. In addition, the thesis provides empirical results from a large dataset about emotion structure in social media.
- Methods for discovering correlation between affective lexicon usage in written texts and mood, and the formulation of a novel use of a term-goodness criterion to discover this sentiment–linguistic association.
- Construction of a novel scheme of sentiment indices for extracting events in the blogosphere and method for incremental burst detection in text streams and novel method for evaluating/ranking detected events.
- Methods for determination of the influence of age and social connectivity using three types of features: topics, linguistic styles and sentiment and method for determining the extent to which the age and connectivity of users can be predicted using these features.
- Methods for discovering latent hyper-groups in social communities using similarity derived from topics, language styles and sentiment. In addition, the thesis provides a comprehensive comparison of using these features to predict the community membership of users and posts.

The significance of this thesis is that it:

- Creates a novel framework for the construction, analysis and mining of web logs, serving as a foundation for the analysis of social media. The solutions to fundamental problems make this work applicable to other new media genres (for example, micro-blogging forums such as Twitter, media warehouses with

community aspects such as Flickr or YouTube, content aggregators such as Digg and dedicated social networks such as Facebook, with their attendant activity streams).

- Provides for better mood classification of text, which is significant in terms of (1) improving mood-based applications, such as Internet search with a mood facet (for example, brand and reputation monitoring, on-line forum surveillance including counter-terrorism, automated and semi-automated monitoring and support for at-risk users) and (2) introducing new features for classification.
- Reveals that global patterns of moods in social networks accord with emotion structures proposed within psychology, such as the core affect model. This data-driven organisation of mood is of interest to a wide range of practitioners in humanities and has many potential uses in sentiment-aware applications. Mood patterns can also be considered to be a source of mood synonyms, which have potential applications in data mining, in which moods are employed as a clustering component for classification. In addition, the association of moods and affective lexicon usage could have potential uses in personalisation applications.
- Develops the concept of ‘sentiment burst’, a novel and incremental implementation of event extraction based on an aggregate emotion measure, which makes it practical for detecting events and accompanying opinions in real-time settings.
- Creates a novel and effective method for evaluating and ranking events, which has a potential use in Topic Detection and Tracking (TDT) research.
- Develops an innovative framework for modelling and mining blog data focused on an egocentric view. The fine-grain information about users extracted here is of great commercial interest to diverse clients, and has applications to personalised information retrieval (which relies on knowledge of user attributes such as age and personality traits to re-rank results) and on-line advertising (in which user profile attributes are employed to better target advertisements based on age, connectivity and user mood).
- Identifies that meta-communities grouped by sentiment characteristics can be a social indicator having potential applications in fields such as mental health

(whereby individuals or groups manifesting negative mood can be targeted for support) or marketing (in which customer communities having the same sentiment on similar topics can be targeted).

## 1.3 Outline of the Thesis

The remainder of this thesis is organised as follows.

*Chapter Two* provides background and literature related to the work in this thesis. It describes the structures and functions of social media systems and the motivations for using sentiment in understanding and analysing social media. It reviews the techniques popular in the mining of social media and the feature spaces used in these data mining algorithms. It then discusses how mood time series can be used in event detection. Finally, the chapter turns to a discussion of how sentiment information can be used to characterise users in an egocentric view, as well as in terms of communities, which manifest networking properties of the social media genre.

*Chapter Three* presents a comprehensive study of different feature selection schemes in machine learning for the problem of mood classification in weblogs. In particular, the thesis introduces two feature sets proposed in psychology, showing that these features are efficient, do not require a feature selection phase and yield classification results comparable to state-of-the-art, supervised feature selection schemes.

*Chapter Four* presents the results of data-driven clustering on a dataset of approximately 18 million blog posts with mood ground truth. These manifest global patterns of mood organisation are found analogous to the core affect model for the structure of human emotion proposed in psychology. The association of mood patterns with an affective lexicon is then investigated.

*Chapter Five* presents approaches to construct time-series sentiment indices, the extrema of which can be correlated with real-world events. A novel concept of sentiment burst is then introduced, employing a stochastic model for detecting bursts in text streams. This provides the foundation for sentiment-based burst detection



and, subsequently, bursty event extraction. An incremental version for a powerful burst detection algorithm is also proposed to detect real-time events in streaming data. Further, effective methods for evaluating and ranking events extracted using a combination of topic modelling and Google search are presented.

*Chapter Six* analyses the effect of age and social connectivity on blog posts. In particular, two hypotheses are made: (1) old and young bloggers demonstrate significant differences in the topics they choose to write about, how they write (linguistic styles) and their moods, and (2) bloggers with different degrees of social connectivity write about different topics and exhibit different writing styles and moods.

*Chapter Seven* introduces a method to group on-line communities in the blogosphere based on the content of their discussion—in terms of topics and linguistic styles—and then proposes a novel approach to addressing hyper-community detection based on users' sentiments. This chapter then identifies what features are useful for community prediction of a blog post or blogger.

*Chapter Eight* concludes the thesis and outlines prospects for future work.

# Chapter 2

## Background

In this chapter we provide a review of relevant literature and a background for the work investigated in this thesis. The chapter starts with a brief introduction to social media, including its structures and functions. It then discusses motivations for using sentiment in understanding and analysing social media. Next, it reviews the specific tools for data mining to be used in this thesis, as well as the feature sets to be used in sentiment classification and clustering in Chapter 3 and Chapter 4. This is followed by a review of how the evolution of mood over time can be used to detect events and bursts, which is the subject of Chapter 5. Subsequently, the chapter looks at *the self* (digital presence) users show in their profile to comment on the associations between personality and moods, languages and the topics that bloggers are interested in. Finally, the chapter turns to a discussion of the networking properties of the social media genre, laying the foundation for the remainder of the thesis.

### 2.1 Social Media: An Overview

Social media consists of two components: media and social relationships. These two components include the following elements.

1. *Users*: Users are the core element of social media, playing various roles. They

are no longer merely consumers of information, as they tend to be in conventional media, such as newspapers, magazines, radio and television. Rather, they are publishers, editors and commentators of ‘news’. Therefore, a social media system should provide the appropriate infrastructure to allow this multi-way conversation, rather than perpetuate the traditional broadcasting model. For these reasons, social media is referred to as user-generated. This is reflected in the move from ‘my home page’ to ‘my channel’.

2. *User profile:* In the user-centric context these new media have brought about, the profile information of users allows people to show their identity. Nearly all social media applications provide a function whereby a user can supply additional information about him or herself—often termed a ‘user profile’ (that is, auxiliary information about a user). This information might include demographic information, such as age, gender and location, a list of interests and hobbies and groups to which the user is subscribed. This information can be found in open-ended declarations, such as in the ‘About Me’ section, or in selections from constrained drop-down lists, such as those of nationality. The user profile element goes a significant way in encouraging the formation of relationships in social networks, as it is in Facebook [Lampe et al., 2007]. The degree of anonymity required when people declare their profile in an on-line community can influence the content interchanged. An effect of this can be the tone of discussion, for example, communities in which real-life identity is known tend to be more polite, as there are real-world consequences to what is said.
3. *Privacy settings:* As with other user-centric systems, social media faces the issue of privacy. To address this, most social media systems allow users to set which information is available to the public and which they would like to keep to themselves. For example, Facebook allows people to share a piece of information with ‘No One’, ‘Friends’, ‘Friends of Friends’, a specific ‘Network’ or ‘Everyone’. Like profiles, privacy effects the content of messages. For example, if your messages are only to be seen by a small group, you might feel safe criticising your boss.
4. *Dominant media:* Though the media part of the new media can be in various kinds of conventional media, a social media application usually has a dominant media or ‘object’ of interest. This helps the system to use the infrastructures

effectively as well as differentiate its ‘products’ in the market for user attention. For example, for Twitter and blogs the object of interest is text, for Flickr it is photos and for YouTube it is video. Further, Delicious focuses on tags (and uniform resource locators, also known as URLs); Last.fm on music and playlists; Upcoming.org focuses on events; and Digg news centres itself on articles. However, social networking services like Facebook have increasing support for other media. The choice of dominant media has resulted in Flickr tending to attract people interested in photography, blogs drawing people with an interest in writing, Facebook drawing those wanting to stay in touch and Twitter drawing those more interested in wider, much less defined ‘conversations’.

5. *Ability to comment:* In multi-way conversation, many types of social media enable their users to comment on the content of others. This has two implications for the current research. First, users should be allowed to decide who is permitted to make comments on their posts, helping them to avoid spam. Second, the structure of the comments (flat or nested; that is, comments on comments) is important. While a flat structure is easily managed, nested commenting allows sub-conversations. Further, while flat comments tend to keep the commentary focused on the original post (for example, the image or video), nested comments encourage sub-discussions that may in fact diverge into completely different topics to the original post. Consequently, if using comments as metadata for the original post, we need to be aware of this effect.
6. *Metadata:* Unlike traditional types of media, which run on broadcasting schemes and rarely receive feedback, social media are often annotated with metadata supplied by the community at large. Specifically, a social media system should allow people to annotate conventional resources with tags (small textual annotations) or rankings (for example, thumbs up/down, favourites and numerical ratings: 1–5). Tags can be free form, as represented by the comments on Flickr images or on Digg news, or can be chosen from a predefined list, as in the case of the current mood tags in Livejournal. Ratings and voting, on the other hand, are often predefined as, for example, five-star rankings accompanied by comments on books in Amazon, *thumbs up/thumbs down* on comments in Digg, or the *Like* option in Facebook. Further, an implicit source of metadata is the user profile, through which people show their interests or preferences.

Through metadata, crowd wisdom can be learned and served, as in the case of manual polls. This kind of data is also said to have potential for searching or personalisation applications [Marlow et al., 2006]. In cases in which tags convey sentiment information, the associated media has the potential for sentiment analysis.

7. *Speed of interaction:* The speed of interaction among users varies in different social media systems. If the dominant media is text, the speed of the corresponding social media is faster than that of a video-sharing media. An example of this is Twitter, in which a tweet is limited to 140 characters. Another policy affecting speed is the censorship mechanism. If the system needs time to check the inputted data before the content becomes accessible to the public, or the data is subject to manual moderation, the speed is lower than comparative media lacking censorship. A reasonable method to address this issue is post-editing, as in YouTube, by which morally or legally objectionable videos are removed. Most social media do not implement this editing policy since it prevents them from attracting users. People who do not know each other well may intentionally choose a medium with a higher latency because it is less intimate.
8. *Networking:* Another inevitable element for a social media system is the connecting capacity offered to its users. For example, for implementing friendship/follower models, Facebook has bi-directional friendship and Twitter has unidirectional followers. It is evident that the ‘friendships’ formed in Facebook are more trusted than those formed in Twitter, since they require mutual acknowledgement. Another way to connect users is through communities: people of common interest can create or join communities they are interested in, as offered in Livejournal.

According to Merriam-Webster, ‘social media’ was first used in 2004.<sup>1</sup> This shows that this kind of media is quite new compared with others, such as newspapers or television. However, the development of this new media has snowballed. As of August 2011, Facebook has more than 750 million active users, accounting for more than 10 per cent of the world’s population. This is despite the fact that Facebook is only one member of the rich ecology of social media.

---

<sup>1</sup><http://www.merriam-webster.com/>, retrieved September 2011.

Different types of social media can be categorised based on the functional building blocks on which they are primarily focused [Kietzmann et al., 2011]. According to these criteria, Facebook centres on relationships, meeting the demand of relating to others. Other social media members primarily concentrate on one of six other functional building blocks, as outlined in Kietzmann et al. [2011]. These include identity (for users to reveal themselves, such as in LinkedIn), presence (for showing the available status of users, such as in Foursquare), sharing (for sharing content, such as via YouTube) and conversations (for communicating among users, such as through Twitter). Further, some social media focus on reputation (for showing the standing of users and content, such as by the number of followers of a given user on Twitter, or the number of ‘thumbs up’ on a book on Amazon) or groups (for forming communities of common interests, as in the case of Livejournal).

Others have categorised social media in terms of functionality. Kaplan and Haenlein [2010] base their classification system on two concepts, being the media aspect (social presence/media richness) and the social aspect (self-presentation/self-disclosure). With respect to the first dimension, social media are considered in terms of the extent of the acoustic, visual and physical content they can convey. This relates to the degree of social presence of the communication partners and to the amount of information (richness) of the communication channels. The second dimension concerns the capacity to show the self that is provided by social media. Based on these measures, social media are classified into six groups: collaborative projects (for example, Wikipedia), blogs and microblogs (for example, Twitter), content communities (for example, YouTube), social networking sites (for example, Facebook), virtual game worlds (for example, World of Warcraft) and virtual social worlds (for example, Second Life). Following this grouping scheme, on the media measure, blogs and collaborative projects score lowest since they are normally in text, which contains a poorer amount of information in comparison with other media. Having the added capacity of sharing multimedia content, content communities and social networking sites score higher for the media aspect, whereas virtual social worlds or virtual game worlds score highest due to their simulations of face-to-face interactions. On the other hand, on the social dimension, blogs and virtual social worlds score highest since users of these kinds of media are free to express themselves unlike users of collaborative projects or virtual game worlds, in which they are forced to follow certain regulations.

## 2.2 Weblogs

Due to their richness in conveying sentiment information, weblogs—a representative of social media—were chosen as the primary domain of investigation in this thesis. We note, however, that several methods developed in this thesis are applicable to social media analysis in general. A weblog (or blog) is defined by Merriam-Webster as

‘a Web site that contains an on-line personal journal with reflections, comments, and often hyperlinks provided by the writer’ or ‘a Web site on which someone writes about personal opinions, activities, and experiences’.<sup>2</sup>

The ‘personal opinions or reflections’ property makes this type of social media appealing for opinion mining and sentiment analysis. In fact, blog text is found second only to emotional writing in affective processes, such as negative emotions, including anxiety, anger and sadness [Pennebaker et al., 2007b]. Further, a weblog is defined in the Routledge Encyclopedia of Narrative Theory as ‘a frequently updated website consisting of dated entries arranged in reverse chronological order so the most recent post appears first’.<sup>3</sup> The temporal nature of blog posts thus raises the problem of detecting events and bursts in time series data. While many blogs are kept confidential from others, as in real-life logs (diaries), blogs are often open to the public and commented on by viewers. Blogs and the conversations they provoke define the blogosphere [Herring et al., 2005].

Weblogs are in fact an old member of social media. According to Merriam-Webster, ‘weblog’ was first used in 1997. As with other kinds of social media, weblogs have seen a rapid development. For example, as of September 2001, Livejournal, a popular weblog, had 41.1 million journals and communities; with 164.8 thousand posts uploaded daily at the site.<sup>4</sup> The growth of blogs has attracted much attention. Research issues in the blogosphere include blog clustering, blog mining, community discovery, influence in blogs and propagation, trust and reputation, and filtering of spam blogs [Agarwal and Liu, 2008]; Hearst et al. [2008] outline three main tasks in

---

<sup>2</sup><http://www.merriam-webster.com/>, retrieved September 2011.

<sup>3</sup>[http://jilltxt.net/archives/blog\\_theorising/final\\_version\\_of\\_weblog\\_definition.html](http://jilltxt.net/archives/blog_theorising/final_version_of_weblog_definition.html), retrieved September 2011.

<sup>4</sup><http://www.livejournal.com/>, retrieved September 2011.

a blog search engine: determining what is currently happening in the blogosphere, finding good blogs to read and making archived blog posts valuable.

From the perspective of information access, Mishne [2006] shows three major computational characteristics of the blogosphere, which could be considered challenges as well as opportunities. The first is the *language* used in blogs. This is often informal, combining written and spoken styles and containing slang, jargon and abbreviations. This may be a valuable means of discovering latent groups that are characterised by a unique underlying language style. The second is the structure of the many types of *links* existing in the network, each of which may contain a somewhat different meaning. The last characteristic of the blogosphere is the *timeline* of blogs. Since the content of blogs is often sensitive to time, we should take into account this factor when detecting topics and tracking their changes over time at the scale of a single blog as well as across the whole blogosphere. Mishne [2006] also points out some concrete tasks on which to focus. These are uncovering the person behind a blog, connecting these profiles to detect a latent community, improving the quality of searching in a large amount of blog data and discarding spam blogs.

To confront these challenges in blogosphere computation, some approaches have been proposed. Since members of the blogosphere can be treated as nodes and their relationships as edges in a social network, existing techniques proposed in network science can be adapted to blog analysis. For example, network-centrality measures, including degree, closeness, betweenness and eigenvector centrality, can be used to gain insight into the structural elements, importance, influence or roles of nodes in the blogosphere [Agarwal and Liu, 2008]. In addition, some models from epidemiology can be used to model the process of information diffusion in the network [Gruhl et al., 2004].

Techniques applied to other types of data, such as documents and emails, can be applied to the same problems in blog text. These techniques include the Content-Time-Relation algorithm proposed by Song et al. [2005] to model and predict the personal behaviour of disseminating information via emails; the Author-Recipient-Topic model [McCallum et al., 2007a], used to learn topics based on the email messages sent between entities; and the Group-Topic model proposed by McCallum et al. [2007b], used to extract joint topics and groups in text and voting record data. The mixed-membership model is another appealing framework [Erosheva et al., 2004].



Under this framework, probabilistic latent semantic indexing [Hofmann, 1999] and LDA [Blei et al., 2003] are the two popular probabilistic topic models. In this thesis, probabilistic approaches which have popularly been used in language models to summarise document content will be used to learn latent topics in a given corpus of blog posts.

## 2.3 Sentiment Analysis in Social Media

So much of the web today is two-way; that is, users are able to not only read commercially owned content but can also respond to it. The facility to comment on what was previously only broadcast media, such as news articles, has led to the creation of an unprecedented amount of associated sentiment-laden text. This adoption of user-contribution by the on-line arms of traditional media and industry seems to have been driven by the emergence of social media—YouTube, Flickr, Facebook, Twitter and blogs—through which whole communities contribute media of various types and transact via on-line communication channels. Social media’s uniquely ‘egocentric’<sup>5</sup> nature means that these communities and the artefacts they produce constitute sentiment-laden corpora in their own right.<sup>6</sup>

Hence, the user-generated content is usually opinionated and/or sentiment bearing, bringing opportunity for a company to gain insight into consumer opinions about its products and those of its competitors. Thus, the ability to identify opinion sources on the Web and monitor them is a growing research field, termed opinion mining. This line of research mainly focuses on subjectivity detection, sentiment classification, joint topic-sentiment analysis and opinion summarisation. In this thesis, we use sentiment in the form of *mood*. Mood is a strong form of sentiment expression, conveying a state of the mind such as being happy, sad or angry (For example, Figure 2.1 shows a list of moods Livejournal recommends its users to tag to their blog posts<sup>7</sup>).

---

<sup>5</sup>For example, blog text was found to have a higher occurrence of the first person singular than conversations [Pennebaker et al., 2007b].

<sup>6</sup>For example, the top reason for writing cited by bloggers is to speak their minds: [www.intac.net/breakdown-of-the-blogsphere/](http://www.intac.net/breakdown-of-the-blogsphere/), retrieved August 2011.

<sup>7</sup><http://www.livejournal.com/moodlist.bml?moodtheme=37>, retrieved September 2011.

### 2.3.1 Subjectivity detection

Subjectivity detection is used to identify whether a given text is subjective or objective. This was the main task of the Blog track challenge at the Text Retrieval Conference (TREC) in 2006 [Ounis et al., 2006]. Contributors were required to retrieve blog posts that expressed an opinion about a given target, with the target being a named entity, a product name or an event.

Subjectivity detection has also been investigated in other media. Using a corpus of Wall Street Journal (WSJ) articles, Yu and Hatzivassiloglou [2003] separated opinions from facts at document and sentence level. They achieved a high accuracy for classification of documents into facts (*Editorial* and *Letter to Editor*) and opinions (*Business* and *News*) genres. Murray and Carenini [2011] investigated subjectivity detection tasks in both spoken and written conversations, using a combination of lexical and conversational features. Using a similar approach, Somasundaran et al. [2007] experimented with subjectivity classification of multi-party meetings. They found that both lexical and discourse features can be helpful for the classification task.

Wiebe et al. [2004] investigated subjective language in a corpus of documents, looking for clues, such as low-frequency words or low-frequency collocations. Their findings indicate that if two words are both subjective, they are likely to be similar in their distributions. Certain techniques in natural language processing can also be utilised for subjectivity detection [Wilson, 2008]. In this thesis, these techniques can be taken into account in the sentiment classification problem since it is found that to certain polarity classifiers, the subjectivity extractions are proved to be more efficient input than the original documents [Pang and Lee, 2004].

### 2.3.2 Sentiment classification

Sentiment classification entails identifying the point of view of given text, for example, to determine whether a movie review is positive or negative. The performance of this task can be improved if only the subjective portions of the text, identified during subjectivity detection, are input to the classifier [Pang and Lee, 2004].

Sentiment classification has been considered at the document level, where the sentiment orientation of a whole document is considered, and at the sentence level [Kim and Hovy, 2004]. It has also been explored at feature levels, where different sentiment orientations for different features of a document are examined [Ding et al., 2008] (for example, in a review such as, ‘the acting is top notch but the script is sparse’, the word ‘acting’ is attributed positive sentiment whereas the word ‘script’ is attributed with negative sentiment).

Beyond identification of binary sentiment orientation (up versus down or positive versus negative), work on sentiment classification has also been considered at ordinal scales (for example, one to five stars, of an author’s review [Pang and Lee, 2005]).

There are three basic existing approaches to determining the opinion of a given text: lexicon-based, dictionary-based and corpus-based. Lexicon-based methods use a list of sentiment bearing words, which are classified as positive/negative. For example, in LIWC, words are categorised into *posemo* (positive emotion) or *negemo* (negative emotion) groups. Lexicons can also rate terms at a finer scale. For example, in ANEW, the valence of words is on a scale of 1, *very unpleasant*, to 9, *very pleasant*. If a text has more positive than negative opinion words (or the text has high valence), it is considered a positive review, and negative otherwise [Dodds and Danforth, 2010]. Dictionary-based methods use synonyms and antonyms to score the sentiment for words not in the set of seed opinion words [Andreevskaia and Bergler, 2006, Kim and Hovy, 2004]. Corpus-based methods use a corpus of documents with labels of sentiment orientations to train a classifier. Various features are introduced to determine which are good for classification. This approach is regularly implemented by various machine learning techniques [Boiy and Moens, 2009, Pang et al., 2002].

### 2.3.3 Mood analysis

Mood analysis can be viewed as a subset of sentiment analysis [Pang and Lee, 2008]. Mood classification in weblogs has been conducted by Mishne [2005], who classifies blog text according to the mood tagged by its author at the time of writing, and by Leshed and Kaye [2006], who predict the user’s state of mind for an incoming blog post. Mood estimation has a wide scope of application (for example, in business



Figure 2.1: Examples of moods suggested by Livejournal.

and government intelligence, product research, ethnographic study of the Internet, community or media recommendation, search with a mood facet and pervasive healthcare). However, mood estimation from text poses challenges beyond those encountered by typical text categorisation and clustering. Leaving aside the complexity of the underlying cognitive processes, the manifestation of mood is coloured by a person's idiosyncratic vocabulary and style, with messages often reflecting a social context, including community norms, history and shared understanding. Consequently, text is often short, informal and punctuated by jargon, abbreviations and frequent grammatical errors.

In addition to classification, clustering mood into patterns is also an important task as it provides clues about human emotion structures, with implications for sentiment-aware applications such as sentiment-sensitive text retrieval. The structure of mood organisation has been the focus of investigation from a psychological perspective for some time. For example, Russell [1980, 2003] proposes the *circumplex model of affect* to represent affect states. Using this model, emotion names can be placed around the perimeter of a circle in two-dimensional space. The dimensions defining the space are pleasantness (or valence) and activation (or arousal). Figure 2.2 shows examples of moods according to their distribution in valence and arousal space. However, the structure of mood formation has not been investigated from a data-driven and computational point of view. This thesis aims to discover intrinsic

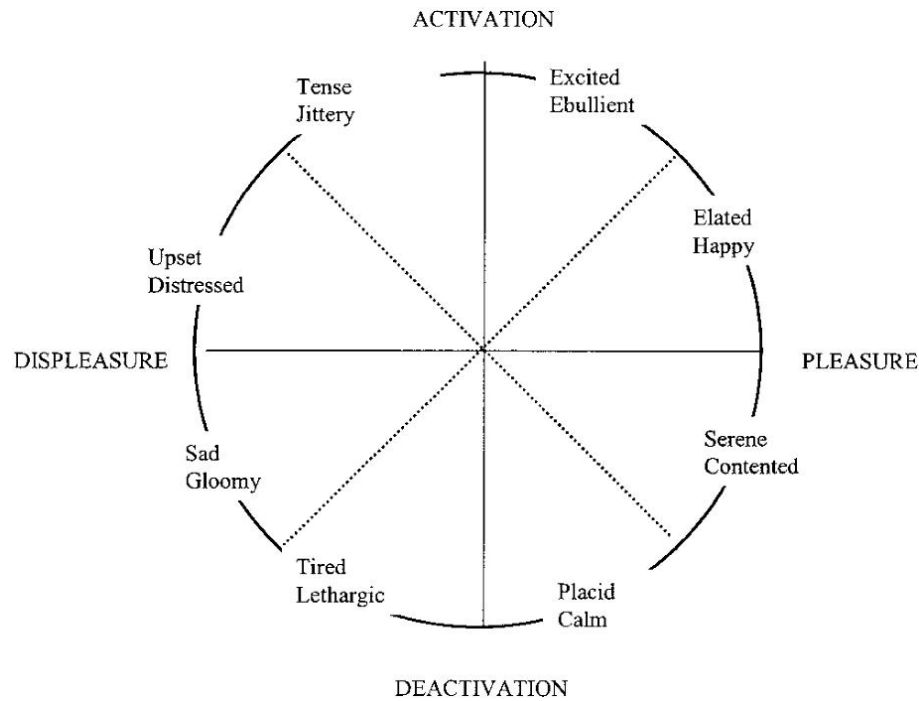


Figure 2.2: Core affect model expressing emotion structure on a 2D plane of valence and arousal (from Russell [2003]).

patterns in mood structure using unsupervised learning approaches. Using a large, ground-truthed dataset of approximately 18 million posts introduced in Leshed and Kaye [2006], it seeks empirical evidence to answer commonly posed questions from psychology. For example, does mood follow a continuum in its transition from ‘pleasure’ to ‘displeasure’ or from ‘activation’ to ‘deactivation’? Is ‘excited’ closer to ‘aroused’ or ‘happy’? Does ‘depressed’ transit to ‘calm’ before reaching ‘happy’? These are interesting and important research questions that have long been the focus of conjecture in different fields, but have not been extensively investigated from a data-driven perspective. The work of Russell [1980], for example, includes only 36 participants, and is thus far smaller in scale than the dataset used here.

Mood clustering has also been found in Hu and Downie [2007], in which the authors performed clustering on music genres, artists and usages in term of moods used, and in Leshed and Kaye [2006], in which the authors grouped blog posts to find mood synonymy. The idiosyncratic nature of mood attributions for the domain of music leads Hu and Downie [2007] to cluster, rather than classify, music genres, artists and usages into data-derived ‘mood spaces’. Some approaches employ a clustering component for mood classification. For example, Sood and Vasserman [2009] use

K-means clustering to obtain a much reduced set of mood classes from the popular blog site, Livejournal's, predefined 132 moods. Mood classification of blog posts is reduced to a ternary classification from among happy, sad and angry, and is achieved using a number of generic text features plus some that are mood specific (for example, emoticons and Internet slang), with an average F-measure of 0.66. Their mood classifier is used as part of a mood-aware search interface.

### 2.3.4 Joint topic-sentiment analysis

A normal setting for sentiment classification assumes the given text (a single document or a collection of documents) is on-topic with the subject of interest, for example, a set of movie reviews on 'Titanic'. However, it is possible that parts of the text are off-topic, for example, 'the weather is terrible but the movie is great', which potentially degrade the performance of sentiment analysis. Thus, a joint topic-sentiment analysis, including identification of topics in the text and the opinions/sentiment associated with each of them, can be an effective approach to improve performance. For example, Riloff et al. [2005] utilised subjectivity analysis in information extraction systems. They found that topic-based and subjectivity filtering are complementary to enhance the performance of the system. Lin and He [2009] proposed a joint sentiment/topic model, a probabilistic modelling framework based on LDA, to detect sentiment and topic simultaneously from text. Jijkoun et al. [2010] presented a method for automatically deriving topic-specific subjectivity lexicons from a general purpose polarity lexicon. It produces subjective on-topic information, improving the performance of an opinion retrieval system.

### 2.3.5 Opinion summarisation

On the Internet, the number of reviews on a popular product can be huge, making it difficult for companies to follow customer opinions, that might otherwise help them to improve their products and services. Opinion summarisation of, say all customer reviews of a product, is a response to this need. Hu and Liu [2004], Liu et al. [2005], Hu and Liu [2006] propose a three-step procedure for this problem of opinion summarisation, given a set of reviews of a product: (1) identify product features that

customers have commented on, (2) identify review sentences and determine whether each of them is positive or negative, and (3) summarise opinions for each feature on the product. The output provided by such a system is not just beneficial for the manufacturers but also for potential customers making decisions about whether to purchase the product. Ku et al. [2005, 2006] extracted opinions on news or writings. They include a step for major topic detection to discard non-relevant sentences before identification of the opinion of sentences, and subsequent summarisation, achieving a high performance for opinion extraction.

## 2.4 Techniques for Mining of Social Media

Two popular techniques for data mining are used in this thesis: classification and clustering.

### 2.4.1 Classification

Given a document  $d$ , such as a blog post, we are interested in inferring the *mood* of the user at the time of composing this document based on textual features extracted from  $d$ . This is somewhat similar to the process of categorising a text document into genre, such as politics or entertainment, but known to be much more challenging due to discrepancies in users' perceptions of mood and differences in composing styles [Pang and Lee, 2008].

This thesis employs two popular classifiers, naive Bayes (NB) and support vector machines (SVM), for classification purposes. Despite differences in the underlying mechanisms of the two algorithms, both were found to work well in generic text categorisation [Sebastiani, 2002]. A logistic regression model was also employed to learn discriminating features, and was found to perform equally well, with the additional advantage of being able to assess the influence of individual features.

### 2.4.1.1 Naive Bayes classifier

Denote by  $\mathcal{M} = \{\text{sad, happy, ...}\}$  the set of all mood categories and by  $K$  the number of moods ( $K = |\mathcal{M}|$ ). Let  $\mathcal{D}$  be the training dataset consisting of  $n$  data points

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^m, y_i \in \{1..K\}\}_{i=1}^n$$

where  $\mathbf{x}_i := \{x_i^1, x_i^2, \dots, x_i^m\}$  denotes the data vector in the feature space,  $y_i$  denotes its class label, and  $m$  denotes the number of features chosen to characterise a blog post. The objective of a NB classification is to assign a class label  $k^*$  to an unseen post  $\mathbf{x}_{new}$ :

$$k^* = \underset{k}{\operatorname{argmax}} \{Pr(y_{new} = k \mid \mathbf{x}_{new}, \mathcal{D})\} \quad (2.1)$$

where

$$Pr(y_{new} = k \mid \mathbf{x}_{new}, \mathcal{D}) \propto Pr(y_{new} = k \mid \mathcal{D}) Pr(\mathbf{x}_{new} \mid y_{new} = k, \mathcal{D}) \quad (2.2)$$

The prior probability for the class  $k$ ,  $Pr(y_{new} = k \mid \mathcal{D})$  is estimated from training data as

$$Pr(y_{new} = k \mid \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n I(y_i = k) \quad (2.3)$$

where  $I(y_i = k) = 1$  if  $y_i = k$  and  $I(y_i = k) = 0$  if  $y_i \neq k$ .

NB imposes an independence assumption among the features so that the term  $Pr(\mathbf{x}_{new} \mid y_{new} = k, \mathcal{D})$  can be factorised into

$$Pr(\mathbf{x}_{new} \mid y_{new} = k, \mathcal{D}) = \prod_{j=1}^m Pr(\mathbf{x}_{new}^j \mid y_{new} = k, \mathcal{D}) \quad (2.4)$$

where  $Pr(\mathbf{x}_{new}^j \mid y_{new} = k, \mathcal{D})$ , the conditional probability of the  $j^{th}$  feature given class  $k$ , is estimated using training data under a maximum likelihood framework depending on the choice of the conditional probability. Additional details on NB



can be found in Lewis [1998].

Despite the independence assumption, NB classifiers have been found to perform well in both text categorisation [Lewis, 1998] and sentiment classification [Pang et al., 2002].

### 2.4.1.2 Support vector machine classifier

In a binary classification, SVM constructs a *linear* classifier to maximise the margin between the two data classes. Let  $\mathcal{D}$  be the training dataset with  $n$  data points

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^m, y_i \in \{-1, 1\}\}_{i=1}^n$$

where  $\mathbf{x}_i$  denotes the data vector in the feature space and  $y_i$  denotes its class label. A linear hyper-plane in the feature space can be written as

$$\mathbf{w} \cdot \mathbf{x} + w_0 = 0 \tag{2.5}$$

where  $w_0$  is a bias term and  $\mathbf{w}$  denotes the classifier parameter vector. Given a linearly separable binary classification problem, (either in original data space or in some feature space), two parallel hyper-planes can be chosen such that there are no points between them. The distance between these hyper-planes can be shown to be  $\frac{2}{\|\mathbf{w}\|}$ . Maximising this margin is equivalent to minimising  $\|\mathbf{w}\|$ . To find the optimal classifier, SVM poses the following quadratic programming problem

$$\min_{\mathbf{w}, w_0} \frac{\|\mathbf{w}\|^2}{2}, \text{ subject to } y_i (\mathbf{w} \cdot \mathbf{x}_i + w_0) \geq 1 \tag{2.6}$$

The optimisation constraint in the above expression ensures that the data points from either class do not fall into the margin of the two hyper-planes. It has been shown that this optimisation problem can be converted to an unconstrained optimisation problem:

$$\min_{\mathbf{w}, w_0} \max_{\alpha} \left\{ \frac{\|\mathbf{w}\|^2}{2} - \sum_i \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + w_0) - 1] \right\} \quad (2.7)$$

whose solution can be expressed as a linear combination of the training data vectors as

$$\mathbf{w} = \sum_{j=1}^p \alpha_j c_j \mathbf{x}_j \quad (2.8)$$

where  $\alpha$  denotes non-negative Lagrange multipliers. The data vectors  $\{\mathbf{x}_j\}_{j=1}^p$  are also called support vectors as they lie on the margin and satisfy  $y_j (\mathbf{w} \cdot \mathbf{x}_j + w_0) = 1$ . Using this condition,  $w_0$  is given as  $w_0 = y_j - \mathbf{w} \cdot \mathbf{x}_j$ .

To perform nonlinear classification, one can transform the training data to a new feature space in which the feature vectors are linearly separable. This is often possible if the feature space is high-dimensional. Although, this allows the data to be linearly separable, high-dimensional spaces have their own challenges due to the curse of dimensionality and computational difficulty. A common solution to this problem is the use of the ‘kernel trick’. Using this trick, the dot products in the feature space are computed by using a kernel function in the original data space.<sup>8</sup>

For traditional text categorisation, SVMs have been found to be effective, outperforming the NB [Joachims, 1998]. In sentiment classifications, they have also been found to be better than NB classifiers in certain feature sets [Pang et al., 2002].

In this thesis, for NB classifiers, the standard NB algorithm is taken from the Waikato Environment for Knowledge Analysis (Weka) suite [Hall et al., 2009]; for SVM classifiers, the Weka implementation of sequential minimal optimisation (SMO) [Hall et al., 2009] is used. All experiments run under all default settings in the Weka package, unless otherwise specified.

---

<sup>8</sup>Many standard kernel functions have been used, such as Gaussian kernel or polynomial kernels. See Burges [1998] for details.

### 2.4.1.3 Logistic classification

Suppose our goal is to predict if the user gender is *male* or *female* given his/her feature vector  $\mathbf{x}_{new} := \{x_i^1, x_i^2, \dots, x_i^m\}$ , the logistic model defines:

$$\text{logit}(p_{\mathbf{x}_{new}}) = \log \frac{p_{\mathbf{x}_{new}}}{1 - p_{\mathbf{x}_{new}}} = \alpha + \sum_{i=1}^m \beta_i x_i^i \quad (2.9)$$

where  $\alpha$  and  $\beta$  are regression parameters, and  $\frac{p_{\mathbf{x}_{new}}}{1 - p_{\mathbf{x}_{new}}}$  is the odds ratio. Then  $p_{\mathbf{x}_{new}}$ , without loss of generality, is the probability for the user being *male* and determined as:

$$p_{\mathbf{x}_{new}} = \frac{e^{\alpha + \sum_{i=1}^m \beta_i x_{new}^i}}{1 + e^{\alpha + \sum_{i=1}^m \beta_i x_{new}^i}} \quad (2.10)$$

If  $p_{\mathbf{x}_{new}} > 0.5$  the user will be classified as *male*, otherwise, the user will be classified as *female*. The likelihood function for the training data consisting of  $n$  data points  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$  is

$$\mathcal{L} = \prod_{j=1}^n p_j^{y_j} (1 - p_j)^{1 - y_j} \quad (2.11)$$

where  $y_j = \{0, 1\}$  is the label of the  $j^{th}$  data point and  $p_j$  is the probability for the  $j^{th}$  user being *male* determined as in Equation 2.10.

The parameters  $\alpha$  and  $\beta$  are estimated from training data using the standard maximum likelihood approach [Wasserman, 2004] using Equation 2.11. The resulting model is then used to classify unseen data. If  $\beta_i$  is greater than zero, an increase (or decrease) in  $x^i$  value is associated with an increase (or decrease) in the logit( $p$ ) (and thus  $p$ ). Conversely, if  $\beta_i$  is less than zero, an increase (or decrease) in  $x^i$  value is associated with a decrease (or increase) in  $p$ . If  $\beta_i$  equals zero,  $x^i$  has no predictive power for predicting the sentiment orientation of the blog post. This explicit interpretation is one of logistic regression's advantages, allowing for easy examination of the predictive effect of each feature once the parameter is estimated.

		Groundtruth	
		$c_i$	$\bar{c}_i$
Prediction	$c_i$	$TP_i$ (true positive)	$FP_i$ (false positive)
	$\bar{c}_i$	$FN_i$ (false negative)	$TN_i$ (true negative)

Table 2.1: The contingency table for category  $c_i$ .

		Groundtruth	
		YES	NO
Prediction	YES	$TP = \sum_{i=1}^{ \mathcal{M} } TP_i$	$FP = \sum_{i=1}^{ \mathcal{M} } FP_i$
	NO	$FN = \sum_{i=1}^{ \mathcal{M} } FN_i$	$TN = \sum_{i=1}^{ \mathcal{M} } TN_i$

Table 2.2: The overall contingency table for classification.

#### 2.4.1.4 Classification evaluation

To evaluate the performance of classification, this thesis uses four widely used metrics: accuracy, precision, recall and F-measure. Taking a mood classification as an example, given  $\mathcal{B}$  the corpus of blog posts and  $\mathcal{M} = \{\text{sad, happy, angry}\}$  the set of mood categories, the ground truth is taken from the current mood tags in the blog posts. Then, the number of true positive occurrences  $TP_i$  in a classification for class  $c_i \in \mathcal{M}$  is defined as the quantity of the blog posts tagged with class  $c_i$  and predicted correctly, also  $c_i$ . Other numbers of false positive ( $FP$ ), false negative ( $FN$ ) and true negative ( $TN$ ) occurrences for class  $c_i$  are defined in Table 2.1.

The quantities of true positive (also known as *hit*), false positive (*false alarm*), false negative (*miss*) and true negative (*correct rejection*) occurrences for the classification of all classes is defined as the total of their corresponding numbers in all classes, as shown in Table 2.2. Given these quantities for all classes, the accuracy, precision, recall and F-measure for the classification are defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (2.12)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.13)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.14)$$

$$F\text{-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.15)$$

With respect to these formulas, *accuracy* implies how correctly a classifier predicts a data point to be or not to be an example of a class. On the other hand, high *precision* means high hit and low false alarms, while high *recall* means high hit and low missing detection, and *F-measure* is the weighted harmonic mean of precision and recall. High precision is desirable in information retrieval contexts in which one desires to have as many relevant documents as possible in the top returned results. Conversely, high recall is important in classification contexts in which the exhaustive detection of all possible correct outcomes is desired.

### 2.4.2 Clustering with affinity propagation

For clustering tasks, parametric clustering methods, such as K-means, are widely applied. However, parametric clustering methods are not always suitable since we do not know the number of clusters in advance. In this thesis, we often make use of the affinity propagation algorithm (AP) [Frey and Dueck, 2007] for clustering since it can automatically discover the number of clusters. In addition, AP also produces an exemplar for each cluster, and is therefore particularly useful for many tasks considered in this thesis.

AP is grounded in probabilistic graphical modelling and discovers latent clusters by exchanging local messages. For each data point, an exemplar node is created and treated as hidden, subject to be inferred. A factor graph is constructed and function potentials are designed to encode the similarity measure between the data points and to enforce specific constraints to make all exemplar nodes a valid configuration. The energy function equals the sum of all potentials. In the sum, two main types of messages are passed: the responsibility  $s(x,y)$  sent from a point to its candidate exemplar, representing how well  $x$  ‘trusts’  $y$  as its exemplar, and the availability  $a(x,y)$  sent from a candidate exemplar  $x$  to  $y$ , representing how good it would be for  $x$  to choose  $y$  as its exemplar. Performing max-sum message passing to minimise the energy function on this factor graph results in the solution for AP.

Another clustering scheme employed in this thesis to enhance the visualisation of

clusters is the self-organising map (SOM) [Kohonen, 1990]. A SOM is a type of artificial neural network using unsupervised learning to generate a low-dimensional representation (called a map) of the input space. A neighbourhood function is used in SOMs to preserve the topological properties of the input space. SOM has been chosen in this thesis for its efficiency in computation and dimensionality reduction, which can effectively preserve the distance in the lower dimension for visualising the structure of clusters.

### 2.4.2.1 Similarity measures

Clustering algorithms often require similarity between the two data points. To compute the similarity of two points, a distance measure is used. In this thesis, the following distances are employed to estimate the similarity of the two data points: the Euclidean distance, the Bhattacharyya coefficient (BC), the Kullback–Leibler divergence (KL) and the Jensen–Shannon divergence (JS).

Given  $x = (x_1, x_2, \dots, x_n)$  and  $y = (y_1, y_2, \dots, y_n)$ , the Euclidean distance from  $x$  to  $y$  is given by

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.16)$$

If the  $x$  and  $y$  are two discrete distributions, any suitable probability distance functions can be used. In this thesis, two popular measures in information theory are used to compute the similarities: the negative KL [Kullback and Leibler, 1951] and the BC [Ramesh and Meer, 2000]:

$$KL(x||y) = \sum_{i=1}^n x_i * \log \frac{x_i}{y_i} \quad (2.17)$$

and

$$BC(x, y) = \sum_{i=1}^n \sqrt{x_i * y_i} \quad (2.18)$$

Note that KL is not a symmetric distance; that is,  $KL(x||y) \neq KL(y||x)$ . Occasionally, the JS divergence, a symmetrised version of the KL divergence, is used:

$$JS(x||y) = \frac{1}{2}KL(x||z) + \frac{1}{2}KL(y||z) \quad (2.19)$$

where  $z = \frac{1}{2}(x + y)$ .

#### 2.4.2.2 Evaluation of clustering

We evaluate the performance of clustering algorithms using two popular metrics: cluster purity (CP) and normalised mutual information (NMI).

**Cluster purity** The purity of a clustering  $\mathcal{C}$  [Manning et al., 2008] is defined as

$$CP = \frac{\sum_{j \in \mathcal{C}} m_j}{n} \quad (2.20)$$

where  $m_j$  denotes the number of points from the major category in cluster  $j$  and  $n$  is the total number of data points. The higher the  $CP$ , the better the clustering is. It is clear that when all data points are correctly clustered,  $\sum_{j \in \mathcal{C}} m_j = n$  and  $CP = 1$ .

**Normalised mutual information** Given the induced partition  $\mathcal{C}$  with  $P$  labels  $c_1, \dots, c_P$  and true partition  $\mathcal{T}$  with  $Q$  ground-truth category labels  $z_1, \dots, z_Q$ , normalised mutual information [Manning et al., 2008] is defined as

$$NMI = \frac{I(\mathcal{C}, \mathcal{Q})}{[H(\mathcal{C}) + H(\mathcal{Q})]/2} \quad (2.21)$$

where  $I(\mathcal{C}, \mathcal{Q})$  is mutual information

$$I(\mathcal{C}, \mathcal{Q}) = \sum_p \sum_q \frac{|c_p \cap z_q|}{n} \log \frac{n * |c_p \cap z_q|}{|c_p| |z_q|} \quad (2.22)$$

$n$  is the total number of data points,  $|c_p|, |z_q|$  denote the number of data points in  $p$ -th cluster from partition  $\mathcal{C}$  and  $q$ -th cluster from partition  $\mathcal{T}$  respectively,  $|c_p \cap z_q|$  denotes the number of common data points between  $p$ -th cluster from partition  $\mathcal{C}$  and  $q$ -th cluster from partition  $\mathcal{T}$ , and  $H(\cdot)$  is the entropy function

$$H(\mathcal{C}) = - \sum_p \frac{|c_p|}{n} \log \frac{|c_p|}{n} \quad (2.23)$$

The higher the *NMI*, the better the clustering quality is.

## 2.5 Features for Sentiment Analysis

The selection of features for use in data mining algorithms plays an important role in the performance of mining tasks. At times, feature selection is considered more important than the choice of learning algorithms [Mishne, 2005]. In this section, we shall discuss features and feature selection methods. Feature sets used in generic text categorisation and those specific to sentiment representation will be presented. We will use mood classification accuracy as the basis for deciding which feature sets to use in a given setting, especially when considering the trade-off between speed and performance. In this thesis we will place an emphasis on scalable feature sets appropriate for deployment on large scale data like social media.

### 2.5.1 Parts of speech

It has been shown that the use of linguistic components such as adverbs, adjectives or verbs can be strong indicators of mood [Pang and Lee, 2008]. In this thesis we use an existing part-of-speech tagger to identify all terms that can be automatically tagged as verbs, adjectives and adverbs. The tagger used is the SS-Tagger [Tsuruoka and



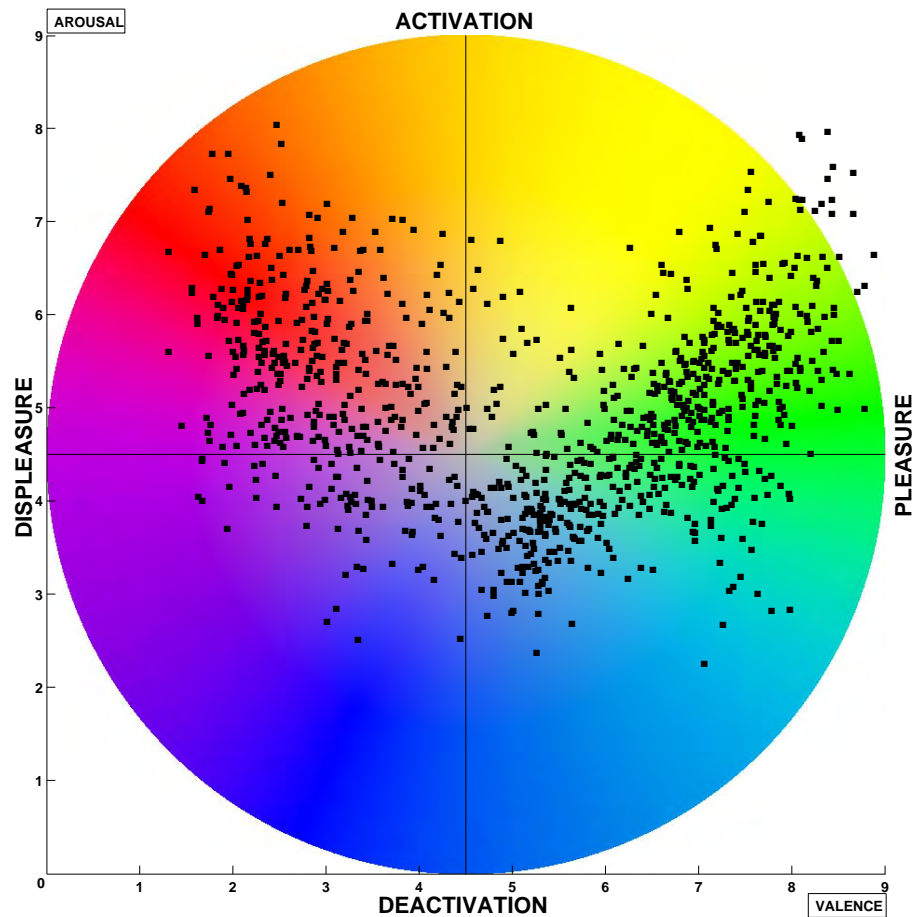


Figure 2.3: Valence and arousal values of 1,034 ANEW words on the affect circle.

Tsujii, 2005] ported to the Antelope Natural Language Processing framework<sup>9</sup> which is reported to achieve reasonable accuracy (for example, the SS-Tagger achieved 97.1 per cent accuracy for the Wall Street Journal corpus, see Tsuruoka and Tsujii [2005] for details).

### 2.5.2 Sentiment bearing lexicon features

For sentiment analysis, some emotion-bearing lexicons that have been subjectively chosen by human labour could also potentially boost performance. In particular we investigate the Affective Norms for English Words (ANEW) [Bradley and Lang, 1999], a set of 1,034 sentiment-conveying English words. ANEW was constructed by

<sup>9</sup>[www.proxem.com](http://www.proxem.com), retrieved September 2011.

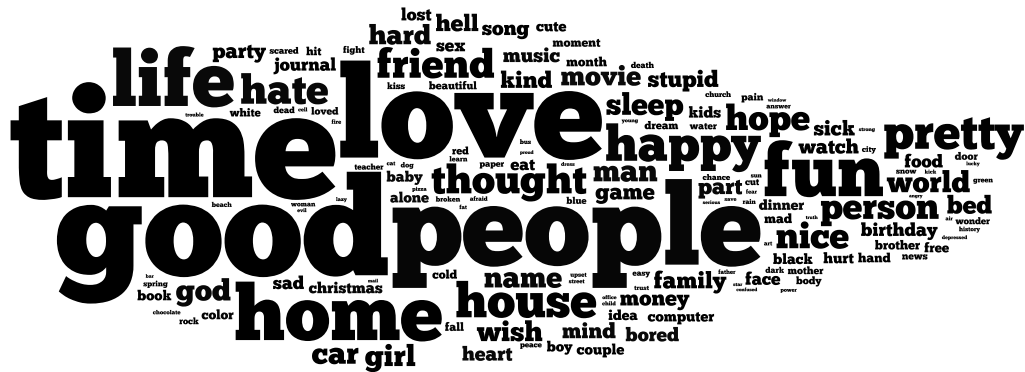


Figure 2.4: Cloud visualisation of ANEW words used in the content of the CHI06 database.

the National Institute of Mental Health of the United States in an effort to create standard verbal materials for use in studies on cognition and emotion. The authors have rated these words in terms of the valence, arousal and dominance they convey. The valence values in ANEW range from 1.25 (*suicide*) to 8.82 (*triumphant*); the arousal values range between 2.39 (*relaxed*) and 8.17 (*rage*); and the dominance values range between 2.27 (*helpless*) and 7.88 (*leader*). The median values for these ranges are 5.29 (*patient*), 5.06 (*sunrise*) and 4.12 (*knife*) respectively.

Two of these dimensions are sufficient to conceptualise an emotion as proposed in psychology [Mauss and Robinson, 2009]. Accordingly, from the dimensional view of emotion formation, Russell [1980, 2009] suggests measuring emotion—a form of expressive sentiment—in terms of valence and arousal. Based on the valence and arousal values provided in the emotion-bearing lexicon ANEW, we plot these words on the affect circle (see Figure 2.3). The figure shows that ANEW words cover all quarters of the circle of the core affect model [Russell, 1980, 2009]. This provides an excellent reference model for understanding and interpreting the emotion patterns extracted in this thesis.

ANEW has been used in sentiment estimation applications. For instance, Dodds and Danforth [2010] used it for estimating happiness levels in three types of data: song lyrics, blogs and the State of the Union addresses. Their work computes the average over valence quantities of ANEW words used in the data to decide their sentiment orientation. Kim et al. [2009] used ANEW to detect sadness in micro-bloggers' responses to the death of Michael Jackson in Twitter. They found that the tweets related to the event used many more negative words than did others.

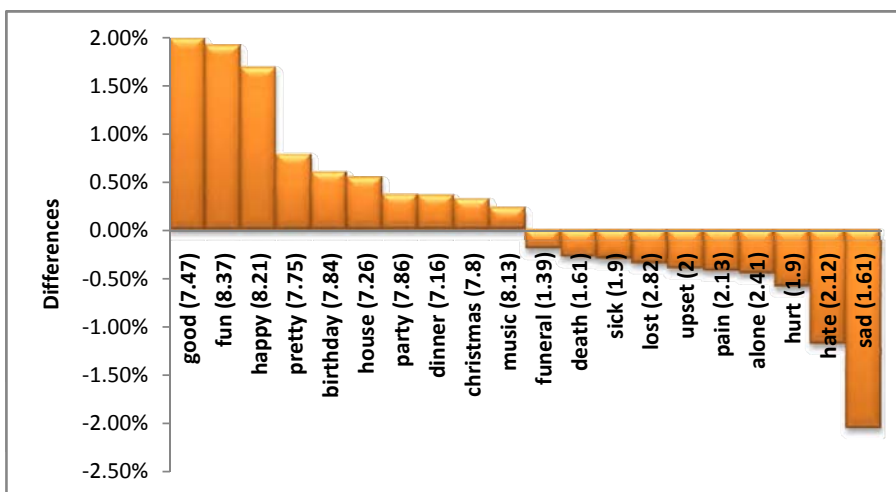


Figure 2.5: Examples of ANEW words (with valence in the parentheses) in favour of happy (above) and sad (below) blog posts. The magnitude shows the difference between the percentage of corresponding ANEW words in the content of happy and sad blog posts.

A cloud visualisation of ANEW words used in the content of approximately 18 million blog posts in the CHI06 dataset, which shall be described in Section 3.3.2, is illustrated in Figure 2.4. In this dataset, on average, a blog post has approximately 14 ANEW words. The top five most frequently occurring ANEW words in this corpus are *time* (4.41%), *good* (3.96%), *love* (3.71%), *people* (3.27%) and *home* (2.23%), consistent with the statistics reported in Dodds and Danforth [2010].

Since the ANEW lexicon contains sentiment-bearing words it is useful for the task of mood classification. Our assumption is that there exists difference in the use of ANEW words among blog posts written by authors having different moods. For example, on average, the proportions of ANEW words in the blog posts tagged with *happy* and *sad* in the CHI06 dataset are not similar, as shown in Figure 2.5. As can be seen, while the blog posts tagged with mood *happy* prefer those ANEW words in high valence such as *good* and *fun*, those tagged with *sad* use more low valence words such as *alone* and *hurt*.

### 2.5.3 Other sentiment related features

In addition to sentiment-bearing words, other features have been found useful in sentiment classification, but otherwise are not popular in generic text categorisation applications. For example, features include length-related, point-wise mutual information [Mishne, 2005], most frequent words [Leshed and Kaye, 2006], or emotion words, emoticons and slang [Sara and Lucy, 2009]. Furthermore, exploiting additional contexts, such as negation or discourse, may also improve the performance of sentiment analysis. Specifically, the appearance of the word ‘NOT’, if taken into account, might cause the sentiment polarity of a sentence to be reversed [Das and Chen, 2007].

### 2.5.4 Psycholinguistic features

Another powerful set of features used in this thesis are psycholinguistic, drawn from the LIWC package [Pennebaker et al., 2007b]. The LIWC package assigns English words to one of four high-level categories: linguistic processes, psychological processes, personal concerns and spoken categories, which are further sub-divided into a three-level hierarchy.<sup>10</sup> The taxonomy ranges across topic (for example, religion and health), mood (for example, positive emotion) and processes not captured by either, such as cognition (for example, causation and discrepancy) and tense. Tausczik and Pennebaker [2010] surveyed the use of the LIWC package, based on the social and psychological meaning of words, across different research areas in sociology and psychology, including status, dominance and social hierarchy, honesty and deception, thinking styles and individual differences.

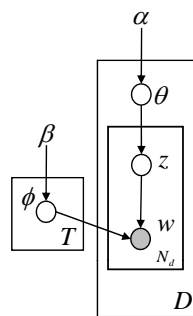
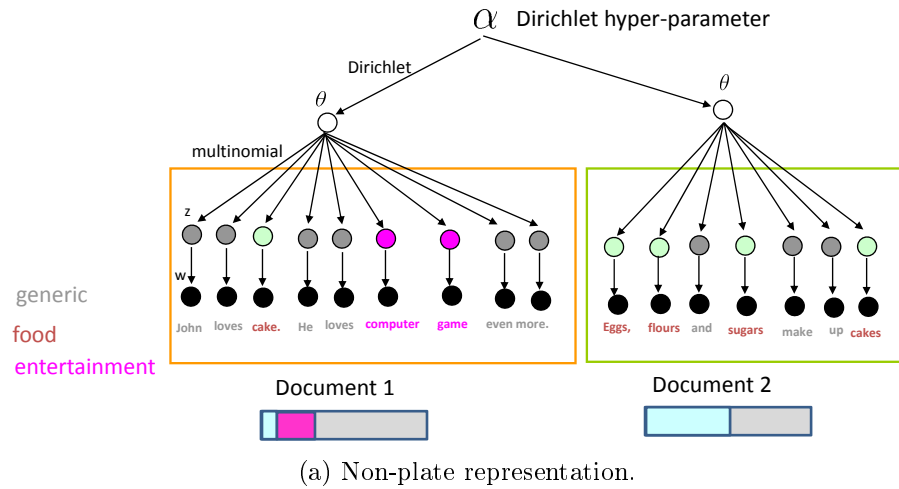
According to Pennebaker et al. [2007a], weblogs are particularly suitable for sentiment analysis since they contain a high rate of words in *affective processes*. For the CHI06 dataset used in this thesis we compute the mean for each of the LIWC groups and present them in Table 2.3. As can be seen, the percentage of words in *affective processes* in the corpus (7.3 per cent) is larger than that of any class of text reported in Pennebaker et al. [2007a], including the emotional writing class (6.02 per cent). This highlights the usefulness of social media-derived corpora for sentiment

---

<sup>10</sup><http://www.liwc.net/descriptiontable1.php>, retrieved July 2011.

Category	Code	%	Category	Code	%
<b><i>Linguistic processes</i></b>			Anger	anger	1.2
Word count	wc	227	Sadness	sad	0.5
Words/sentence	wps	13.3	<b><i>Cognitive</i></b>	cogmech	15.3
Dictionary words	dic	83.4	Insight	insight	1.9
Words >6 letters	sixltr	12.4	Causation	cause	1.3
Function words	funct	53.4	Discrepancy	discrep	1.5
<b><i>Total pronouns</i></b>	pronoun	17.2	Tentative	tentat	2.5
<b><i>Personal pron.</i></b>	ppron	11.9	Certainty	certain	1.3
1st pers singular	i	7.9	Inhibition	inhib	0.4
1st pers plural	we	0.7	Inclusive	incl	4.5
2nd person	you	1.6	Exclusive	excl	2.8
3rd pers singular	shehe	1.2	<b><i>Perceptual</i></b>	percept	2.3
3rd pers plural	they	0.4	See	see	0.9
<b><i>Impers. pron.</i></b>	ipron	5.3	Hear	hear	0.6
Articles	article	4.5	Feel	feel	0.7
Common verbs	verb	15.8	<b><i>Biological</i></b>	bio	2.6
Auxiliary verbs	auxverb	9.3	Body	body	0.9
Past tense	past	4.0	Health	health	0.6
Present tense	present	9.8	Sexual	sexual	0.8
Future tense	future	1.0	Ingestion	ingest	0.4
Adverbs	adverb	5.8	<b><i>Relativity</i></b>	relativ	13.7
Prepositions	prep	10.6	Motion	motion	2.1
Conjunctions	conj	6.3	Space	space	4.9
Negations	negate	1.9	Time	time	6.3
Quantifiers	quant	2.5	<b><i>Personal concerns</i></b>		
Numbers	number	0.7	Work	work	1.6
Swear words	swear	0.7	Achieve	achieve	1.2
<b><i>Psychological processes</i></b>			Leisure	leisure	1.5
<b><i>Social</i></b>	social	8.4	Home	home	0.5
Family	family	0.4	Money	money	0.5
Friends	friend	0.3	Religion	relig	0.4
Humans	human	0.8	Death	death	0.2
<b><i>Affective</i></b>	affect	7.3	<b><i>Spoken</i></b>		
<b><i>Positive</i></b>	posemo	4.6	Assent	assent	1.2
<b><i>Negative</i></b>	negemo	2.7	Nonfluency	nonflu	0.5
Anxiety	anx	0.3	Fillers	filler	0.1

Table 2.3: LIWC language groups of the CHI06 corpus.



- For each document:  $\theta_d \sim \text{Dir}(\alpha)$
- For each topic:  $\phi_k \sim \text{Dir}(\beta)$
- For each position  $i$  in document  $d$ :
  - Sample topic:  $z_i \sim \text{Mult}(\theta_d)$
  - Sample word:  $w_i \sim \text{Mult}(\phi_{z_i})$

(b) Plate representation.

Figure 2.6: Latent Dirichlet Allocation [Blei et al., 2003].

analysis.

## 2.5.5 Topics and topic modelling

Probabilistic approaches have popularly been used in language models to summarise document content. To extract topics in the content topic-modelling, tools such as probabilistic latent semantic analysis (PLSA) [Hofmann, 2001] or LDA [Blei et al., 2003] can be used. In this thesis we use LDA to learn latent topics in a given corpus of texts.

Let  $d_i$  be a post and  $D$  the corpus of all blog entries. LDA is a Bayesian framework in which a topic is defined as a distribution over words (for example, the ‘education’ topic would assign high probabilities to the words ‘student’ or ‘teacher’). Assume there are  $V$  terms among all blog entries, then each topic is expressed quantitatively

as a real-valued vector in which each element is a probability assigned to a word, summing up to 1. A graphical model illustration for LDA is shown in Figure 2.6.

In this figure,  $D$  is the total number of blog entries and for each blog entry  $d$ , a mixing topic proportion  $\theta_d$  is sampled from a Dirichlet distribution with hyper-parameter  $\alpha$ . Each word  $w$  is generated by first sampling a topic  $z$  from  $\theta_d$  and then sampling from  $Pr(w|z, \phi)$ , where  $\beta$  is the Dirichlet hyper-parameter for  $\phi$ . Therefore, LDA models each document as a mixture of topics and each topic as a mixture of words.

Since exact inference is intractable for LDA, we use the Gibbs sampling proposed in Griffiths and Steyvers [2004] for learning post-topic distribution. Further details for LDA and its inference procedure shall be given throughout the thesis as needed.

### 2.5.6 Feature selection methods

For generic text categorisation, a wide range of feature selection methods in machine learning has been studied. Yang and Pedersen [1997] conducted a comparative study of different feature selection schemes, including information gain (IG), mutual information (MI) and  $\chi^2$  statistic (CHI). Their study concluded that IG and CHI are the most effective at dimensionality reduction for text categorisation without compromising classification accuracy, while MI has inferior performance by comparison.

An alternative to a term-class interaction approach to selecting features is to consider term statistics. Thresholds for term frequency (TF) or document frequency (DF) are commonly used in feature reduction in data mining. The joint term frequency-inverse document frequency (TF-IDF) scheme, popular for retrieval settings, is also used in text mining and often outperforms TF and DF.

This study is motivated by the question of whether these methods have similar effects when applied to mood classification and what changes should be made to use them effectively in the social media context.

It is to be noted that the difference in the representation of feature vectors might affect the performance of sentiment classification. It is well known in text mining that the bag-of-words counting representation (that is, the number of appearances of

a term) is an effective practice. However, in sentiment analysis, it has been found that simple binary representation (that is, the use of 1 if the term appears in the document and 0 otherwise) is more effective at movie review classification [Pang et al., 2002]. In addition to its use in classification, a binary feature representation can be better compressed, making it suitable for very large datasets.

## 2.6 Sentiment-based Detection of Events and Bursts

In the previous sections we discussed background related to sentiment classification and clustering. In this section we review works related to using sentiment information as a means for tracking events and detecting bursts.

### 2.6.1 Event detection

TDT has received much research attention. First sponsored by the Defence Advanced Research Projects Agency in 1996 and then the National Institute of Standards and Technology, TDT-related evaluation tasks have been organised with special interest in the transcription and understanding of broadcast news. An emerging part of the TDT initiative is event detection. An *event*, as defined in Allan et al. [1998], is something that happens at a certain place and time. Event detection in stream data aims to extract the first story of a new event, while event tracking groups stories discussing the same event. Much work in this topic has followed a content-based approach [Allan et al., 1998, Kumaran and Allan, 2004, Zhang et al., 2007]. For text documents, content features are extracted using unigrams or n-grams. Subsequently, a document in a text stream is represented as a vector of features. Event detection and tracking is then cast as a simple nearest-neighbour classification problem based on the similarity of the newly arrived document with the existing pool of events. If the similarity exceeds a predefined threshold, the document is classified into the closest event and updated accordingly to include the new document. Otherwise, a new event is formed.

One of the most popular and useful representations for text is the bag-of-word model, which assumes interchangeability among words in the document. The content



represented in this form can be used to answer *what* happens for an event, either by explicitly examining the words or via topic-modelling tools. Other properties characterising an event, such as *where* the event happens and to *whom* the event is related, can also be estimated through entity recognition [Kumaran and Allan, 2004, Zhang et al., 2007]. Das Sarma et al. [2011] used information about relationships among entities in documents to detect events. Spatial and temporal information can also be exploited to detect context-specific events. For example, Makkonen et al. [2003] represented events as vectors of terms and additionally incorporate temporal, location and person identity information. Sakaki et al. [2010] utilised spatiotemporal information in Twitter to find the centre and trajectory of events. Zhao et al. [2007] included information flow patterns and underlying social network information to detect finer granularity events; that is, events are defined by keywords (from content-based clustering results) and temporal segments along with community information (based on information flow in a social network).

Sentiment information conveyed in social media data can be utilised to detect significant events. For example, using Livejournal data, Balog et al. [2006] found the spike time of events based on sudden changes in a mood time series. Gilbert and Karahalios [2010] estimated the level of an ‘*anxious*’ index shown to predict the trend of Standard and Poor’s 500 index (S&P 500). Using Twitter data, Thelwall et al. [2011] found that the occurrence of popular events is linked to increases in negative sentiment strength and the index in stock markets. Tumasjan et al. [2010] predicted the 2009 German election based on political sentiment contained in Twitter. Using Facebook data, Kramer [2010] found that the sentiment contained in status updates varied with occurrences of events. Mishne and Glance [2006] utilised blogger sentiment from weblog posts appearing in Blogpulse [Glance et al., 2004] to predict movie sales.

### 2.6.2 Burst detection

As defined in Kleinberg [2003], a *burst* can be understood as an unusual growth in intensity of an observation of interest. A simple approach is to count the relative frequency and apply a threshold for detection. Balog et al. [2006] use the thresholding method to detect peak times of moods and explain the spikes by events defined

by overused words in the interval. One obvious problem with this approach is the uncertainty in determining a good threshold. Another method for learning peak times for time series is to learn the hidden states generating the observations of the time series. A popular example is the hidden Markov model [Rabiner, 1989]. A seminal work of detecting bursty periods for time series using the state-based approach is that of Kleinberg [2003], who models the generative process of messages in streams with an infinite state automaton: the higher the state, the higher the rate of generating messages. When the automaton is in high state, a burst is detected at the time interval. Originally, Kleinberg observed that certain topics in his email corpus were more easily characterised by a sudden confluence of message sending, rather than by the textual features of the messages themselves. These high-intensity periods are called bursts, and a term is in a bursty state when it grows in intensity for a period. The Kleinberg model (KLB) is a simple and efficient algorithm to localise a term's burst time.

The KLB has been used to detect events. For example, He et al. [2007a,b] use the KLB to detect bursts for unigrams in content and then to explore the co-occurrence of these burst features for event detection. Kumar et al. [2003] use the KLB to identify and rank bursts in communities by detecting bursty periods of linked structure creation among bloggers in communities. Fujiki et al. [2004] modify the Kleinberg model to impose different initial rates on daytime and nighttime periods. Gruhl et al. [2005] apply the model of detecting bursts to find correlation between on-line content (blog post mentions) spikes and customer behaviour (sales) spikes.

While working well in practice, the KLB is not suitable for large-scale and on-line data since it requires knowledge of the entire data to determine the ratio of emitting relevant documents during non-bursting periods and the transition cost. Since data in the blogosphere is normally large and streams on-line, the need for a mechanism for detecting bursty periods incrementally arises.

Further, since the vocabulary used in the blogosphere is large, the process of determining bursty periods for all terms is time consuming and impractical. However, since the number of sentiment expressions people can perceive is limited, looking for bursts for this vocabulary of sentiment, known as sentiment bursts, should be a feasible practice. From these sentiment bursts, the events associated can be detected and traced.

## 2.7 Identity in Social Media

The above sections discuss work aimed at classifying, clustering and event detection, by the manifest mood of texts. If the focus of analysis is shifted from texts to their authors, different questions arise, including: Can users be characterised by the mood of the messages they author? Are some authors more alike than are others in terms of the mood of their discussions?

As discussed in Section 2.1, social media are a promising arena for investigating *the self* that people express on-line since users and their profiles are usually accessible to the public. A number of works have explored the identity that users show in social media, with many of them considering demographic information, such as age and gender. For example, Schler et al. [2006] examined the relationship between post content and the age and gender of the authors. While the content in this study is considered under a simple bag-of-word representation, a number of richer language models have also been used in the literature for analysing the relationship between textual content of blog posts and user demographics. One example is the LIWC package, which categorises words into psycholinguistic groups. Using the LIWC utility, Pennebaker and Stone [2003] found that older people use language that is less *self-focused* and is more *present* and *future* oriented. Focusing on gender, Newman et al. [2008] demonstrate that women tend to use more *social* words and *references* to others, whereas men use more *complex* language. For sentiment-laden text, Stirman and Pennebaker [2001] and Rude et al. [2004] found that depressed and suicidal individuals are more *self-focused* (using more *first person* singular), express more *negative* emotions and occasionally use more *death-related* words. Further, Nowson [2006] found correlations between gender and personality and the manifestation of features in the LIWC.

It is also assumed that there are differences in topical interests among users in different gender or personality groups. Various claims related to *identity* in social media will be hypothesised, substantiated and empirically evaluated in this thesis.

## 2.8 Networking Aspect of Social Media

As mentioned in Section 2.1, *networking* is another key feature that a social medium must meet to attract users. For example, Livejournal allows people of common interests to establish their own spaces via *communities*.

The topic of community structure discovery in complex networks has been investigated in numerous studies. In particular, link structure, which can be explicitly declared as *friendship* or *membership* in subscribers' profiles, has been exploited. An example of a link-based approach for detecting community structure in social and biological networks is Girvan and Newman [2002]. For on-line social networks, Kumar et al. [2006] use friendship links and other information user profiles, annotated with timestamps, to learn the network structure and its evolution using Flickr, Yahoo! 360 and blogosphere data. Backstrom et al. [2006] use co-authorship and publication information in the Digital Bibliography and Library Project, and friendship and community membership in Livejournal to learn group formation in the two networks. A disadvantage of link-based approaches is that they make strong assumptions about the availability of link structure and the stability of communities, which is not always true in practice.

Another approach to learning community structure is to use the content created by users. For example, McCallum et al. [2007a] presented the Author-Recipient-Topic model to exploit the content in emails sent among Enron employees. The model can discover relevant topics, predict people's roles and give lower perplexity on previously unseen messages. They also introduced the Group-Topic model to detect groups in streams using relationships between entities and the properties of the relationships [McCallum et al., 2007b]. In Song et al. [2005], the Content-Time-Relation model was applied to the Enron email corpus to discover roles, predict the senders, infer the receivers and describe the content of information exchanged between people in the company.

In addition to the content, tags can also be exploited for the task of community detection. For instance, Negoescu et al. [2009] used the tags and membership information of Flickr users in a bag-of-words representation to learn Flickr hyper-groups. Berendt and Hanser [2007] showed that tags could potentially enrich information

related to the posts and the bloggers. The authors reveal that tags complement content, reflect differences between annotators and provide additional information. Hayes and Avesani [2007] used tag information to find topic-relevant blogs.

A joint link–content approach has also been used in studying social networks. An example of this practice is that of Nallapati and Cohen [2008], who introduce Link-PLSA-LDA to model the relationships between the citing/linking and the cited/linked documents on specific topics.

In this thesis, community detection is performed using topic-based, mood-based and psycholinguistic-based features in a comparative analysis that, to our knowledge, is the first of its kind.

## 2.9 Conclusion

In this chapter, work related to this thesis has been discussed and surveyed. The anatomy of a social media system and a representative of social media—weblogs—were described. The potential of this kind of media for use in a sentiment-based approach to data mining was explained and sentiment analysis—in terms of the aspects that have received sufficient consideration and those areas of the field still in need of examination—was reviewed.

Next, the machine learning techniques used in this thesis were reviewed. We then discussed features and feature selection methods, noting the desirability of cheap selection schemes for data on the scale of social media.

A literature survey on event and burst detection, related to the temporal dimension of social media, was provided. Existing methods for detecting events were also reviewed; in particular we highlighted how sentiment information has been used. A popular technique for detecting bursts and the need to adapt that algorithm to work for streaming data was reviewed, as was the literature on the digital presence aspect of social media.

Next we examined the egocentric view of social media, discussing user demographics

and personality. This is followed by a discussion on the networking aspect of the social media domain, and a review of work related to community detection.

# Chapter 3

## Feature Selection and Mood Classification

In the previous chapter, an argument was made for the use of sentiment as a novel and promising approach to analysing social media. In this chapter, the ‘building blocks’ for sentiment analysis will be constructed for use in the chapters developed in sequel. An emotion high-order construct termed ‘mood’ will be examined. The problem will be approached from a classification perspective, resulting in the presentation of a comprehensive study of different feature selection schemes in machine learning for mood classification in weblogs. This is complemented by an empirical analysis of the effect of different feature subsets and the way this thesis represents features of mood classification, for example binary versus counts. The thesis introduces a novel use of two feature sets proposed in psychology: the ANEW lexicon and the LIWC. We show that these features are efficient, do not require a feature selection phase and yield classification results comparable to state-of-the-art, supervised feature selection schemes.

### 3.1 Mood Classification

As mentioned in the previous chapter, mood is a state of mind, such as being happy, sad or angry. It is a complex cognitive process that has been the focus of

extensive research efforts and debate by psychologists on its nature, formulation and structure [Mauss and Robinson, 2009, Russell, 1980, 2009]. However, better scientific understanding of what constitutes a ‘mood’ has ramifications beyond psychology; for neuroscientists, it might offer insight into the functioning of the human brain; and for medical professionals working in the domain of mental health it might enable better monitoring and intervention for individuals and communities. For instance, one recent study found empirical evidence for the spread and influence of mood among friends and within cohorts [Christakis and Fowler, 2009]. This study has generated tremendous interest among sociologists.<sup>1</sup>

Research like that cited above aims to understand psychological drives and structures behind human mental states and typically does so with expensive methodologies involving questionnaires or interviews that limit the number of participants. By contrast, this work aims to classify and cluster mood based on pre-existing content generated by users, which is collected unobtrusively: a sub-problem known as *mood analysis* in sentiment analysis [Pang and Lee, 2008]. However, the goal and scope of mood analysis differs from that found in psychology literature. The goal of mood analysis is to formulate computational models for analysing users’ moods based on content created by them, such as blog posts, embedded images, personal videos, audio or the structure of their own social network. The volume of media, and the ease of collecting it, means that the size of a study can theoretically be very large (that is, millions as opposed to dozens). Therefore, this study restricts itself to a subset of social media genres: the textual content of blog posts.

Text-based mood classification and clustering, as a sub-problem of opinion and sentiment mining, have many potential applications, as identified in Pang and Lee [2008], such as automated recommendation for product websites, as a sub-component of web technology in business and government intelligence or for the collection of empirical evidence for studies in psychological and behavioural sciences. Specifically, in the blogosphere, mood classification can be used to filter search results, to ascertain the mental health of communities or to gain detailed insight into patterns of how bloggers behave and relate to one another.

---

<sup>1</sup>Results show that a person is 15 per cent more likely to be happy if directly connected to another happy person. However if that person is unhappy, then the likelihood of happiness decreases by 7 per cent. It also shows that happy and unhappy people tend to cluster themselves.



However, text-based mood analysis poses additional challenges beyond standard text categorisation and clustering. The complex cognitive processes of mood formulation make it dependent on the specific social context of the user; their idiosyncratic associations of mood and vocabulary; their syntax and style, which reflects language usage (for example, the order of linguistic components); and the specific genre of the text. In the case of the weblogs studied in this thesis, these challenges are reflected in the diverse styles of expression of the bloggers, the relatively short text length and the use of informal language, such as jargon, abbreviations and non-standard grammar. This leads us to investigate whether machine learning-based feature-selection methods for general text classification remain effective when applied to blog text.

Feature selection methods available in machine learning are often computationally expensive, relying on labelled data to learn discriminative features. However, the blogosphere is vast (reaching almost 130 million users<sup>2</sup>) and continuing to grow, making it desirable to construct a feature set that works without requiring a supervised feature selection stage to classify mood. To this end, it is necessary to look to the results of studies that intersect psychology and linguistics. Doing so reveals two potentially useful systems: (1) the sentiment-bearing lexicon known as ANEW [Bradley and Lang, 1999] and (2) psycholinguistic features drawn from the LIWC [Pennebaker et al., 2007b]. It is proposed that these systems are suitable for use in mood classification.

The contribution of this chapter is twofold. First, it provides a comparative study of machine learning-based text feature selection for the specific problem of mood classification, elucidating insights into what can be transferred from a generic text-categorisation problem for mood classification. Second, it formulates a novel use of two psychology-inspired sets of features for mood classification that do not require supervised feature learning and are thus useful for large-scale mood classification.

The rest of this chapter is organised as follows. The machine learning-based feature selection schemes, together with the proposed ANEW and LIWC feature sets, are presented in Section 3.2. Section 3.3 describes three Livejournal datasets that are used for mood classification experiments in this chapter. Section 3.4 presents the results of mood classification in the three datasets, followed by some concluding

---

<sup>2</sup>From the state of the blogosphere 2008 at <http://technorati.com>.

remarks.

## 3.2 Feature Selection for Mood Classification

Denote by  $\mathcal{B}$  the corpus of all blog posts and by  $\mathcal{M} = \{\text{sad, happy, ...}\}$  the set of all mood categories. In a standard feature selection setting, each blog post  $d \in \mathcal{B}$  is also labelled with a mood category  $c_d \in \mathcal{M}$  and the objective is to extract from  $d$  a feature vector  $\mathbf{x}^{(d)}$ , being as discriminative as possible, for  $d$  to be classified as  $c_d$ . For example, if we further denote by  $\mathcal{V} = \{v_1, \dots, v_{|\mathcal{V}|}\}$  the set of all terms, then the feature vector  $\mathbf{x}^{(d)} = [\dots, \mathbf{x}_i^{(d)}, \dots]$  might take a simple counting with its  $i$ -component  $\mathbf{x}_i^{(d)}$  representing the number of times the term  $v_i$  appears in document  $d$ , a scheme widely known as bag-of-word representation.

The problem of generic text document classification has been investigated extensively in the text-mining domain. It is generally agreed among researchers that, despite the strong independence assumption among features conditional on the class, the simple NB classifier remains the state-of-the-art for this task: a fact that is verified by the experimental results for the problem of mood classification presented in this chapter. However, different feature selection methods are found to influence classification performance greatly [Yang and Pedersen, 1997]. Taking the view that the work of Sebastiani [2002], Yang and Pedersen [1997] represents most widely used feature selections for generic text categorisation tasks, we question whether the findings in the work hold for our mood classification problem. We shall briefly describe commonly used feature selection schemes, including those in Sebastiani [2002], Yang and Pedersen [1997].

### 3.2.1 Term-based selection

These are features derived with respect to a term  $v$ . Two common features are TF and DF, where  $TF(v, d)$  represents the number of times the term  $v$  appears in document  $d$ , and  $DF(v)$  the number of blog posts containing the term  $v$ . It is also well known in text mining that the  $TF.IDF(v, d)$  weighting scheme can

potentially improve discriminative power where  $TF.IDF(v, d) = TF(v, d) \times IDF(v)$  with  $IDF(v) = |\mathcal{B}|/DF(v)$ , which is the inverse document frequency. In this work, depending on the scheme, a term  $v$  will be selected if it has a high  $DF(v)$  value or high average values of  $TF(v, d)$  or  $TF.IDF(v, d)$  across all documents  $d$  over a threshold.

### 3.2.2 Term-class interaction-based selection

The essence of these methods is to capture the dependence between terms and corresponding class labels during the feature selection process, thereby capturing the relevance of terms for classification tasks at the corpus level. Three common selection methods falling into this category are  $IG(v)$ ,  $MI(v, c)$  and  $CHI(v, c)$  [Yang and Pedersen, 1997].  $IG(v)$  captures the information gain (measured in bits) when a term  $v$  is present or absent;  $MI(v, c)$  measures the mutual information between a term  $v$  and a class label  $c$ ; and  $CHI(v, c)$  measures the dependence between a term  $v$  and a class label  $c$  by comparison with one degree of freedom  $\chi^2$  distribution. Denote  $A$  the number of times term  $v$  and class  $c_i$  appear together in a blog post in the corpus,  $B$  for the times term  $v$  is present without class  $c_i$ ,  $C$  for the times class  $c_i$  occurs without term  $v$ ,  $D$  for the times both term  $v$  and class  $c_i$  are absent from the corpus and  $N$  is the number of data points ( $N = A + B + C + D$ ). Three term-class interaction-based selection methods are defined below.

- Information Gain

$$\begin{aligned}
 IG(v) &= - \sum_{i=1}^{|\mathcal{M}|} Pr(c_i) \log Pr(c_i) \\
 &\quad + Pr(v) \sum_{i=1}^{|\mathcal{M}|} Pr(c_i|v) \log Pr(c_i|v) \\
 &\quad + Pr(\bar{v}) \sum_{i=1}^{|\mathcal{M}|} Pr(c_i|\bar{v}) \log Pr(c_i|\bar{v})
 \end{aligned} \tag{3.1}$$

where  $Pr(c_i) = \frac{A+C}{N}$  is the probability of class  $c_i$ ,  $Pr(v) = \frac{A+B}{N}$  and  $Pr(\bar{v}) = \frac{C+D}{N}$  are the probabilities for term  $v$  found and not found in blog posts respectively, and

$Pr(c_i|v) = \frac{A}{A+B}$  and  $Pr(c_i|\bar{v}) = \frac{C}{C+D}$  are the conditional probabilities of class  $c_i$  given  $v$  and  $\bar{v}$  respectively.

- Mutual Information

The MI criterion for a term  $v$  and a class  $c$  is defined as

$$MI(v, c) = \log \frac{Pr(v \wedge c)}{Pr(v)Pr(c)} \quad (3.2)$$

where  $Pr(v \wedge c) = \frac{A}{N}$  is the probability that  $v$  and  $c$  are present together in blog posts.

We compute  $MI$  for term  $v$  by

$$MI_{avg}(v) = \sum_{i=1}^{|\mathcal{M}|} Pr(c_i)MI(v, c_i) \quad (3.3)$$

or

$$MI_{max}(v) = \operatorname{argmax}_{c \in \mathcal{M}} \{MI(v, c)\} \quad (3.4)$$

In practice, both these *avg* and *max* options are used to compute the MI and the better result is reported.

- CHI-square ( $\chi^2$ ) statistic

The  $\chi^2$  statistic between a term  $v$  and a class  $c$  is defined as

$$\chi^2(v, c) = \frac{|\mathcal{B}| \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (3.5)$$

We compute the  $\chi^2$  statistic for  $v$  by

$$\chi_{avg}^2(v) = \sum_{i=1}^{|\mathcal{M}|} Pr(c_i) \chi^2(v, c_i) \quad (3.6)$$

or

$$\chi_{max}^2(v) = \operatorname{argmax}_{c \in \mathcal{M}} \{\chi^2(v, c)\} \quad (3.7)$$

In practice, as in computing the MI, both formulas are used to compute the CHI-square statistics and the better result is reported.

## 3.3 Datasets

In this thesis we use data crawled from Livejournal, a weblog hosting site. Livejournal allows people to tag their *mood* when they are blogging, thus providing an excellent source of ground truth data for sentiment analysis. The host provides a comprehensive set of 132 moods for users to specify their current emotion at the time of blogging. The provided moods range diversely in the emotion spectrum; for example, *cheerful* and *grateful* for *happiness*, or *discontent* and *uncomfortable* for *sadness*. A sample post containing the current mood tag is shown in Figure 3.1. For mood classification, we use three Livejournal datasets described below. Table 3.1 shows the top 10 moods in these datasets based on their frequency counts.

### 3.3.1 The IR05 dataset

The IR05 dataset, introduced by Mishne [2005], contains 815,494 blog posts from Livejournal. This dataset can be considered the first Livejournal corpus created for the purpose of mood classification. Mishne [2005] performs emotion analysis on this blog post corpus. He uses a number of feature sets such as frequency counts, lengths,

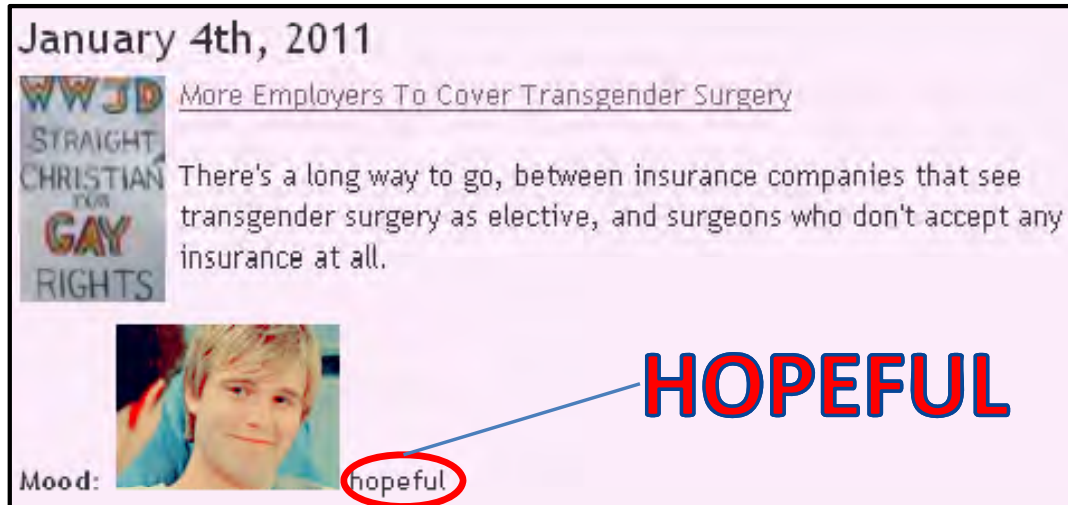


Figure 3.1: An example of blog posts in Livejournal with the ‘current mood’ tag (*hopeful* in this case).

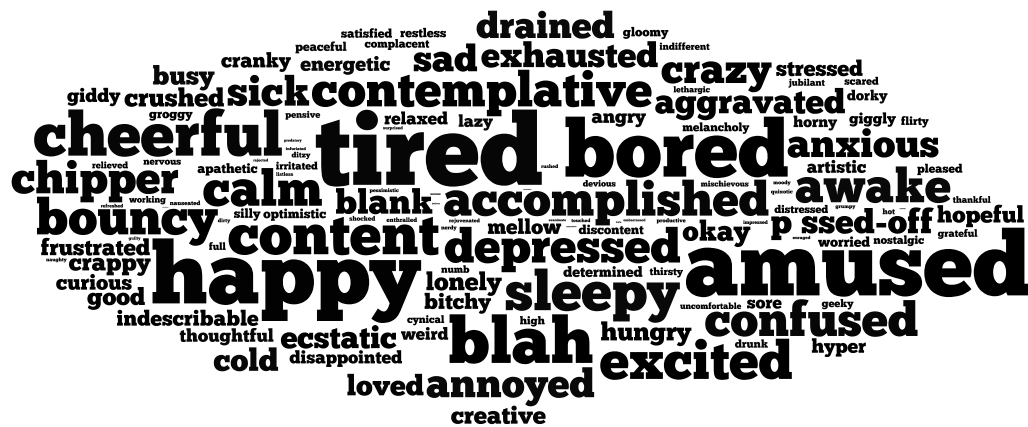


Figure 3.2: Cloud visualisation of 132 predefined mood labels tagged in the CHI06 dataset.

IR05		CHI06		WSM09	
Moods	Occurrences	Moods	Occurrences	Moods	Occurrences
amused	24,686 (4.6%)	tired	671,427 (3.7%)	amused	26,781 (4.8%)
tired	20,210 (3.8%)	amused	604,946 (3.4%)	tired	23,344 (4.1%)
happy	16,407 (3.1%)	happy	579,518 (3.2%)	cheerful	18,767 (3.3%)
cheerful	12,939 (2.4%)	bored	536,098 (3.0%)	accomplished	16,340 (2.9%)
bored	12,718 (2.4%)	blah	434,459 (2.4%)	happy	15,640 (2.8%)
accomplished	12,138 (2.3%)	cheerful	412,265 (2.3%)	calm	15,003 (2.7%)
sleepy	11,505 (2.1%)	content	384,412 (2.1%)	excited	13,495 (2.4%)
content	11,149 (2.1%)	sleepy	371,713 (2.1%)	contemplative	13,041 (2.3%)
excited	11,045 (2.1%)	excited	358,905 (2.0%)	busy	12,890 (2.3%)
contemplative	10,705 (2.0%)	calm	336,654 (1.9%)	sleepy	12,741 (2.3%)

Table 3.1: Top 10 moods in the datasets.

sentiment orientations, emphasised words and special symbols as input to an SVM classifier. The classification accuracy is modest, being slightly above baseline. This dataset is also used in Keshtkar and Inkpen [2009], in which the authors introduce an approach using the hierarchy of possible moods, achieving better results than flat classifications.

Of the total, 535,844 posts are tagged with predefined moods. We disregard the posts annotated with non-predefined moods.

### 3.3.2 The CHI06 dataset

The CHI06 dataset was introduced by Leshed and Kaye [2006]. This dataset contains approximately 18 million blog posts written in English by about 1.6 million bloggers and tagged with the predefined moods. These journals range in time from 1 May 2001 to 23 April 2005. A cloud visualisation of moods tagged in this dataset is shown in Figure 3.2.

Leshed and Kaye [2006] perform emotion classification on 50 of the most frequent moods appearing in the corpus. They use the TF-IDF feature selection method to select only the first 5,000 features. Blog entries are represented in the ‘bag-of-word’ model of information retrieval and subjected to an SVM classifier. The average accuracy of the system is reported to be 78 per cent [Leshed and Kaye, 2006].

### 3.3.3 The WSM09 dataset

The WSM09 dataset was provided by Spinn3r (spinn3r.com) as the benchmark dataset for the ICWSM 2009 conference.<sup>3</sup> It contains 44 million blog posts crawled between August and October 2008. A subset from this dataset was extracted for this thesis consisting of only blog posts from Livejournal that include the mood ground truth entered by the user when the post was composed. Again, only the moods predefined by Livejournal are considered and all others discarded, resulting in approximately 600,000 blog posts.<sup>4</sup>

The WSM09 dataset is used in Sood and Vasserman [2009] for a two-step data-mining application. At first, all moods are categorised into three classes: happy, sad and angry, using K-mean clustering. Blogs posts in these three groups are then subjected to a NB classifier. The feature set considered in this task consists of unigrams, bigrams, stems, emotion, emoticons and slang. The highest recall, precision and F-measure are 67.1 per cent, 65 per cent and 66.1 per cent respectively.

## 3.4 Classification Results

Three term weighting-based (TF, DF, TF-IDF) and three term-class interaction-based (IG, MI, CHI) selection methods are employed in this experiment. These feature selection methods will be applied on all terms (unigrams) or with respect to a subset of terms tagged with a specific part of speech (for example, only verbs). For comparison, the number of features in all cases, except for LIWC, equals 1,034, which is the number of ANEW words. For the LIWC case, all 68 categories returned from the LIWC package are used as features.

The purpose of the experimental design described here is to compare and evaluate which feature selection methods work best in terms of both performance and computational efficiency, and to examine the effect of specific linguistic components in the context of mood classification. For the classification method we experimented with many off-the-shelf classifiers implemented in the Weka package [Hall et al.,

---

<sup>3</sup><http://www.icwsm.org/2009/data/>, retrieved November 2011.

<sup>4</sup>Consistent with what is reported in Sood and Vasserman [2009].



2009] such as NBC, SVM, IBK and C4.5 and reported the best results the proposed feature sets can achieve among the classifiers. The NB classifier consistently outperforms comparative methods in all feature sets, except for LIWC, and therefore the results for these features with respect to NBC shall be reported.

For LIWC features, the SVM classifier outperforms the others, including the NBC. This may be because certain LIWC groups consist of others. For example, *affect* (affective processes) consists of *negemo* (negative emotion) and *posemo* (positive emotion), resulting in high correlations among these LIWC groups. This obviously violates the assumption of independence among features by NBC, reducing the performance of the probabilistic classifiers. Therefore, for LIWC features, we shall report the results with respect to SVM. The Weka implementation of SMO [Hall et al., 2009] is used for classification using LIWC features.

For each run we use 10-fold cross-validation, repeat 10 runs and report the average result. To evaluate the results we report two commonly used measures: accuracy and F-score (which is measured based on recall and precision, see Section 2.4.1.4).

### 3.4.1 Effect of feature selection schemes and linguistic components.

We use three datasets: IR05, CHI06 and WSM09 for the task of mood classification. For comparison with previous results in Sood and Vasserman [2009], we examine three popular moods {sad, happy, angry} in this experiment.

We run the experiment over all possible combinations of the six feature selection methods (TF, DF, TF-IDF, IG, MI and CHI) on five different linguistic subsets (all terms, adjectives, verbs, adverbs and a combination of these three parts of speech). Mood classifications are also performed with ANEW and LIWC as feature sets. The top results are reported in Table 3.2.

With respect to feature selection schemes, IG is observed to be the best selection scheme. Other term-class interaction-based methods do not perform as well. Noticeably, MI does not appear in any of the top results. These observations are consistent with the findings of Yang and Pedersen [1997] for the text-categorisation problem.

However, in contrast to the conclusions in Yang and Pedersen [1997], we found that CHI performs badly for mood-classification tasks and does not appear in any of the top results. Surprisingly, both TF and DF perform better than TF-IDF in all-term (unigram) cases, challenging the prevailing perception of these schemes in text mining, which holds that IF-IDF is often superior, although much more computationally expensive. Therefore, TF or DF should be the substitution candidates for IG for the trade-off of computational cost.

The performance of the selected feature selection schemes is also acceptable across the three datasets, as can be seen in Table 3.2. The best results stand at 78.8 per cent F-score for IR05, 74.7 per cent for CHI06 and 77.4 per cent F-score for WSM09, which is higher than that reported in Sood and Vasserman [2009] (66.1 per cent).

With respect to the effect of linguistic components (which are not tested in Sood and Vasserman [2009] or Leshed and Kaye [2006]), a combination of adjectives, verbs and adverbs (AjVbAv) dominates the top results and gives performance very close to that achieved when using all terms. Using verbs or adjectives alone also produces a good performance.

### 3.4.2 Performance of LIWC

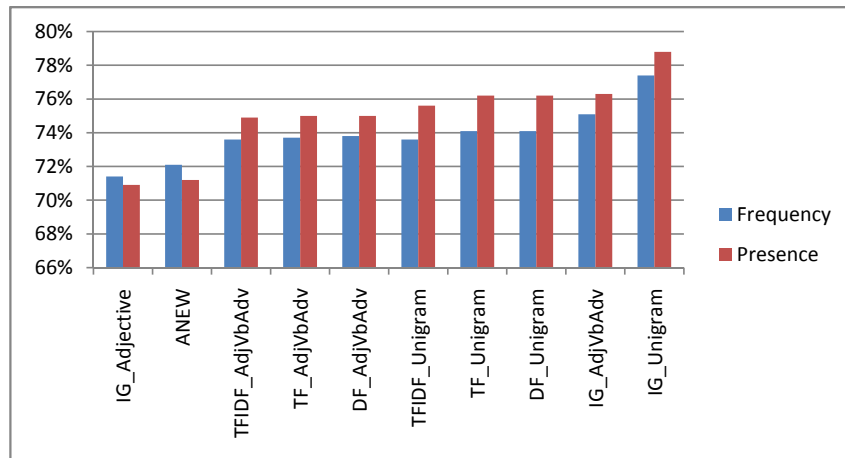
In comparison with more than 30 combinations of selection methods and feature spaces, though without the need for a supervised feature selection stage, the result of LIWC features is found to be very encouraging, appearing among the top results across the three datasets. This result reveals differences in the use of the psycholinguistic features (categorised by the LIWC package) among posts tagged with different moods.

To examine the difference, a subset of the CHI06 dataset is sampled, consisting of 10,000 posts tagged with current mood *happy* (valence value=8.21) and 10,000 posts tagged with current mood *sad* (valence value=1.61). It was found that all LIWC features are significantly different between *happy* and *sad* posts (Mann-Whitney U tests [Mann and Whitney, 1947],  $ps < .05$  two-tailed,  $n_1 = n_2 = 10,000$ ), barring five exceptions, *percept* ( $p < .717$ ), *space* ( $p < .693$ ), *number* ( $p < .584$ ), *relig* ( $p < .516$ ) and *humans* ( $p < .158$ ). Figure 3.4 graphs the difference of LIWC categories between

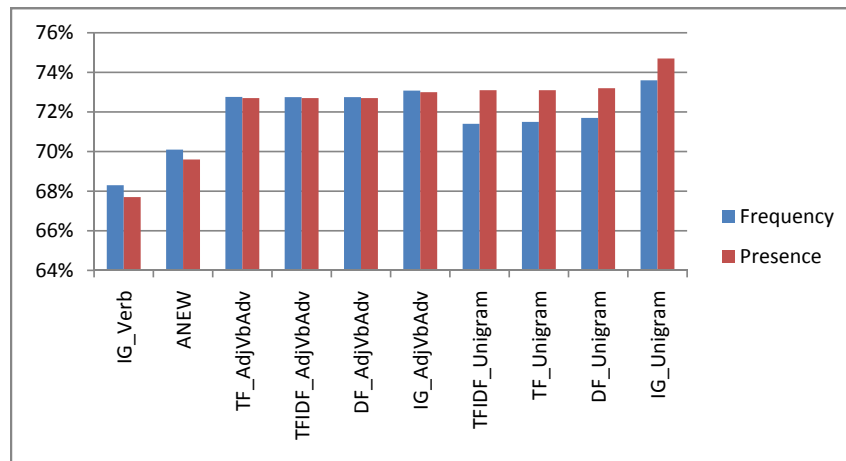
IR05				CHI06			
Selection method	Linguistic subsets	Accuracy	F-score	Selection method	Linguistic subsets	Accuracy	F-score
LIWC		0.733	0.692	LIWC		0.678	0.677
IG	Adjective	0.738	0.709	IG	Verb	0.695	0.677
ANEW		0.734	0.712	ANEW		0.715	0.696
TF.IDF	AjVbAv	0.759	0.749	TF	AjVbAv	0.735	0.727
TF	AjVbAv	0.76	0.75	TF.IDF	AjVbAv	0.736	0.727
DF	AjVbAv	0.76	0.75	DF	AjVbAv	0.736	0.727
TF.IDF	unigram	0.765	0.756	IG	AjVbAv	0.739	0.73
DF	unigram	0.765	0.762	TF.IDF	unigram	0.732	0.731
TF	unigram	0.765	0.762	TF	unigram	0.733	0.731
IG	AjVbAv	0.773	0.763	DF	unigram	0.734	0.732
IG	unigram	0.791	<b>0.788</b>	IG	unigram	0.748	<b>0.747</b>

WSM09			
Selection method	Linguistic subsets	Accuracy	F-score
ANEW		0.713	0.697
IG	Verb	0.714	0.7
LIWC		0.744	0.73
TF-IDF	unigram	0.744	0.738
DF	AjVbAv	0.75	0.745
TF	AjVbAv	0.751	0.745
TF.IDF	AjVbAv	0.754	0.748
DF	unigram	0.753	0.752
TF	unigram	0.753	0.752
IG	AjVbAv	0.762	0.756
IG	unigram	0.776	<b>0.774</b>

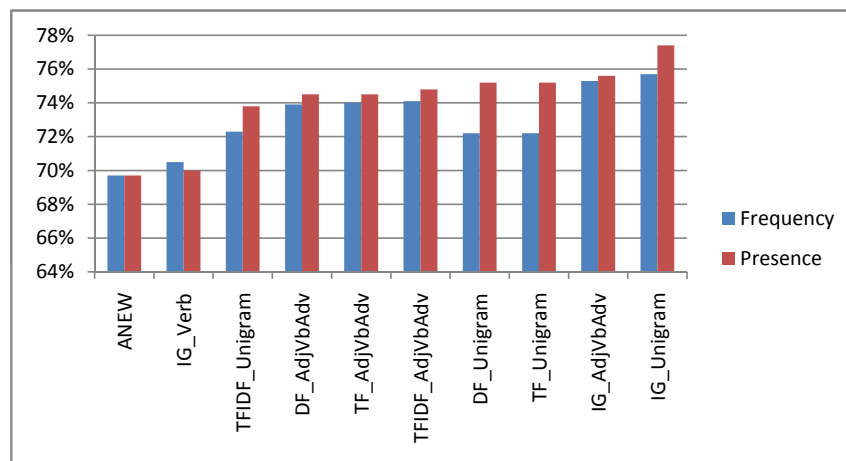
Table 3.2: Mood classification results for different feature selection schemes and for different feature subsets. Different combinations of selection methods and feature spaces are run, but we report only the top results sorted in ascending order of F-score.



(a) IR05 database.



(b) CHI06 database.



(c) WSM09 database.

Figure 3.3: Performances of binary versus counting features for mood classification in the top best results. The performance is measured in F-score and sorted in increasing order of performance.

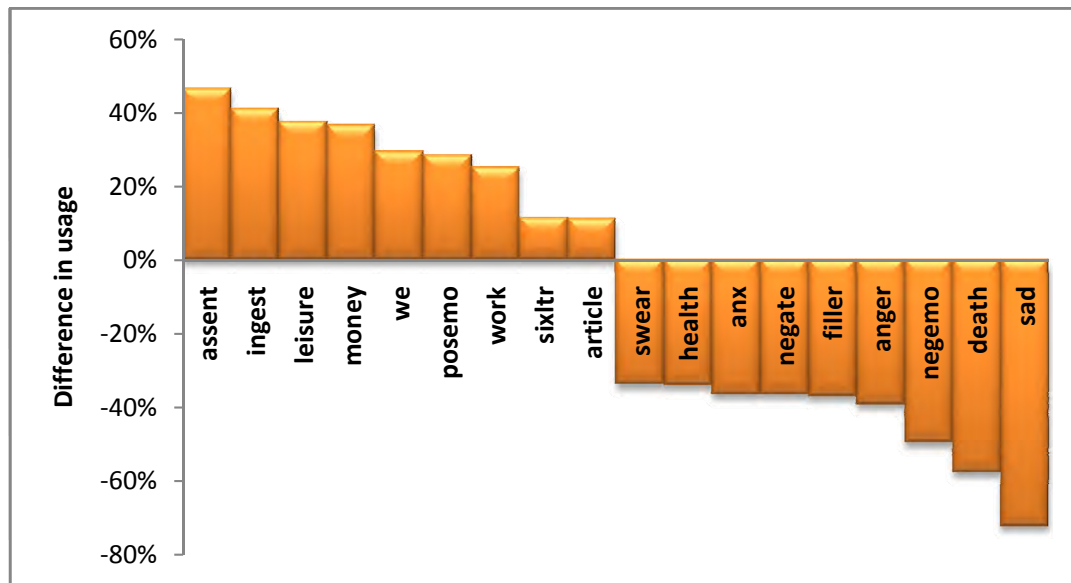


Figure 3.4: The LIWC features above the zero line are in favour of *happy*; otherwise, they are in favour of *sad* ( $ps < 0.001$ ).

*happy* and *sad* posts. Not surprisingly, *posemo* and *leisure* words are used more often in *happy* posts, while *negemo*, *sad* and *death* words dominate *sad* ones.

### 3.4.3 Performance of ANEW

Similar to LIWC, in not needing a supervised feature selection stage, the results of ANEW features are also found to appear among the top results across the three datasets. The results across the three datasets are also consistent, standing at approximately 70 per cent F-score. On the benchmark dataset, WSM09, the result from ANEW words as features are better than the best result reported in Sood and Vasserman [2009] (66.1 per cent).

Therefore, it is argued that there is a difference in the use of ANEW words among posts in different moods. Figure 3.5 shows an example of this difference. As can be seen, the post tagged with a high-valence mood (*happy*) uses ANEW words in high valence, as opposed to the post tagged with a low valence mood (*sad*).

Thank you!  
 Jou and Seto, thank you for the flowers! They're absolutely gorgeous.  
 John, thank you for the flowers and the gift certificate. The flowers  
 are lovely, and I'm going to have a lot of fun trying to decide what to  
 buy.  
 Everyone else who wished me a happy birthday, thanks so much! :D  
 Sweetheart, I know you aren't going to see this until you get home  
 tomorrow, but thank you for making my birthday so wonderful  
 tonight. I had a great time out at Golden Gate park and I really liked  
 the restaurant, and the earrings and necklace that you gave me are so  
 beautiful. I love them. \*kiss\*

(a) ANEW words used in a post tagged with mood *happy*. The average valence for this post is 7.75.

Co-worker is DEAD  
 So last year a co-worker of mine never came back to work. Rumor had it he  
 was deported due to his visa running out. Well that is only part of the story.  
 Thanks to the current state of fear in our country, human rights abuse  
 and hatred for immigrants my co-worker died in detention awaiting  
 his appeal. The cancer got him but his broken spine was not due to cancer.  
 I am really angry to say the least. Sad and upset most definately.  
 There is alot I could say but it will not change the world I live in or  
 the country I call home.  
 I just ask that others spread the story. I don't want my co-worker to  
 have died and no one hear about his story.  
 R.I.P. Hiu Lui "Jason" Ng

(b) ANEW words used in a post tagged with mood *sad*. The average valence for this post is 3.12.

Figure 3.5: Examples of the use of ANEW words in happy and sad blog posts. The colour reflects the valence and arousal values an ANEW word conveys, as shown in Figure 2.3.

### 3.4.4 Binary versus counting features

It is well known in text mining that the bag-of-word counting representation (that is, count the number of appearances of a term) is an effective feature. However, in sentiment analysis, it has been found that a simple binary representation (that is, use 1 if the term appears in the document and 0 otherwise) is more effective at movie review classification [Pang et al., 2002]. Further, a binary feature representation can be better compressed, making it suitable for dealing with large datasets. Therefore, we are motivated to investigate whether a binary representation is effective for mood classification.

Figure 3.3 shows the results for the IR05, CHI06 and WSM09 datasets when the classification was performed on the binary and counting features respectively. For each dataset, the classification F-score is plotted for both types of representation, displaying the top results in an increasing order of performance. The results reveal that binary representation outperforms its counterpart for four top performances in all datasets. This result again confirms the superiority of binary representation over counting for mood classification, suggesting that recording the appearance of a term in a document is sufficient and recommended for mood classification tasks.

## 3.5 Conclusion

In seeking to create tools for analysing content in social media under the impact of users' moods, we have investigated the problem of mood classification in weblogs. While the problem of machine learning-based feature selection for text categorisation has been intensively explored, little work has been done on textual-based mood classification, which is often more challenging. This chapter's contribution is a comprehensive comparison of different feature selection schemes in combination with a range of linguistic subsets as feature spaces across three large datasets, elucidating insights into what can be transferred from a generic text-categorisation problem to mood classification.

In addition, a novel use of a new set of psychology-inspired features (ANEW) and psycholinguistic features (LIWC) is proposed that does not require a supervised

---

feature selection phase and can therefore be applied for mood analysis at a much larger scale. The results support similar findings in previous research, but have also brought to light discoveries particular to the problem of mood classification. The newly proposed feature sets have also performed comparatively well at a fraction of the computational cost of supervised schemes. These findings will serve as the base for subsequent sentiment-related analysis in the following chapters.



## Chapter 4

# Mood Patterns and Affective Lexicon Access in Weblogs

The preceding chapter addressed the problem of sentiment classification in social media, using sentiment in the form of *mood*. It was proposed to use a psychology-inspired set of features—ANEW—for the classification, and it was demonstrated that the lexicon has the advantage of being computationally efficient while maintaining accuracy comparable to other state-of-the-art feature sets.

This chapter explores another important problem of pattern discovery in social media: sentiment clustering. Since sentiment classification requires class labels for training examples, clustering is suitable on data without sentiment annotations. It investigates the possibility of unearthing intrinsic patterns of mood using unsupervised approaches. The ANEW feature set is used in unsupervised learning to cluster 18 million blog posts, providing a unique view of mood patterns in the blogosphere, in which each group is a set of synonymic moods. These manifest global patterns of mood organisation are found analogous to the pleasure–displeasure and activation–deactivation dimensions of the core affect model for the structure of human emotion proposed independently in the psychology literature. In addition, discovery of mood–lexicon correlation in the blogosphere indicates the predictive power of the affective lexicon for mood prediction in social media.

The contribution of this chapter is twofold. First, it provides empirical results for

mood organisation in the blogosphere on the largest dataset with mood ground truth available today. To our knowledge, it is the first to consider the problem of data-driven mood pattern discovery at this scale. Second, it explores a novel problem of detecting the mood–affective lexicon usage correlation in the new media and proposes a novel use of a term-goodness criterion to discover this sentiment–linguistic association.

## 4.1 Data-driven Discovery of Mood Patterns

### 4.1.1 Basic emotions

The basic emotion theory in psychology assumes that our emotions vary but stem from basic ones. For example, the Big Six, proposed by Ekman and Friesen [1971], specifies six basic emotions: happiness, anger, fear, sadness, disgust and surprise. Existing work in sentiment analysis in social media is often limited to finding the sentiment sign in the dipole pattern for a given text, after which it is classified into one of two sentiment polarities: for example, negative versus positive or ‘thumbs up’ versus ‘thumbs down’. Extensions to this task include the three-class classification (adding neutral to the polarity) and locating the value of the emotion carried by the text across a spectrum of valence scores. However, the number of sentiment expressions is not limited to that quantity. For example, in the dataset CHI06, described in Section 3.3.2, all 132 predefined moods are used to express authors’ feelings in 18 million blog posts, from the least *intimidated* (7,029 times, 0.04 per cent) to the most popular *tired* (671,427 times, 3.75 per cent).

This chapter describes a study associated with mood patterns. It departs from a data-driven approach to discover ‘emotion patterns’ automatically. Each of the clustered mood groups can be considered as a set of moods originating from a basic emotion. It is inspired by the question as to whether these structures can be discovered directly from data, without the cost of involving human participants, as in traditional psychological studies. Next, it aims to study the relationship between the data-driven emotion structures discovered and those proposed by psychologists.

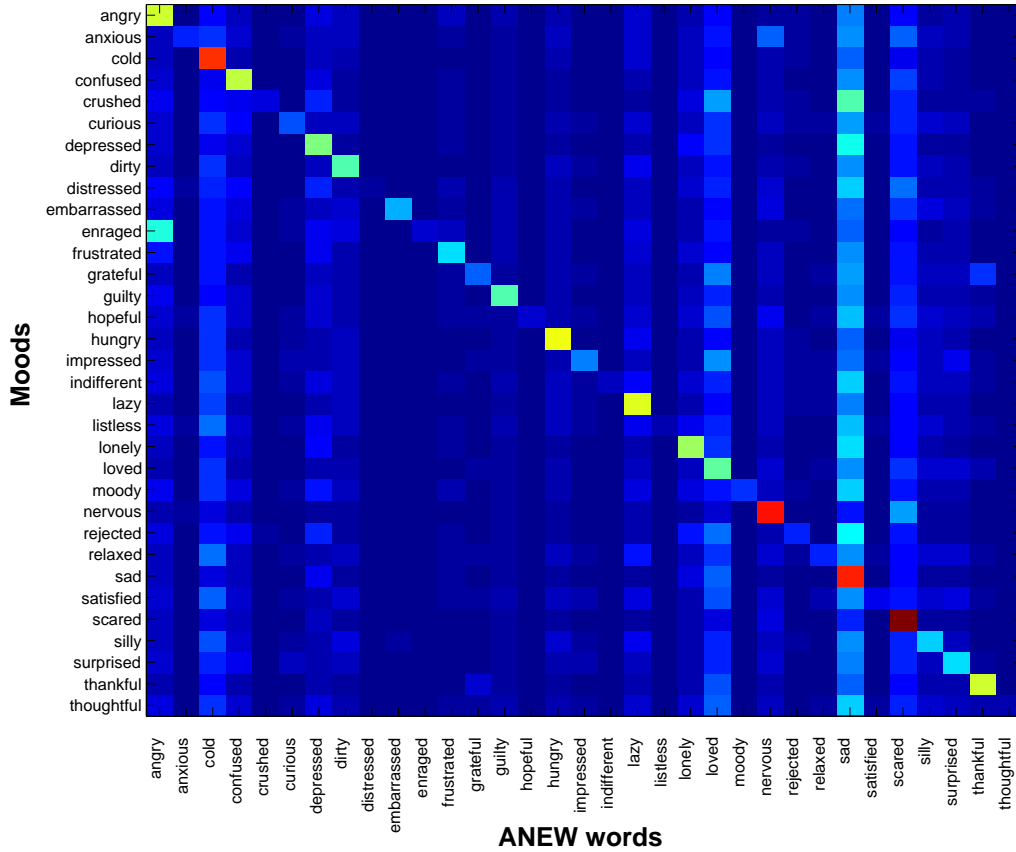
### 4.1.2 Proposed framework

Livejournal provides a comprehensive set of 132 moods for users to tag their moods when blogging. The provided moods cover a broad emotion spectrum, but they are typically observed to fall into soft clusters such as happiness (*cheerful* or *grateful*) or sadness (*discontent* or *uncomfortable*). We call each cluster of these moods an *emotion pattern* and aim to detect them automatically. We take, as a promising basis for this analysis, the ANEW feature vectors used in Chapter 3, which gave classification performances comparable with those employing the expensive feature-selection schemes.

Let us denote by  $\mathcal{B}$  the corpus of all blog posts and by  $\mathcal{M} = \{sad, happy, \dots\}$  the predefined set of moods ( $|\mathcal{M}| = 132$ ). Each blog post  $d \in \mathcal{B}$  in the corpus is labelled with a mood  $l_d \in \mathcal{M}$ . Denote by  $n$  the number of ANEW words ( $n = 1034$ ). Let  $\mathbf{x}^{(m)} = [\mathbf{x}_1^m, \dots, \mathbf{x}_i^m, \dots, \mathbf{x}_n^m]$  be the vector representing the usage of ANEW words by mood  $m$ . Thus,  $\mathbf{x}_i^m = \sum_{d \in \mathcal{B}, l_d = m} c_{id}$ , where  $c_{id}$  is the counting of the ANEW item  $i$ -th occurrence in the blog post  $d$  tagged with the mood  $m$ . The usage vector is normalised so that  $\sum_{i=1}^n \mathbf{x}_i^m = 1$  for all  $m \in \mathcal{M}$ . A part of this  $m \times n$  matrix is illustrated in Figure 4.1a. As shown, generally, the same words with the tagged moods are often used in the blog posts. An example of this representation is plotted in Figure 4.1b, which shows a cloud of ANEW words used in the blog posts tagged with mood *happy*. As can be seen from the figure, more high valence ANEW words, for example *good*, *love*, *time*, *fun* and *happy*, are used in the *happy* blog posts.

For the Livejournal corpus, we make the initial observation that blog posts tagged with moods from the same emotion pattern use similar proportions of words from the ANEW lexicon. For example, in Figure 4.2, it can be seen that the ANEW usage patterns of *happy/cheerful* and *angry/p\*ssed off* are well separated. *Hate*, *hell* and *stupid* are most likely to be found in the *angry/p\*ssed off* tagged posts and least likely to be found in the *happy/cheerful* ones. In contrast, words such as *good*, *happy*, *fun* and *love* are less commonly used in the posts tagged with *angry/p\*ssed off* than in the *happy/cheerful* ones, suggesting that the similarity between ANEW usage patterns can be used as a foundation for the study of the structure of mood space.

To discover the grouping of the moods based on the usage vectors, we use an arti-



(a) A part of the matrix of moods  $\times$  ANEWs.



(b) A plot of the ANEW usage vector for mood *happy*.

Figure 4.1: The ANEW word usage in the blog posts tagged with different moods.

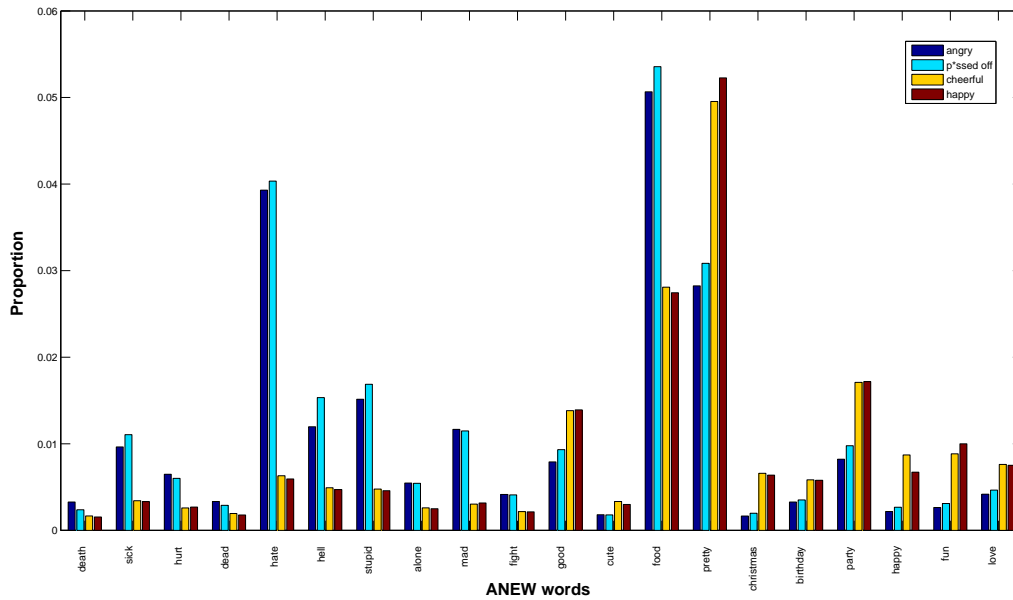


Figure 4.2: The ANEW word proportion in two groups of moods:  $\{angry, p^*ssed\ off\}$  and  $\{happy, cheerful\}$ . The valence of ANEW words is increasing on the horizontal axis.

ficial neural network, SOM [Kohonen, 1990], and a non-parametric algorithm, AP [Frey and Dueck, 2007]. The algorithms only require the pairwise similarities between moods, which we compute based on the Euclidean distances for simplicity (see Section 2.4.2).

To map the emotion patterns detected to their psychological meaning, we measure the sentiment scores of those  $|\mathcal{M}|$  mood words. In this research, we use the sentiment scores provided by the ANEW lexicon. Specifically, the valence and arousal of moods are assigned as those of the same words in the ANEW lexicon. For those moods not in ANEW, their values are assigned according to the nearest ‘parent’ words in the mood hierarchical tree,<sup>1</sup> in which those moods conveying the same or similar meaning are on the same branch of the tree. Examples of this tree are shown in Figure A.1 (see Appendix A.1). Therefore, each member of a mood cluster can be placed onto a 2D representation according to its valence and arousal dimensions, making it possible to compare them with the core affect model theorised by psychologists (see Section 2.3.3).

<sup>1</sup><http://www.livejournal.com/moodlist.bml>, retrieved November 2011.

### 4.1.3 Self-organised mood patterns

We use the CHI06 database, described in Chapter 3, for the experiments. The SOM algorithm [Kohonen, 1990] was chosen for its efficiency in computation, dimensionality reduction and ability to preserve the distance in the lower dimension effectively for visualising the structure of clusters. We use the SOM-PAK package [Kohonen et al., 1996] and the SOM Toolbox for Matlab [Vesanto et al., 2000] to train and visualise the map. For training, a  $9 \times 7$  map is used, which accounts for nearly half of the mood classes. Using the recommended parameter settings in Kohonen [1990], the horizontal axis is roughly 1.3 that of the vertical axis; node topology is hexagonal and the number of training steps is 32,000 (about 500 times the number of nodes). We omit coarse-level results and present the structures of the clusters discovered in Figure 4.3, in which the top six moods in each cluster are included.

Several interesting patterns emerge from this analysis. At the highest level, one can observe the general transition of mood from an extreme of *pleasure* (clusters II, III and V) to *displeasure* (clusters IV, VI and VII). On the *pleasure* pole, we observe moods that have very high-valence values<sup>2</sup> such as *good* (valence: 7.47), *loved* (8.64) or *relaxed* (7). Conversely, on the *displeasure* pole, we observe low valence values such as *enraged* (2.46) or *stressed* (2.33). Certain mood transitions are also evident, for example the cluster path IV-II-III represents a transitional pattern from *infuriated* to *relaxed* and then to *good*. Though not emerging as strongly as the case of *pleasure* to *displeasure*, a global pattern of *activation* to *deactivation* is also observed, based on the analysis of arousal measures as shown in Figure 4.3. Our results support the core affect model for human emotion structure studied in psychology and generally agree with the global mood structures also proposed by Russell [1980, 2009].

### 4.1.4 Affinity propagation mood patterns

The SOM algorithm [Kohonen, 1990] is useful for visualising clusters. However, it requires the map topology as well as the size to be specified in advance. Therefore, it motivates us to use the AP algorithm [Frey and Dueck, 2007] to discover the

---

<sup>2</sup>Measured based on a study on ANEW reported in Bradley and Lang [1999].

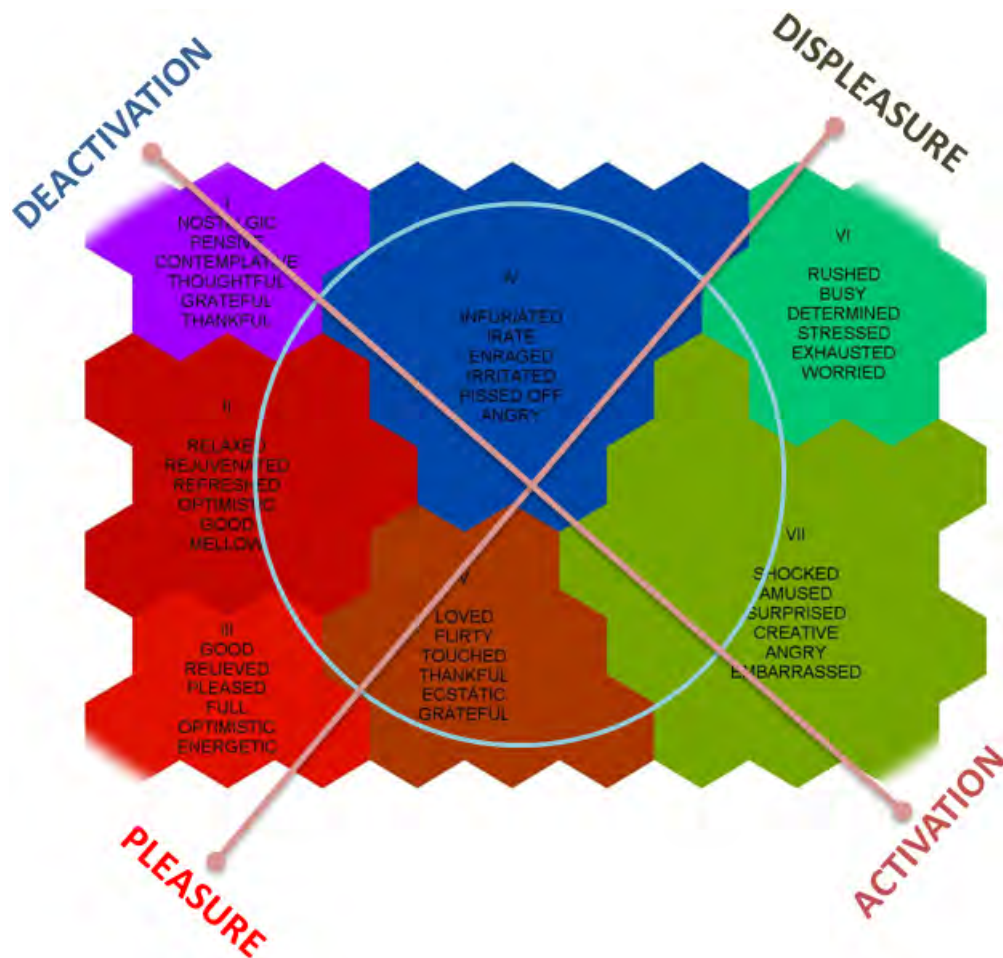


Figure 4.3: Discovered mood structure map. Each cluster is annotated with the top six mood categories.

number of clusters and their exemplars automatically.

For this study, the same dataset as in Section 4.1.3 (CH106) is used. After running the AP algorithm, 16 clusters were detected. Table 4.1 lists the discovered clusters, with exemplar moods in caps and the remaining members in lower case. Clusters 1–7 typically contain moods of high valence or pleasure, while clusters 8–16 contain moods of low valence or displeasure.

To visualise the mood patterns on the affect circle, the valence and arousal for each pattern are computed by averaging the values of those moods in the pattern. The clusters can then be projected into the (valence–arousal) space, as shown in Figure 4.4. The exemplar in each cluster is represented in upper case by its mood label.

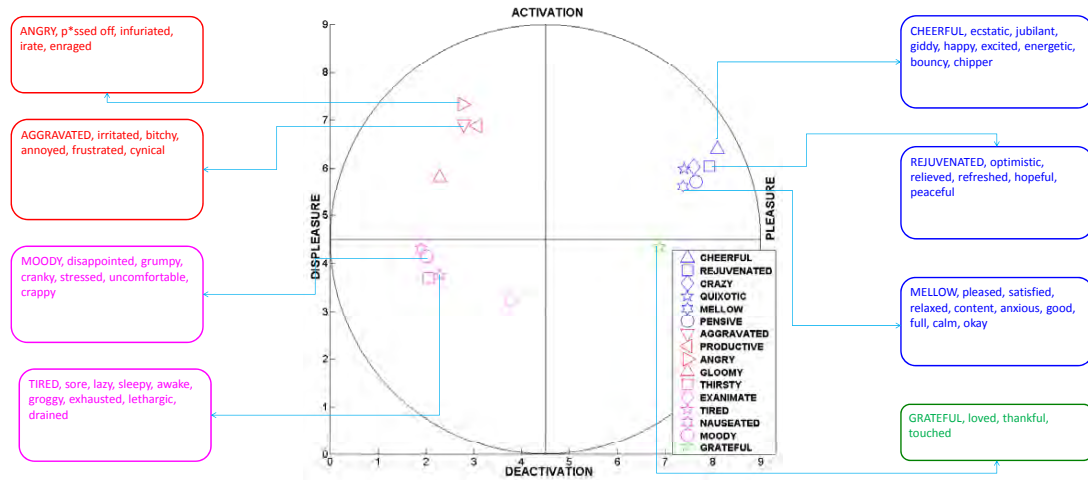


Figure 4.4: Discovered emotion patterns on the affect circle.

Sixteen mood clusters are detailed on the affect circle. These patterns cover all four quarters of the circle and correspond to intuitive mood groupings.

Those moods in the first quarter of the affect circle ( $0^{\circ} - 90^{\circ}$ ) can be considered as representative of high-pleasure and high-activation emotions; for example, *happy* (valence: 8.21, arousal: 6.49), *hopeful* (7.1, 5.78) or *good* (7.47, 5.43) (recall that the valence and arousal values of words in ANEW lexicon are on a scale of 1, *very unpleasant* and *least active*, to 9, *very pleasant* and *most active*). Meanwhile, moods in the second quarter ( $90^{\circ} - 180^{\circ}$ ) express the feeling of sadness at the height of activation; for instance, *angry* (2.85, 7.17), *enraged* (2.46, 7.97) or *frustrated* (2.48, 5.61). Moods in the third quarter ( $180^{\circ} - 270^{\circ}$ ), for example, *moody* (3.2, 4.18) or *lazy* (4.38, 2.65), could be representative of the moods of sadness and deactivation. Further, the *grateful* group could be representative of moods that are both low in pleasure and in the degree of activation ( $270^{\circ} - 360^{\circ}$  of the affect circle), for example, *grateful* (7.37, 4.58) or *thankful* (6.89, 4.34).

The above data-driven affect analysis is, to the author's knowledge, the first of its kind. The coherence of the discovered implicit mood structures validates the use of ANEW-based features for inferring mood in unlabelled text. Such a mood sensor is relatively inexpensive to apply and has a potential role in many applications, including text surveillance, mood-related search facets, characterisation and recommendation of media and communities and ethnographic study.



---

<b>Cluster</b>	<b>Exemplar</b>	<b>Members</b>
1	CHEERFUL	ecstatic, jubilant, giddy, happy, excited, energetic, bouncy, chipper
2	PENSIVE	determined, contemplative, thoughtful
3	REJUVENATED	optimistic, relieved, refreshed, hopeful, peaceful
4	QUIXOTIC	surprised, enthralled, devious, geeky, creative, recumbent, artistic, impressed, amused, complacent, curious, weird
5	CRAZY	horny, giggly, high, flirty, hyper, drunk, naughty, dorky, ditzy, silly
6	MELLOW	pleased, satisfied, relaxed, content, anxious, good, full, calm, okay
7	GRATEFUL	loved, thankful, touched
8	AGGRAVATED	irritated, bitchy, annoyed, frustrated, cynical
9	ANGRY	p*ssed off, infuriated, irate, enraged
10	GLOOMY	jealous, envious, rejected, confused, worried, lonely, guilty, scared, pessimistic, discontent, distressed, indescribable, crushed, depressed, melancholy, numb, morose, sad, sympathetic
11	PRODUCTIVE	accomplished, working, nervous, busy, rushed
12	TIRED	sore, lazy, sleepy, awake, groggy, exhausted, lethargic, drained
13	NAUSEATED	sick
14	MOODY	disappointed, grumpy, cranky, stressed, uncomfortable, crappy
15	THIRSTY	nerdy, mischievous, hungry, dirty, hot, cold, bored, blah
16	EXANIMATE	intimidated, predatory, embarrassed, restless, nostalgic, indifferent, listless, apathetic, blank, shocked

---

Table 4.1: Livejournal moods clustered by similarity to ANEW word use.

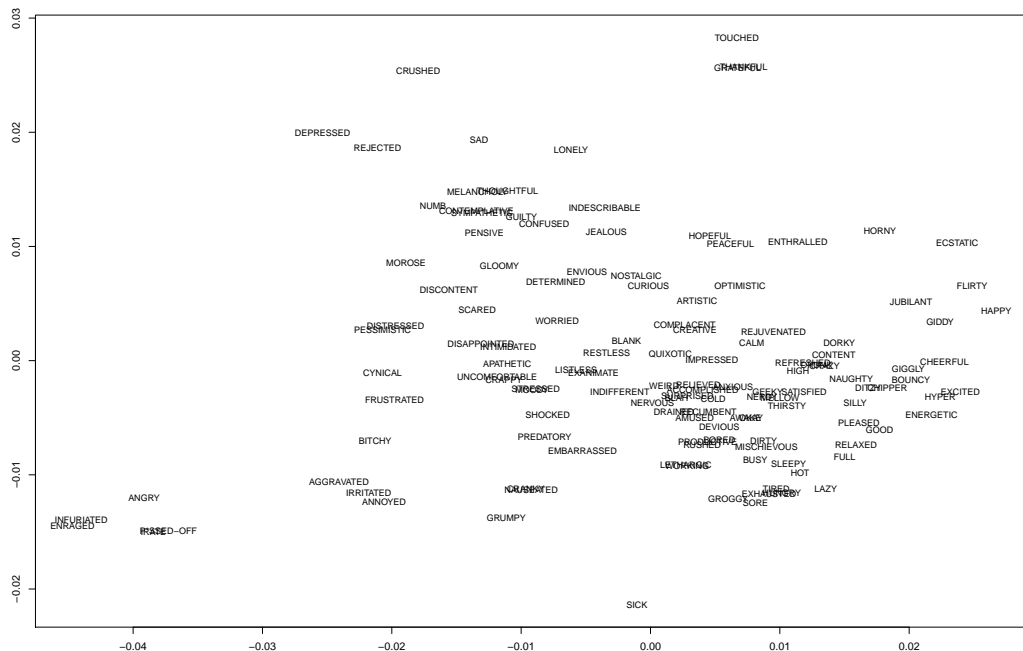


Figure 4.5: Projection of moods onto a two-dimensional mesh using multidimensional scaling.

### 4.1.5 Mood distance

To visualise the similarity of moods, the ANEW usage vectors are subject to multidimensional scaling [Borg and Groenen, 2005] and a hierarchical clustering. Figure 4.5 shows the distance between moods by use of ANEW words through scaling the data onto a two-dimensional mesh. As can be seen, while close to *p\*ssed off*, *infuriated*, *irate* and *enraged*, *angry* is far from *cheerful*, *ecstatic*, *jubilant*, *giddy*, *happy*, *excited*, *energetic*, *bouncy* and *chipper*. In line with this distance, these moods are found in two affinity propagation mood patterns detected in Section 4.1.4, with *angry* and *cheerful* as exemplars.

Figure 4.6 shows the dendrogram of a hierarchical clustering of moods based on their similarity in the use of ANEW words. It can be seen from the figure that *p\*ssed off*, *infuriated*, *irate*, *enraged* and *angry* are in a branch while *cheerful*, *ecstatic*, *jubilant*, *giddy*, *happy*, *excited*, *energetic*, *bouncy* and *chipper* are in another branch of the dendrogram tree.

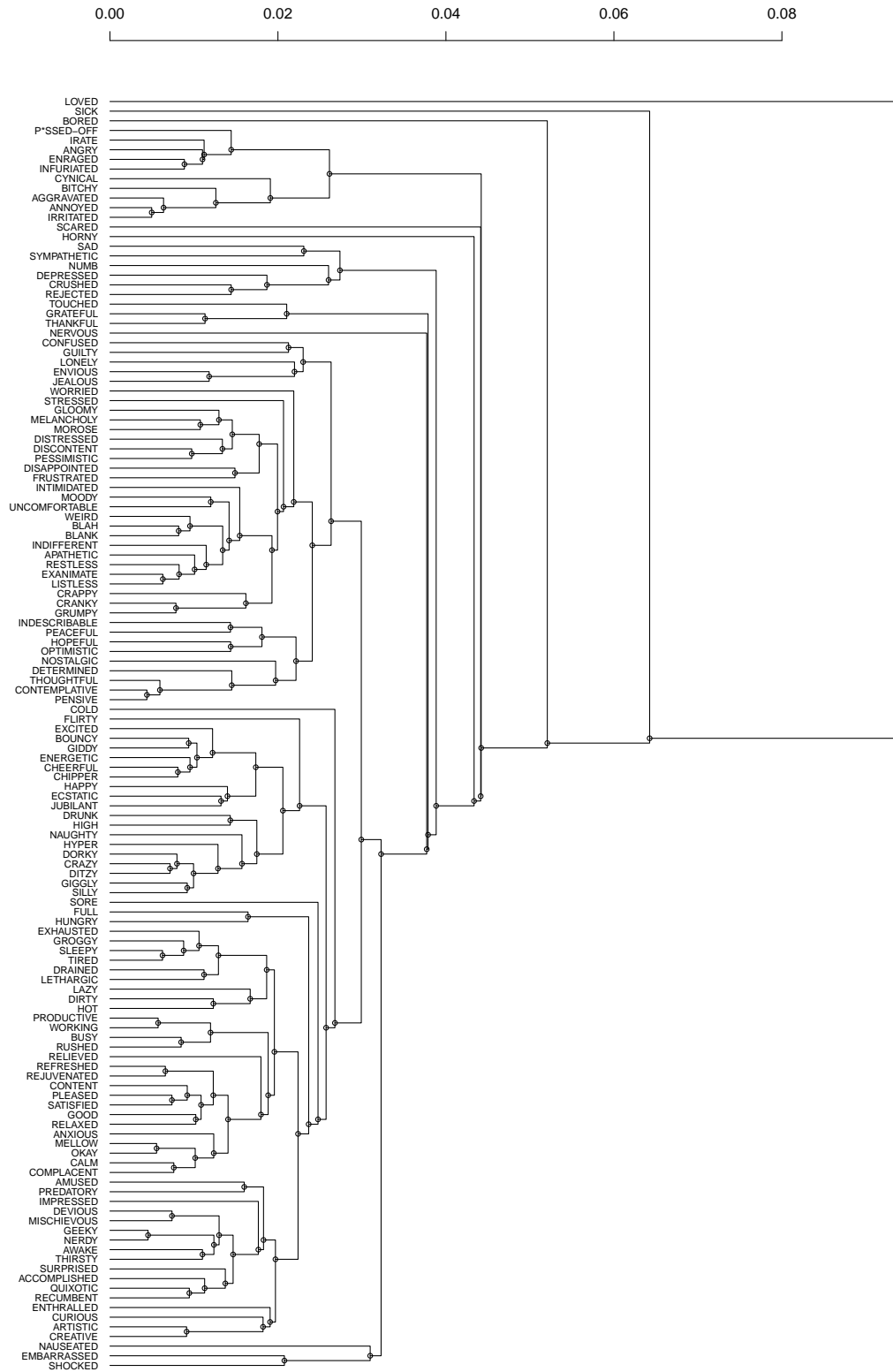


Figure 4.6: Moods in a dendrogram using a hierarchical clustering.

## 4.2 Moods and Affective Lexicon Access

In contrast to the aforementioned process that aims to infer basic emotions from written texts, it is equivalently interesting to question the influence of emotion on the lexicon usage of a social media user during the process of composing a post. From a psychological perspective, there has been some work on the analysis of sentimental influence on writing styles; Chastain et al. [1995], for instance, investigated how mood would affect the usage of an affective lexicon. However, to our knowledge, the connection between affective lexicon access and emotions in the social media domain remains largely unexplored. This motivates us in the next part of this chapter to study the association between mood and the use of affective words during the blogging process.

As seen in Section 2.5.4, Livejournal blog posts carry a large amount of affective information about the bloggers at the time they are blogging. The annotated moods accompanying the blogged text enable research into an association between the two. In seeking which words are likely to appear in a blog post tagged with a given mood, we may find a new lexicon characterising affect information. Conversely, given a word (an entity name, for example) and finding which moods are likely to be tagged when that word is employed, we are gathering attitudes towards the word. We use ANEW as an example for this exploration of mood–lexicon usage correlation.

### 4.2.1 Mood and ANEW usage association

To study the statistical strength of an ANEW word with respect to a particular mood, IG is used [Mitchell, 1997]. Given a collection of blog posts  $\mathcal{B}$  consisting of those tagged or not tagged with a target class attribute mood  $m$ , the entropy of  $\mathcal{B}$  relative to this binary classification (tagged or not tagged with  $m$ ) is

$$\mathcal{H}(\mathcal{B}) = -p_{\oplus} \log_2(p_{\oplus}) - p_{\ominus} \log_2 p_{\ominus} \quad (4.1)$$

where  $p_{\oplus}$  and  $p_{\ominus}$  are the proportions of the posts tagged and not tagged with  $m$  respectively.

The entropy of  $\mathcal{B}$ , relative to the binary classification given a binary attribute  $A$  (for example, if the word  $A$  is present or not) observed, is computed as

$$\mathcal{H}(\mathcal{B}|A) = \frac{|\mathcal{B}_\oplus|}{|\mathcal{B}|} \mathcal{H}(\mathcal{B}_\oplus) + \frac{|\mathcal{B}_\ominus|}{|\mathcal{B}|} \mathcal{H}(\mathcal{B}_\ominus) \quad (4.2)$$

where  $\mathcal{B}_\oplus$  is the subset of  $\mathcal{B}$  for which attribute  $A$  is present in the corpus and  $\mathcal{B}_\ominus$  is the subset of  $\mathcal{B}$  for which attribute  $A$  is absent in the corpus.

The IG of an ANEW attribute  $A$  in classifying the collection with respect to the target class attribute mood  $m$ ,  $IG(m, A)$  is the reduction in entropy caused by partitioning the examples according to attribute  $A$ .

$$IG(m, A) = \mathcal{H}(\mathcal{B}) - \mathcal{H}(\mathcal{B}|A) \quad (4.3)$$

Therefore, with respect to a given mood  $m$ , those ANEW words having high IG are considered likely to be associated with the mood. This measure, also often considered a term-goodness criterion, outperforms others in feature selection in text categorisation [Yang and Pedersen, 1997].

### 4.2.2 Results

Based on the IG values between moods and ANEW, we learn the correlation of moods and the affective lexicon. With respect to a given mood, those ANEW words having high IG with it are considered most likely to be found in the blog posts tagged with the mood. The list of top ANEW words for certain moods is shown in Table 4.2a. Conversely, with respect to an ANEW word, those moods having high IG with it are considered most likely to be tagged to the blog posts containing the word. The list of most likely and least likely moods (low IG) for certain ANEW words is shown in Table 4.2b.

The ANEW words used in the blog posts tagged with moods in the same pattern are more similar than the ones in the posts tagged with moods in different patterns. In

Mood	Top ANEW words associated
Cheerful	fun, happy, hate, good, christmas, merry, birthday, cute, sick, love
Happy	happy, hate, fun, good, birthday, sick, love, mind, alone, bored
Angry	angry, hate, fun, mad, love, anger, good, stupid, pretty, movie
P*ssed off	hate, stupid, mad, love, hell, fun, good, god, pretty, movie
Gloomy	sad, depressed, hate, wish, life, alone, lonely, upset, pain, heart
Sad	sad, fun, heart, upset, wish, funeral, hurt, pretty, loved, cancer

(a) Moods and the most associated ANEW words

ANEW	Most likely moods	Least likely moods
Desire	contemplative, thoughtful	enraged, drained
Anger	angry, p*ssed off	nauseated, grateful
Accident	sore, bored	exanimate, indifferent
Terrorist	angry, cynical	rejuvenated, touched
Wine	drunk, p*ssed off	ditzy, okay

(b) ANEW words and the most associated moods

Table 4.2: Mood and ANEW correlation.

Table 4.2a, the most associated ANEW words in the blog posts tagged with *cheerful* are more similar to those in posts tagged *happy*, than in ones tagged with either *angry* or *p\*ssed off*.

For a given mood, the majority of the ANEW words used in the blog posts tagged with that mood have a similar valence to the mood. In some cases, the ANEW words are different from the tagged mood; for example, the ANEW word *hate* appearing in the posts tagged with *cheerful* or *happy* moods. However, this might be the result of a negation construction used in the text, or of other context.

For a given ANEW word, the most likely moods tagged to the blog posts containing the word are similar to the word in the affective scores. In addition, the least likely moods are different from the ANEW word in the affect measure.

In addition to the use of the ANEW words conveying abstract concepts, for example

*desire* or *anger*, those ANEW words expressing more concrete ideas, for example *terrorist* or *accident*, could be used to learn opinions from social networks towards the ideas. In the corpus, the posts containing the ANEW word *terrorist* are most likely tagged with *angry* or *cynical* moods. Also, the posts containing the ANEW word *accident* are most likely tagged with *bored* and *sore* moods.

### 4.3 Conclusion

Our analysis reveals interpretable and interesting patterns in the organisation, transition and continuum of moods, suggesting valuable empirical evidence about the structure of human emotion. The patterns contain mood synonyms, which can be used interchangeably, for instance, in terms of the sentiment score. The results have additional potential applications, such as in mood-sensitive indexing and retrieval.

In addition, this chapter has revealed an association between moods and the affective lexicon in the blogosphere, indicating the predictive power of moods based on affective lexicon usage in social media. This finding could have potential uses in personalisation applications.

## Chapter 5

# Event Extraction Using Behaviours of Sentiment Signals and Burst Structure in Social Media

In the previous two chapters we investigated the emotion structure of mood in the blogosphere, characterising mood by different feature sets and clustering them into different sentiment patterns. In its first application, in this chapter, the sentiment information being conveyed in on-line diaries is utilised to learn important events being discussed by bloggers. This is motivated by the observation that significant world events often cause the behavioural convergence of the expression of shared sentiment. Therefore, we examine the use of the blogosphere as a framework to study user psychological behaviours, using their sentiment responses as a form of ‘sensor’ to infer real-world events of importance automatically. Firstly, we shall present a novel temporal sentiment-index function using a quantitative measure of the valence of words in blog posts in which the set of affect bearing words is inspired from psychological research in emotion structure. The annual local minimum and maximum of the proposed sentiment signal function are utilised to extract significant events of the year and corresponding blog posts are further analysed using topic-modelling tools to understand their content. The chapter then examines the correlation of topics discovered in relation to world news events reported by the mainstream news service provider, Cable News Network (CNN), and by using the



Google search engine. Next, to understand sentiment at a finer granularity, the stochastic burst detection model of Kleinberg, the KLB, is extended to work incrementally with stream data, to extract bursts in specific sentiments (for example, a burst of observing ‘shocked’). The blog posts at those time indices are analysed to extract topics and these are compared to real-world news events. Our comprehensive set of experiments conducted on a large-scale set of 12 million posts from Livejournal shows that the proposed sentiment index (SI) function coincides well with significant world events, while bursts in sentiment allow us to locate finer-grain external world events.

## 5.1 Sentiment Reactions in Social Media

Recall that social media is a new type of media wherein users assume a myriad of roles, including consumers of information, publishers, editors and commentators. The blogosphere is one example of this new decentralised and collaborative forum, through which people participate, exchange opinions, generate content and interact with others. This user-generated content mirrors subjective user sentiment, much more so than other written genres. New opportunities exist for opinion mining and sentiment analysis and thus this topic has attracted much recent interest [Coontz, 2009, Giles, 2010, Pang and Lee, 2008]. Sentiment information in social media has also been used to explain real-world trends such as by mapping the proportion of *anxious* posts in Livejournal with the S&P 500 index [Gilbert and Karahalios, 2010] and predicting the 2009 German election based on political sentiment contained in Twitter [Tumasjan et al., 2010]. However, joint sentiment–topic investigation, in which sentiment and topic interaction is taken into account together, has received little attention.

Detecting sentiment-based events, Balog et al. [2006] found spikes of events based on the mood tags in Livejournal posts. Using a simple threshold approach, they extracted blog posts tagged with bursty moods and used frequent words to interpret related events. Since the bursty intervals were detected based on a simple threshold, a large number were discovered, often occurring over short periods. State–space approaches have been used for burst detection of terms in a document corpus. Kleinberg [2003] observes that certain topics in emails are characterised by a sudden

increase in textual features. These high-intensity periods of terms, called *bursts*, grow in intensity over a period, with the KLB burst detection algorithm being a simple and efficient means of detecting them. However, predefined sets of parameters obtained from offline data are required to calculate the probability of generating relevant documents and their state changes. While suitable for offline settings, this method is not scalable for situations in which diverse and changing data is added over time, as in the case of the blogosphere.

Therefore, open problems include constructing robust sentiment detection and incremental burst detection algorithms that can be applied to large-scale blog data. Research into emotional responses towards traumatic or significant events has been explored in psychology [Back et al., 2010, Bracken et al., 1995, Silver et al., 2002]. Triggered by events in their lives, people write about their experiences in blogs and footprints of their emotional state often emerge in on-line journals. At a larger scale, the blogosphere is a rich source for studying collective emotional response to external events. The questions we ask are: Can we use the affect of the blogosphere as indicative of significant world events? Can we use the sentiment in the blogosphere as an indicator of collective emotional states? Can we perform fine-grain analysis on a specific mood and examine its correlation to external events?

We examine the sentiment at two levels of granularity—aggregated sentiment and specific sentiment across a corpus. To understand the utility of aggregated sentiment, we estimate the total affective score of all blog posts, constructing a temporal SI. Affective scores of on-line journals are computed in two ways: first using the valence of mood tagged by the user for each post; and second, by extracting the sentiment-bearing words according to the LIWC [Pennebaker et al., 2007b] for ‘positive’ and ‘negative’ mood categories. These time-indexed sentiment scores are then analysed to detect maxima and minima, coincident with maximal points in sentiment shifts. The corresponding time indices are used to extract real-world events (from news sites), which are then compared against the topics of the corresponding blog posts, extracted with LDA [Blei et al., 2003]. In the second part of our work, we examine the extraction of ‘bursts’ in a particular sentiment across the corpus. We extract time periods of the mood burst based on KLB [Kleinberg, 2003], making adaptations to the algorithm to apply it incrementally to stream data. We extract the topics of the corresponding blogs and examine how the topics correlate with the top news stories extracted from CNN. Our experiments are performed using more

than 12 million Livejournal posts and our results demonstrate that the SI is a rich signal, whose maxima and minima correlate well with real-world events. We show that this holds true at a finer level when detecting specific sentiment bursts, and once again, extreme points coincide with finer-grain world events. We demonstrate that the KLB can be efficiently adapted to handle streaming data, detecting salient bursty structures in blog data in real-time without compromising performance.

Our contribution in this chapter is three-fold. Firstly, we present approaches to construct time-series sentiment indices, whose extrema can be correlated to real-world events. We validate the topics and events extracted using a list of top stories as voted by CNN. Secondly, we introduce the novel concept of *sentiment burst* and employ a stochastic model for detecting bursts in text streams. This provides the foundation for sentiment-based burst detection and subsequent bursty event extraction. Thirdly, we implement an incremental version of the KLB algorithm, which has wide applicability in real-time burst detection in stream data. Evaluating the events extracted is a challenging task. Therefore, an additional contribution is an effective method for evaluating and ranking events extracted using a combination of topic modelling and Google search.

The rest of this chapter is organised as follows. In Section 5.3, the sentiment score of the whole sphere in a given time is shown to act as a signal for detecting events discussed among bloggers. Then, the role of each mood in event detection is studied in Section 5.4. This section presents the results of bursty pattern discovery of moods and the related facts and figures in different algorithms. It is followed by concluding remarks.

## 5.2 Dataset and Evaluation

**Dataset.** We experiment using the CHI06 database, described in Chapter 3. A subset of this database, ranging from 1 May 2001 to 31 December 2004, is used. This subset contains more than 12 million blog posts tagged with mood labels from a set of 132 moods predefined by Livejournal. The number of blog posts each day in the period is shown in Figure 5.1. As can be seen, the number increased with time. We note that our method is applicable even when mood tags are not available and

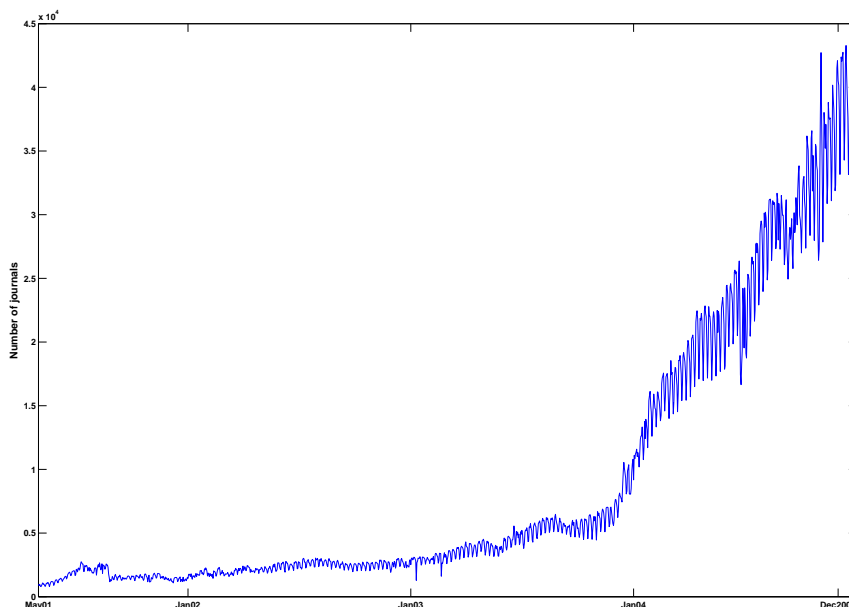


Figure 5.1: Daily number of blog posts over the corpus period.

thus can be applied to alternative datasets.

**Evaluation method.** Evaluation of events detected using our proposed algorithm is challenging due to the scale of the dataset and the unconventional setting of the problem considered in this chapter. To the best of our knowledge, no benchmark dataset is available. For example, assume our mood burst detection algorithm returns 1 February 2003 as a significant time index. How do we evaluate if an external event of importance occurred? Two questions remain: *What* is the event? *How* do we know it is significant? One obvious approach is to conduct questionnaires. However, the scope and scale of the dataset quickly rules out the feasibility of this approach.

In this chapter, to infer about topics within a post we use a popular Bayesian probabilistic method for topic modelling known as LDA [Blei et al., 2003]. In the above example, we would collect all blog posts, each of which is a document, posted for the date of 1 February 2003 and run LDA on this corpus, using the returned topics as the content for the posts. To run LDA, we use Gibbs sampling as proposed in Griffiths and Steyvers [2004], which is guaranteed to converge in a limited number of iterations. The topic model requires that the number of topics be specified in advance. A commonly accepted rule of thumb is to choose the number of topics  $k$

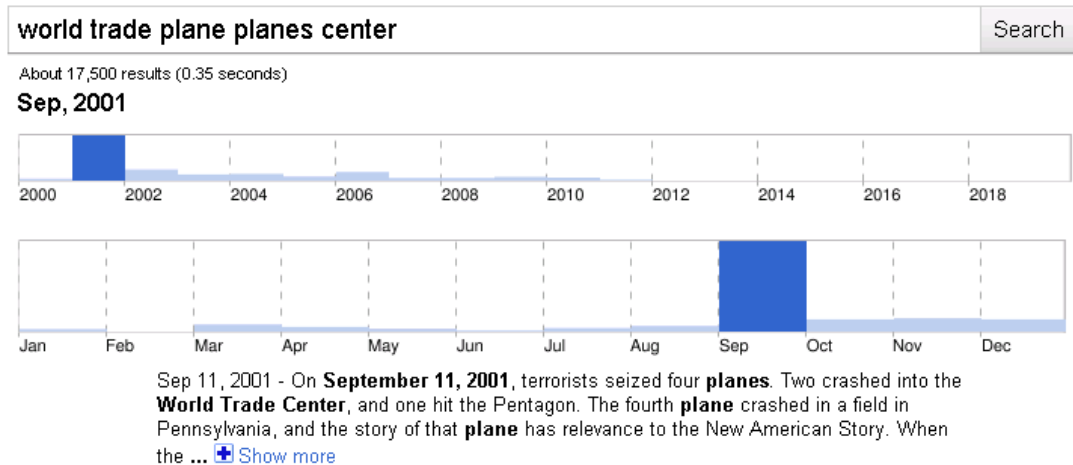
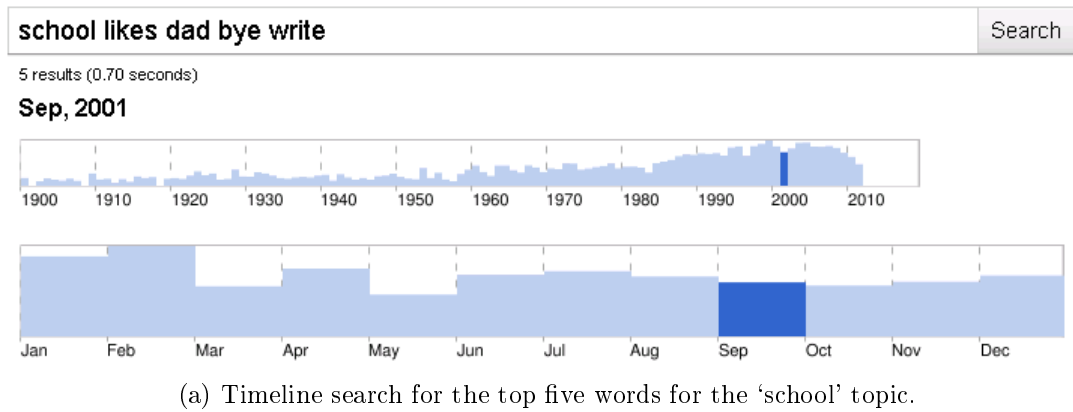


Figure 5.2: Examples of querying Google for top five words for two topics for the date of 11 September 2001. By general search, we cannot determine which topics mention an event since both return millions of results. By restraining the search results to September 2001, ‘school’ receives five results, whereas ‘WTC’ returns 17,500 results. Within the timeframe of 10 September–12 September 2001, the search engine returns zero for ‘school’ and 6,830 results for ‘WTC’.

in the scale  $\alpha \log V$  where  $V$  is the vocabulary size and  $\alpha$  is a constant. For our dataset, the vocabulary size is on average equal to 50,000. A common practice is to use  $\alpha = 1$ , then  $k = \alpha \log V \approx 10$ . This suggests the number of topics to be 10. Since Gibbs sampling is a special case of Markov Chain Monte Carlo, the early samples depend on the initialisation value. To obtain samples that are independent from the initialisation values, it is customary to discard some initial samples as *burn-in*. In our implementation, we discard the first 1,000 samples for *burn-in* and use the next 5,000 samples for inference.

We then propose two approaches to evaluate the significance and impact of the topics returned from our detection algorithms. Firstly, for topics extracted using the SI, we use the set of annual top stories reported by CNN from the corresponding year to evaluate against our extracted topics (see Section 5.3 for more details). CNN ranking has been chosen as groundtruth for event detection, for example, in Chua et al. [2011].

However, ‘significant event’ is an ill-defined term due to its context-specific nature and is often biased by personal knowledge, experience, cultures and preferences. The CNN ranking is likely biased to a US-centric view of which events qualify as significant, and a populist one at that, as the ranking is poll-driven. We therefore supplement the event groundtruth using a method based on Google Timeline, which returns the news articles associated with a given topic within a period of time. Since no polls are involved, this approach offers a better degree of objectivity when compared to the CNN ranking.

In this second approach, we use the volume returned from Google queried by the top five words per topic during the timespan of the event to examine its significance and impact. To evaluate whether the impact is significant, we construct a baseline measure by randomly selecting five words from the vocabulary set and querying Google to get the returned volume. We do so 1,000 times, taking the average as the reference baseline. An example of this approach is shown in Figure 5.2. Our intuition is that, by restricting the search results to within a limited timeframe (dictated by the time span of the event), the probability of five words simultaneously occurring is extremely unlikely (as validated by the empirical results). Therefore, the impact of events extracted can be justified by the Google results referring to the randomness. We found that this approach is very effective for the task. The average return

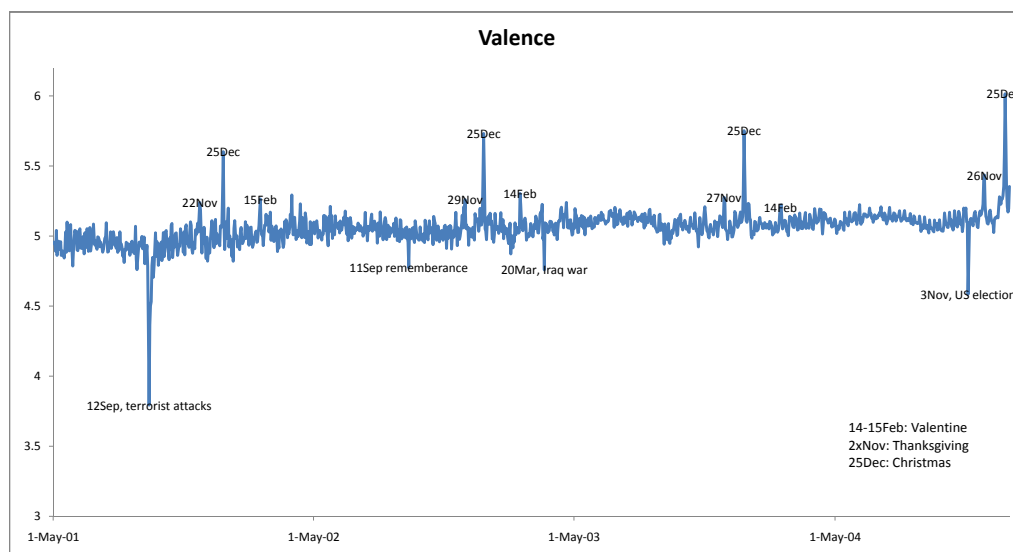


Figure 5.3: Sentiment indices computed using the valence value of the tagged moods.

from Google for the random baseline count is  $0.01 \pm 0.28$ , whereas the results from our topics are significantly higher (see Section 5.4). A similar approach has been used in Luo et al. [2012], in which the significance of events is estimated based on the number of documents discussing them. The greater the number of documents reporting about an event, the more significant it is.

### 5.3 Sentiment Index and Event Extraction

In this section, we formulate a *temporal SI*. At the global level, sentiment indices serve as a psychological signal, accumulating the average community sentiment across time. The signal provides rich analysis opportunities. For example, a large deviation, local minima or maxima, in this time-series signal could correlate psychological shifts to real-world events. We present two approaches to construct the SI. In the first approach (see Section 5.3.1), since blog hosting sites (for example, Livejournal) provide a mechanism for the user to tag their mood when the message is posted,<sup>1</sup> we use mood tags and their valence values to construct the SI. However, mood tags are not always available. Therefore, in the second approach (see Section

<sup>1</sup>That is, the users can indicate if they are sad, happy or angry and so on when they compose the message.

5.3.2) we construct the SI directly using the affective words used in the blog posts. It is interesting to note at the outset that both approaches provide meaningful results that are comparable, suggesting that our sentiment-based methods are applicable in wider contexts.

### 5.3.1 Mood-based sentiment index extraction

We describe a method using the emotion information in the mood label tagged to a blog post to discover significant events. Our observation is that a sudden increase or fall in the intensity of emotion expression could be correlated to a real-world event. For example, a flood of *angry* or *shocked* posts is expected following the 9/11 event. To identify this, we formulate a novel emotional signal, termed SI, to capture the normalised average sentiment over time.

To express mood quantitatively, we need a method to measure the sentimental level of a mood. Recall that, in the domain of psycholinguistics, one widely accepted sentiment measure is *valence*, which indicates the degree of *happiness* a word conveys.

There is some debate about the use of valence—the degree of pleasure—for measuring emotions. For example, Colombetti [2005] states that pleasures are rarely pure; most of the time, they are mixed; that is, they contain pain. Solomon [2003] argues that the polarity of positive and negative affects is simple-minded and opposes the idea of defining emotion in terms of their valence.

However, valence has seen common use as a principle factor to measure emotion. For example, studies on the structure of English emotion words have revealed that valence is an important dimension [Fontaine et al., 2007, Russell, 1983, Smith and Ellsworth, 1985]. In other languages, valence has also proven to be a primary factor in estimating emotions. For example, Galati et al. [2008] discovered that hedonic valence, distinguished by the degree of happiness and sadness, is a major dimension in the organisation of the emotional lexicons in Italian, French, Spanish, Catalan, Portuguese and Romanian. Gehm and Scherer [1988] analysed typical German emotion words using multi-dimensional scaling (MDS) and found that one fundamental dimension is pleasure/displeasure. Yoshida et al. [1970] applied MDS to Japanese emotional words and found that the first dimension is pleasantness-unpleasantness.



Shaver et al. [2001] argued that different cultures lead to different fine-grained distinctions in the structure of the emotion lexicons (subordinate-level emotion concepts) in American English and Indonesian. However, they conclude that emotion concepts in both languages are similar at the superordinate levels; that is, they belong to two categories: positive emotions and negative emotions. Church et al. [1998] compare emotion concepts across cultures, employing samples from English and Filipino. They argue that the nature and range of the emotion lexicon of Filipino and English people is similar. In addition, they conclude that the dimensional structure of English emotion terms, in pleasantness and arousal dimensions, is also supported by Filipino emotion words.

We use the valence value that has been estimated following a consensus basis in ANEW [Bradley and Lang, 1999].<sup>2</sup> This lexicon contains 1,034 English words, tabled with corresponding sentimental measures of *valence* and *arousal*. These two dimensions have also been used by psychologists in the circumplex model of affect [Russell, 1980, 2009] to explain the structure of emotion when emotion states are conceptualised as combinations of these two factors. We then use the valence values as building blocks to compute the SI for the task of event extraction.

Formally, the SI is defined as follows. Denote by  $\mathcal{B} = \bigcup_{j=1}^T \mathcal{B}_j$  the collection of the entire dataset where  $\mathcal{B}_j$  denotes the set of blog posts arriving for  $j$ -th date and  $T$  is the total number of days in the corpus. Denote by  $\mathcal{M} = \{sad, happy, \dots\}$  the set of mood labels predefined by Livejournal. Each blog post  $b \in \mathcal{B}$  in the corpus is labelled with a mood  $m(b) \in \mathcal{M}$ . Denote by  $v_m$  the valence value of the mood  $m \in \mathcal{M}$ . We then compute the SI  $I(j)$  for  $j$ -th date as:

$$I(j) = \frac{\sum_{b \in \mathcal{B}_j} v_{m(b)}}{|\mathcal{B}_j|} \quad (5.1)$$

In other words, the SI for a given date is an accumulation of valence values from all posts appearing on that date, normalised by the total count. We turn to examining this index under the day-of-the-week and monthly effects and then observe its

---

<sup>2</sup>Mood labels that are not listed in the ANEW lexicon are assigned the valence values of their siblings or parents from Livejournal’s shallow mood taxonomy (<http://www.livejournal.com/moodlist.bml>).

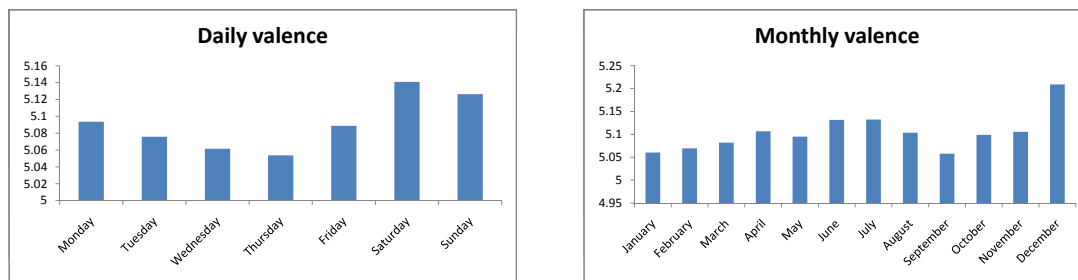


Figure 5.4: Weekly and monthly patterns of mood-based SI.

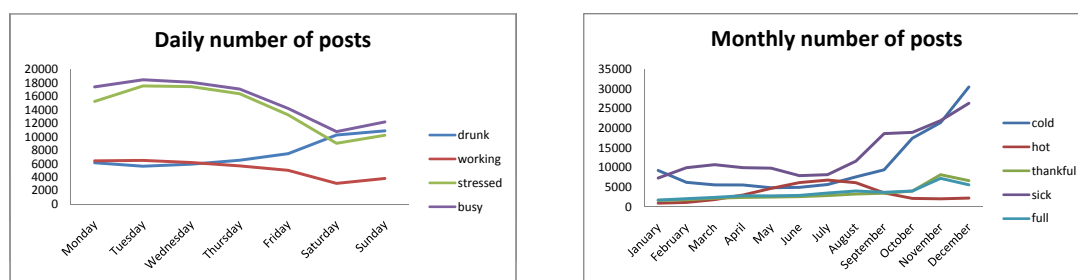


Figure 5.5: Weekly and monthly number of posts tagged with the moods.

variation across the period of the dataset.

### 5.3.1.1 Daily and monthly valence

To examine periodic events or bloggers' behaviours, we aggregate distributions of the valence values by weekdays and by months. As shown in Figure 5.4, the SI is lowest on *Thursdays* and highest on *weekends*. When looking at monthly patterns, *December* is top in the index, being a month of many celebrated days, such as Christmas and New Year, while the drop in *September* was likely due to 9/11.

Variations in SI can be explained by the fact that periodic events or bloggers' habits can cause a cyclic rise of tagging certain moods. For example, in the case of weekdays, Figure 5.5 shows that '*busy*' and '*stressed*' are quite similar in tagging distribution, peaking on Tuesday and gradually decreasing for the following days ('*Tuesday at 11:45 is the most stressful time of the week*'<sup>3</sup>). The behaviour of '*working*' and

<sup>3</sup><http://www.telegraph.co.uk/news/newsttopics/howaboutthat/5113653/Tuesday-at-1145-is-most-stressful-time-of-the-week-survey-suggests.html>, retrieved September 2011.

‘*drunk*’ makes for an interesting observation. ‘*Working*’ is highest at the beginning of the week, decaying slightly as the week progresses. ‘*Drunk*’ is almost a mirror image of ‘*working*’, with its lowest point at the beginning of the week, rising to peak at the weekend.

For the monthly case, as shown in Figure 5.5, it can be seen that ‘*hot*’ and ‘*cold*’ have an inverse relationship. Interestingly, the distribution of ‘*cold*’ is quite similar to ‘*sick*’ (‘*Cold noses reduce ability to fight virus attacks*’<sup>4</sup>) and ‘*cold*’ alternates with ‘*hot*’ across a year. Of interest too is the number of ‘*thankful*’ and ‘*full*’ posts, which are lowest at the beginning of the year and increase moderately to peak at the end of the year.

### 5.3.1.2 Abnormal valence

Despite the periodic nature of SI by day of the week and by month, anomalies are still found in the SI time-series. Examining the structure of the time-series SI  $I(j)$  for these anomalies might reveal events of significance that have caused a psychological shift in the population. A variety of different behaviours manifest in the SI could be explored. In this chapter, in particular, we are interested in the date when the SI reaches its maximum and minimum every year. Figure 5.3 shows the sentiment indices computed using the valence value of the tagged moods. The four highest peaks correspond to highest sentimental dates with the most positive sentiment for 2001 to 2004 respectively. Likewise, the four lowest peaks coincide with the dates with the most negative sentiment.

To gain insight into the context of the real-world events correlated with these dates, we retrieve all the blog posts during the corresponding time to form a collection of documents, upon which topic extraction is applied to describe its content. Our results show that the highest valence sentiment at the annual granularity re-occurs on 25 December—Christmas Day. The highest probability topics learned by LDA from the set of related posts from these dates are dominated by ‘Christmas’. These results are quite intuitive, as one would expect many bloggers are ‘*happy*’ during this time of the year. Other public holidays are also found to be high in valence, for example Valentine’s, Easter and Thanksgiving (see Figure 5.3).

<sup>4</sup><http://edition.cnn.com/2005/health/11/14/cold.chill/index.html>, retrieved September 2011.

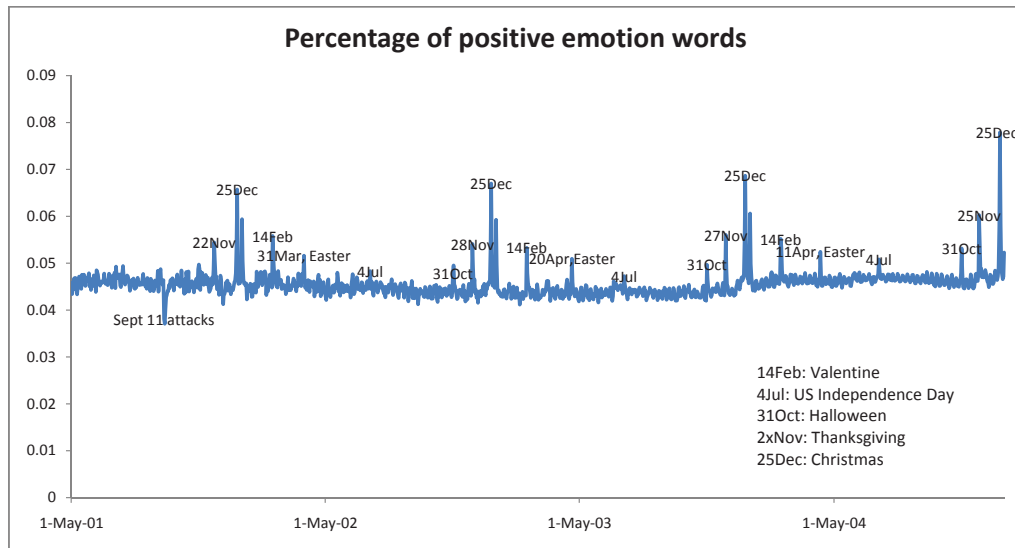
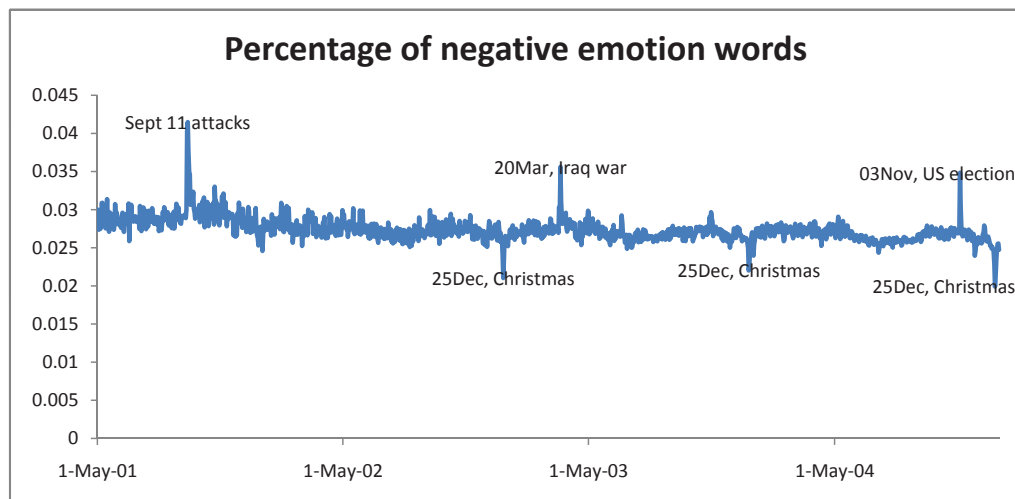
(a) Sentiment indices computed using *posemo* from the LIWC feature set.(b) Sentiment indices computed using *negemo* from the LIWC feature set.

Figure 5.6: Sentiment indices using negative and positive words in blog posts and the corresponding events.

Year	Day	Highest ranked topics	Related events	CNN
2001	12 Sep	<b>world</b> trade plane center pentagon planes building towers buildings war york hit school america crashed heard tv watching tower watched	Terrorist attacks	1st
2002	11 Sep	<b>remember september world</b> lost country lives america died 11th american watching bless family live trade plane forget anniversary watch attacks	9/11 anniversary	10th
2003	20 Mar	<b>war</b> iraq bush world country asia - - - - -	War in Iraq	1st
2004	03 Nov	<b>bush country years america war president world</b> kerry american voted gay vote election iraq america signs asia europe india	2004 election	1st

Table 5.1: The events detected for the dates of lowest sentiment (based on the topics highest ranked by Google Timeline) and their annual ranking by CNN.

In contrast, we find that a majority of the lowest sentiment dates occur during sobering events. For example, in 2002, the lowest sentiment date is 11 September. To learn which topics were discussed on the day, we run LDA over all blog posts made that day (2,973 posts) for 10 topics. The list of topics are ranked using the Google method mentioned in Section 5.2 and are illustrated in Figure 5.2 with the timeframe of 10 September 2002–12 September 2002. The topic mentioning the 9/11 anniversary scores highest in the Google method by querying ‘remember, September, world, lost, country’.

The same process is conducted for other lowest sentiment dates for other years. The topics returned from LDA (see Table 5.1) and ranked highest by the Google method on these dates are compared against the set of top stories ranked by CNN for the corresponding years.<sup>5</sup> The results demonstrate a surprisingly effective set of extracted events: all of the returned events coincide with the first ranked story on the CNN list, with the exception of 2002. However, even here the detected event is ranked tenth by CNN.

### 5.3.2 Content-based sentiment index extraction

Instead of using mood labels, which are not always available, we use sentiment words in the content of the blog post directly to construct our SI. To this end, we utilise

<sup>5</sup><http://edition.cnn.com/specials/2001/yir/>, <http://edition.cnn.com/specials/2002/yir/>, <http://edition.cnn.com/specials/2003/yir/>, <http://edition.cnn.com/specials/2004/yir/>, retrieved September 2011.

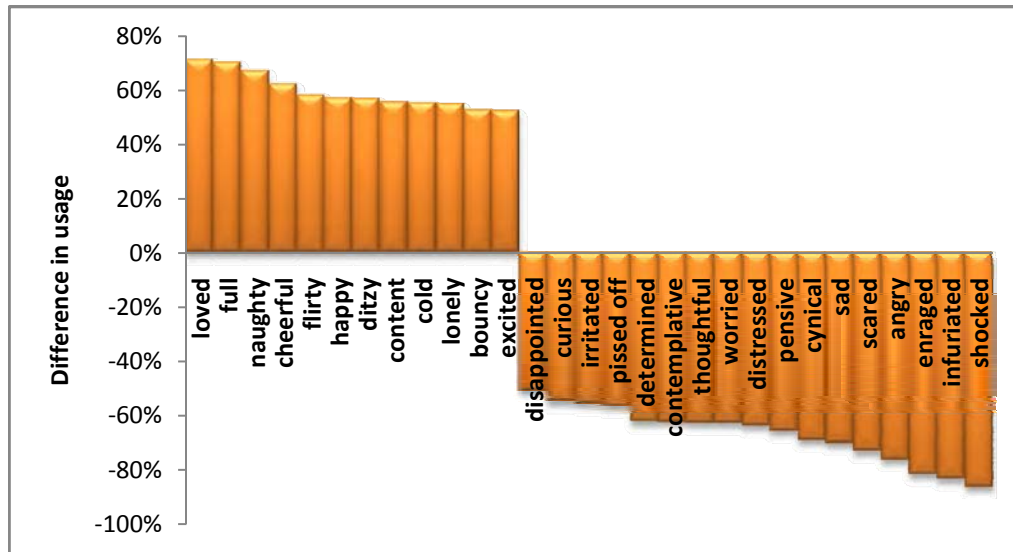


Figure 5.7: Happy versus sad events: the moods above the zero line are used more often for happy events and those below the zero line are for traumatic or sad events ( $ps < 0.005$ ).

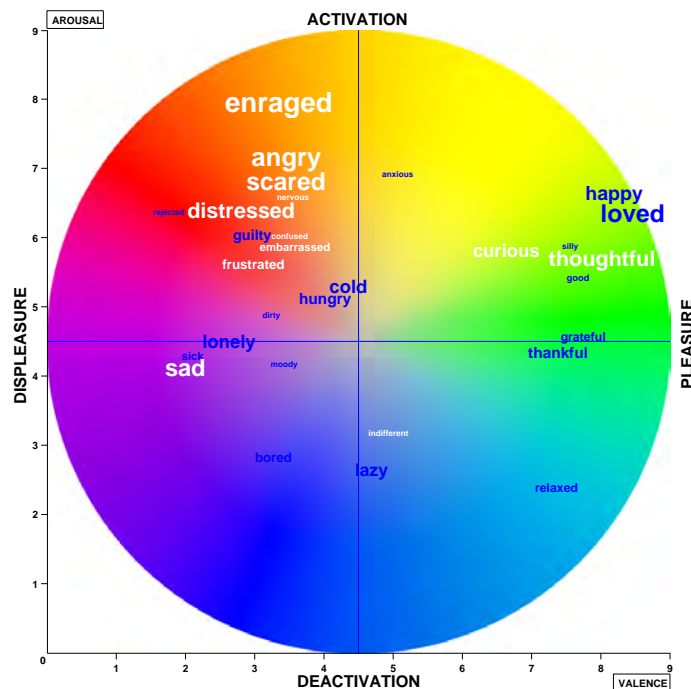


Figure 5.8: Happy versus sad events: visualisation of corresponding mood labels in the core affect model [Russell, 1980, 2009]. Those in blue are used more often for happy events, whereas moods in white are used more often for sobering or sad events. The size of the label reflects the extent of the difference ( $ps < 0.005$ ).

words from the LIWC [Pennebaker et al., 2007b], which classifies words into a set of linguistic and psychological processes. We then compute the percentage of affective words in a blog post as a measure for sentiment. Two LIWC categories—*posemo* (positive emotion words) and *negemo* (negative emotion words)—are used here as an aggregated indicator of the sentiment reaction of bloggers shown in the content of their blog posts. This is analogous to the two extreme ends of valence values presented earlier: one is for happiness and the other is for sadness.

As expected, as shown in Figure 5.6a, the highest percentages of positive emotion words in the blog posts are found on public holidays, such as Christmas, Thanksgiving and Valentine’s Day, with the lowest percentage of *posemo* words occurring on the anniversary of the September 11 attacks. Conversely, as shown in Figure 5.6b, *negemo* usage reaches its highest points on the days of traumatic events, such as the September 11 attacks, the Iraq war and the US election, and reaches its lowest points on Christmas Day. This is an almost identical result to that found for the mood-based approach, with the exception of the anniversary of September 11 in 2002, suggesting that by exploiting the textual content alone, without recourse to manually annotated mood labels, the proposed method is still effective.

### 5.3.3 Happy events versus sad events

Our observation from the two experiments above is that there is a separation between mood labels attached to events at the two extreme ends of emotion, namely happy and sad, which in turn correspond to the two ends of the valence spectrum. To gain an empirical understanding of how the structure of these moods is distributed, we manually collected 211 topics in which the top words correspond to happy events, including: ‘4th July’, ‘Christmas’, ‘Easter’, ‘Halloween’, ‘Thanksgiving’ and ‘Valentines’; and 71 topics related to known traumatic events, including ‘Iraq war’, ‘the loss of space shuttle Columbia’, ‘the attacks of September 11’ and ‘Ronald Reagan’s death’. For this experiment, we run the topic model LDA for every day in the corpus and, using the post-topic distributions returned by LDA, we compute the proportion of mood labels for each topic. We then use a standard non-parametric statistical test, Mann–Whitney U, and find that 96 moods are significantly different among happy versus sad events (Mann–Whitney U tests,  $ps < .005$  two-tailed,  $n_1 = 211$ ,

$n_2 = 71$ ).

Moods with the largest differences are shown in Figure 5.7. In accordance with our intuition, *shocked*, *angry*, *sad* and *disappointed* are used more in traumatic events while *happy* and *excited* are used more on holidays. Figure 5.8 further illustrates the separation of these mood labels using the core affect model proposed by psychologists [Russell, 1980, 2009]. We note the interesting result that mood labels used on happy events are located mostly in the top right quarter of the affect circle, indicating states of *high valence* and *high arousal*. In contrast, for traumatic and sad events, the corresponding mood labels are mostly *low valence* and *high arousal* and are situated in the top left quarter of the affect circle. That is, moods in both types of events are of high arousal, but are differentiated by valence.

## 5.4 Mood-based Burst and Event Extraction

In the previous sections, we have presented different methods to construct the global SI as a time-series function, and have extracted events based on the extremes of these functions. In this section, we present another approach to extract events based on the burst structure detected with respect to individual moods. While the previous approach characterises aggregated sentiment of the whole population, and is therefore more likely to discover events at a macro level, the burst-based event extraction presented in this section tends to yield more subtle and focused events regulated by the mood label being used. Again, we use the evaluation approach presented in Section 5.2 to validate performance.

### 5.4.1 Burst-based event extraction

Recall our previous notation that the blog dataset is assumed to grow over time  $\mathcal{B} = \bigcup_{t=1}^T \mathcal{B}_t$  where  $\mathcal{B}_t$  denotes the set of blog posts arriving at time  $t$ . To be precise, it is a group of blog posts arrived at over time interval  $t$ , when the time axis is discretised into regular time spans. This is a more realistic scenario in practice than dealing with each individual blog post over time. In the extreme case though, we can



still treat each  $\mathcal{B}_t$  to contain exactly one blog post. Recall that  $\mathcal{M} = \{sad, happy, \dots\}$  the set of moods predefined by Livejournal; each blog post  $b \in \mathcal{B}$  in the corpus is labelled with a mood  $m(b) \in \mathcal{M}$ . The objective is to detect whether there is any burst of posts related to a given mood  $m \in \mathcal{M}$  observed over time in  $\mathcal{B}$ . We appeal to a finite automaton burst detection approach by Kleinberg [2003]. Denote by  $|\mathcal{B}|$  the number of elements in the set  $\mathcal{B}$ , the total number of blog posts up to time  $t$  is thus:

$$N_t = \sum_{i=1}^t |\mathcal{B}_i| = \sum_{i=1}^t n_i \quad (5.2)$$

where we further use  $n_i = |\mathcal{B}_i|$  for brevity. A blog post  $b$  is said to be relevant to the mood  $m$  if its mood is tagged as  $m$ . Therefore, the number of blog posts relevant to mood  $m$  arriving at time-slice  $t$ , denoted as  $r_t(m)$ , is

$$r_t(m) = |\{b \mid b \in \mathcal{B}_t, m(b) = m\}| \quad (5.3)$$

Thus, at time  $t$  we are given  $n_t$  blog posts, and among them there are  $r_t(m)$  relevant to mood  $m$ . Therefore, the total number of blog posts relevant to mood  $m$  up to and including time  $t$  is:

$$R_t(m) = \sum_{i=1}^t r_i(m) \quad (5.4)$$

A simple approach to detect bursts is to plot the histogram of  $R_t(m)$  versus  $n_t$  and, applying a threshold (THD), declare bursts at those points at which the threshold is exceeded [Balog et al., 2006]. However, as pointed out in Kleinberg [2003], THD might easily generate short spurious bursts. The KLB algorithm [Kleinberg, 2003], a state-based approach to detect bursts, addresses this point. The essential idea of KLB is to model bursting as a generative process using a finite automaton. In the simple two-state model, one state is responsible for generating blog posts during non-burst periods and the other is for when a burst occurs. The input to the algorithm is an observation sequence of pairs  $\{R_t, N_t\}$  for a mood and the ratio of the emission of relevant documents at state  $\pi_0$  ( $\pi_0 = R_T/N_T$ ) and at state  $\pi_1$  ( $\pi_1 = \lambda\pi_0$ ). Two costs incurred at time  $t$  are the emission probability of the pair  $\{R_t, N_t\}$  and the cost of moving from non-bursty to bursty state—transition  $\gamma = \ln(T+1)$ . The KLB method reduces to finding the optimal sequence of states by minimising the two

---

**Algorithm 5.1** Two-state automaton to detect bursts incrementally (iKLB).

---

```

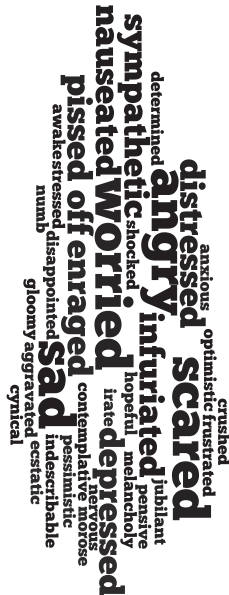
for  $t = 1$  to  $T$ 
*  $\pi_o = R_t/N_t$  : non-burst probability.
*  $\pi_1 = \lambda\pi_o$  : burst probability.
*  $\gamma = \ln(t + 1)$  : transition cost.
if ( $t = 1$ )
*  $c_1(s) = -\ln [\text{Binomial}(N_1, R_1, \pi_s)] + \gamma s$ 
*  $q_1^* = \underset{s}{\text{argmin}} c_1(s)$ 
else
*  $c_t(s) = -\ln [\text{Binomial}(N_t, R_t, \pi_s)]$ 
if  $q_{t-1} = 0$ 
*  $c_t(s) = c_t(s) + \gamma s$ 
endif
*  $q_t^* = \underset{s}{\text{argmin}} c_t(s)$ 
endif
endfor

```

---



(a) The set of bursty moods returned by the burst detection algorithm.



(b) The set of 35 moods found to be significant for event extraction selected from the top 50 ranked events.



(c) Significant moods selected by THD (manually adjusted to obtain 10 bursts per mood label).

Figure 5.9: The set of bursty moods (that is, moods with at least one burst detected) and significant moods that are found to be suitable for event extraction (cloud visualisation size is proportional to the number of bursts detected for the respective mood).

Score	CNN	Mood – bursty time	Topics	Related events
6390	1	shocked: 11–15Sep01	<b>world trade plane center</b> crashed planes pentagon towers attacks twin york attack president bush terrorist washington united south fight american	September 11
3040	1	angry: 18–20Mar03	<b>war</b> bush country iraq saddam coming america president american xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx	War in Iraq
1750	10	jubilant: 28Oct04	<b>sox red world</b> boston series June --	Red Sox win
909	4	sympathetic: 27–31Dec04	<b>tsunami dead sri death toll</b> earthquake disaster thousands bodies lanka ocean imagine natural thailand goes sri affected waves countries sri	Natural disasters
527	2	sad: 1Feb03	<b>space shuttle columbia</b> lost challenge crew ----- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --	Shuttle Columbia
264	1	scared: 30Oct–01Nov04	<b>bush vote kerry election</b> service abortion draft bill voting vote agree not children tax support health military vote choice money	Election 2004
140	1	stressed: 5–07Oct04	<b>cheney debate edwards country</b> vice second answer vote outside president experience watching listen lack presidential interview enjoy offer involved relax	Election 2004
130	10	sad: 11–12Sep02	<b>lost remember world september</b> lives bless watch watching ones forget hit united stand american real nation died america country meet	9/11 anniversary
71	4	nervous: 30Aug–02Sep04	<b>hurricane</b> frances house hit florida miss mississippi sds coming deal storm ty miss south miss miss miss miss miss miss	Natural disasters
67	30	scared: 26–31Oct03	<b>fire san fires california</b> live sri homes smoke deep house miss family mississippi miss miss miss miss miss miss miss	California wildfires

Table 5.2: Top 10 bursty events, according to the Google Timeline results, accompanied by CNN’s annual ranking, detected from the bursty time of moods.

costs. The output is an optimal sequence state  $q_1^*, \dots, q_T^*$  where  $q_t^* = 1$  implies burst at time  $t$ . We refer readers to Kleinberg [2003] for further details and the underlying stochastic explanation for this algorithm.

#### 5.4.1.1 Incremental burst detection

Since data in the blogosphere is normally large and streams on-line, the need for a mechanism for detecting bursty periods incrementally for moods arises. While the KLB approach works well in practice, it is not suitable for large-scale and on-line data since, in addition to requiring a multiplier  $\lambda$  to be set manually, it requires advance knowledge of the entire data to determine the ratio of emitting relevant documents during non-bursting periods  $\pi_0$  and the transition cost  $\gamma$ .

We modify the KLB to operate on-line: the parameter  $\pi_0$ —the rate the automaton emits relevant blog posts during non-bursting periods—is calculated incrementally; that is,  $\pi_0$  is defined as  $R_t/N_t$ , where  $R_t$  is the number of relevant blog posts until the time slice  $t$  and  $N_t$  is the total number of blog posts until the time slice  $t$ . The

Mood	Bursty period	Topics
P*ssed off	11Sep – 13Sep	<b>world</b> trade died center war country terrorists lost united states lives american die innocent attack government buildings live middle pentagon
Shocked	11Sep – 15Sep	<b>world trade plane center</b> crashed planes pentagon towers attacks twin york attack president bush terrorist washington united south flight american
Angry	11Sep – 16Sep	<b>world plane planes center</b> trade crashed towers pentagon tower twin scared tuesday low canton hour wtc hijacked tv building hit
Sympathetic	11Sep – 19Sep	<b>world lost pray families lives</b> goes peace country family wish terrible died stop victims stand helping terrorism horrible loss tragedy
Worried	11Sep – 21Sep	<b>war world</b> big america city end lost country lives president pray nation ones bush scared heard died scary planes fact
Enraged	11Sep – 25Sep	<b>lost heart family nation center wtc</b> pentagon planes buildings war prayers act lives thoughts trade lucky reason race crashed hell

Table 5.3: The set of moods found bursty in the 9/11 event and the related topics found in the blog posts tagged with the bursty moods in the bursty periods.

transition cost  $\gamma = \ln(t+1)$  is also incrementally calculated according to  $t$  instead of  $T$ . The pseudo code for detecting incrementally bursty intervals of moods is given in Algorithm 5.1 (iKLB), where  $\text{Binomial}(n, r, \pi) = \binom{n}{r} \pi^r (1-\pi)^{n-r}$  is the usual binomial probability mass function.

#### 5.4.1.2 Bursty event detection

For each mood  $m$  in the set of 132 predefined mood labels, we perform burst detection using three approaches: THD, KLB and our modification, iKLB. Using the decoded state sequence, a burst is declared when the decoded state transits from low to high. Each bursty interval of given mood  $m$  is deemed to be associated with a real-world event; and to infer *what* the event is about, we retrieve the blog posts tagged with  $m$  during the bursty period and apply topic model LDA to infer the topics, subject to the evaluation method presented in Section 5.2.

## 5.4.2 Experimental results

Among 132 predefined mood labels, 76 are bursty (that is, there is at least one burst detected with the corresponding mood label), resulting in 326 intervals. A cloud visualisation of these mood labels is shown in Figure 5.9a and a demonstration system is provided (see Appendix A.2). We learned topics from blog posts tagged with the bursty mood during the bursty period and then applied Google evaluation (see Section 5.2) to rank the events and topics. Our detection results are validated as being well correlated with real-world events. For example, we found that many events detected were listed in the top 10 CNN stories, as shown in Table 5.2. All top 10 events detected by our method are in the annual list of top 10 CNN events, except for the last event. (However, it is in the list of the top 30 events ranked by the CNN poll).<sup>6</sup>

### 5.4.2.1 Inferring impact of mood labels for event extraction

Table 5.2 shows that high-ranked events tend to be associated with mood labels whose sentiment reflects extreme behaviour, such as *shocked*, *sad* or *angry*. This suggests that the impact of mood labels is not the same when inferring the significance of the derived events. For example, on the 9/11 attacks event, only six moods are found bursty (see Table 5.3). On the other hand, *drunk*, *cold* and *hot* are found bursty in many intervals (see Figure 5.10), but their associated events are not found in the top 50.

Using the top 50 events ranked by our Google evaluation method as the reference, we extract 35 moods that are deemed suitable for the task of event extraction (see Figure 5.9b). Among these moods, *angry*, *sad*, *scared* and *worried* are each responsible for three events from the top-50 list; *depressed*, *distressed*, *enraged*, *infuriated*, *nauseated*, *p\*ssed off* and *sympathetic* are responsible for a further two events. Consistent with the findings in Section 5.3, a majority of the moods indicating significant events when bursty are low in valence and high in arousal. High-valence moods—*ecstatic*, *awake* and *jubilant*—are effective at detecting happy events, for example the release of the Harry Potter books or the success of the Red Sox. Likewise, *hopeful* and *opti-*

---

<sup>6</sup><http://edition.cnn.com/specials/2003/yir/reader.poll.html>, retrieved September 2011.

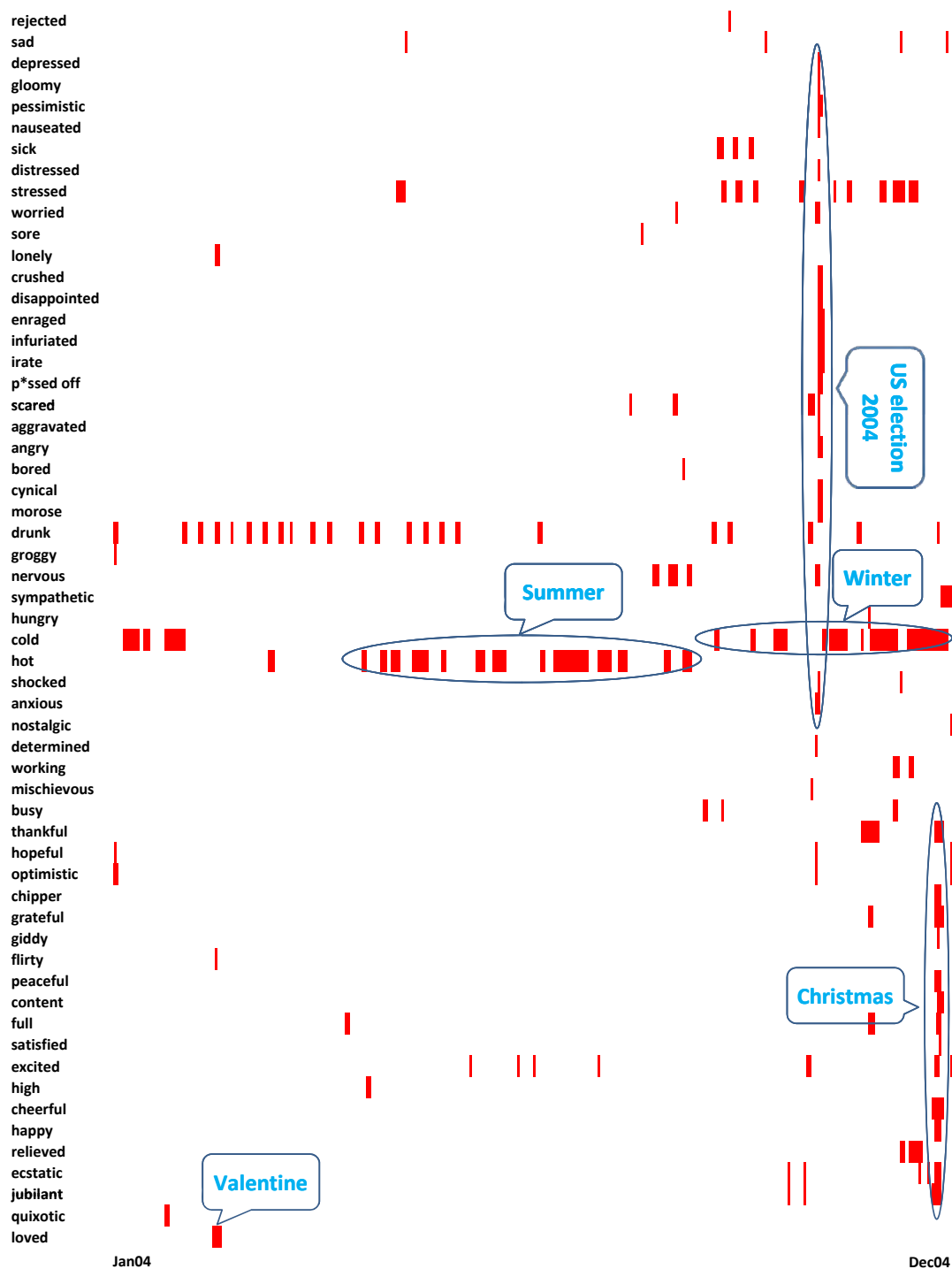


Figure 5.10: Bursty moods (sorted by their valence increasingly from top to bottom) and their bursty tagged intervals in 2004.

*mistic* prove effective for detecting controversial events, for example the US election, 2004.

For comparison, we have applied THD methods with manually adjusted THD so that 10 bursts are detected for each mood label. The results shown in Figure 5.9c are consistent with the findings in the KLB approach: that moods better at indicating significant events are also focused and that a majority of them are low in valence and high in arousal. High-valence moods indicate that important events are *ecstatic* (for ‘Red Sox win World Series’) and *hopeful* (for ‘US election’).

#### 5.4.2.2 Top events and burst detection algorithm comparison

To perform a comprehensive evaluation of the proposed algorithms and the quality of the events discovered, we further construct an evaluation dataset on the events as follows: 10 peak time periods from each mood (with a total of 132) are selected, then corresponding blog posts for each time period within each mood are gathered for topic analysis by applying LDA. The number of topics selected is proportional to the size of the vocabulary—typically 10. For each topic learned by LDA, the top five words and an extended time period (peak period  $\pm 1$  day) were submitted to Google Timeline to get the number of pages discussing the topic within the specified time period. We then ranked them using the Google volume results and selected the top 70 events as the representative events for our evaluation. In addition, we manually examined each event and labelled the corresponding real-world events. These events and the capacity of both KLB and iKLB to detect them are shown in Tables 5.4 and 5.5.

From 326 bursts detected by KLB, we found that 53 intervals match with the time of the 53 top events. Meanwhile, running iKLB on the set of 132 moods resulted in 329 bursts, among them 55 intervals matched with 55 top events, suggesting a close performance between the offline and the proposed incremental method. The fact that a high proportion (more than 75 per cent) of the top 70 events manually constructed was detected by the algorithm suggests that the proposed scheme performs well.

Score	Mood	Peak Period	Topic words	KLB	iKLB
7740	sympathetic	11-15/Sep/01	world lives trade hit ones	hit	hit
7050	sad	01/Feb/03	shuttle space mission nasa columbia	hit	hit
6370	worried	11-13/Sep/01	world trade planes plane towers	hit	hit
6260	scared	11-16/Sep/01	world trade plane center towers	hit	hit
5410	p*ssed off	03/Nov/04	bush kerry election vote percent	hit	hit
5150	enraged	11-13/Sep/01	world trade center crashed planes	hit	hit
4730	angry	11-16/Sep/01	world trade plane center planes	hit	hit
4370	sad	11-15/Sep/01	world trade plane center planes	hit	hit
4370	shocked	11-15/Sep/01	world trade plane center planes	hit	hit
4230	contemplative	20/Mar/03	war iraq world country live	miss	miss
3880	cynical	03/Nov/04	bush kerry vote election voted	hit	hit
3120	nervous	02-03/Nov/04	bush president war country america	hit	hit
2500	crushed	03-04/Nov/04	bush country kerry vote election	hit	hit
2390	shocked	01/Feb/03	shuttle space nasa crew columbia	miss	hit
1490	disappointed	03-04/Nov/04	war american bush iraq world	hit	hit
1220	disappointed	20/Mar/03	war support bush countries country	miss	miss
1090	enraged	03/Nov/04	election votes florida electoral states	hit	hit
972	pessimistic	03/Nov/04	bush voted country kerry vote	hit	hit
963	angry	03-04/Nov/04	bush country vote voted kerry	hit	hit
883	pessimistic	11-12/Sep/01	center trade started lives nuclear	miss	miss
843	crushed	12-13/Sep/01	america live coming innocent united	hit	hit
842	cynical	20/Mar/03	oil war gets united world	miss	miss
772	grateful	25-26/Nov/04	turkey family thanksgiving food pie	hit	hit
710	grateful	11-12/Sep/02	september events remember lives died	miss	miss
709	angry	20/Mar/03	military international small sites world	hit	miss
699	worried	20/Mar/03	war iraq return world start	hit	hit
288	grateful	27/Nov/03	family thanksgiving food coming big	hit	hit
285	cheerful	22-27/Dec/04	christmas house mom family merry	miss	miss
264	grateful	25-26/Dec/02	couple buy christmas bought wanted	hit	hit
216	thankful	27-28/Nov/03	family thanksgiving especially matter big	hit	hit
203	thankful	28-30/Nov/02	family kind thanks especially light	hit	hit
185	nervous	20/Mar/03	war iraq seeing coming iraqi	miss	hit
176	sympathetic	11/Sep/03	terrorism york left gone lives	miss	hit
174	cheerful	23-26/Dec/01	christmas house mom family eve	miss	miss
163	p*ssed off	11-13/Sep/01	united states stop american act	hit	miss
146	full	27-28/Nov/03	thanksgiving turkey family ate house	hit	hit
130	thankful	25-26/Dec/03	christmas merry family mom gave	hit	hit
120	sympathetic	11-12/Sep/02	knew country died events free	hit	hit
119	scared	20/Mar/03	war world country end away	hit	hit
110	full	25-26/Nov/04	thanksgiving turkey family food eat	hit	hit

Table 5.4: The capacity of KLB and iKLB to detect the top 70 events.



Score	Mood	Peaky Period	Topic words	KLB	iKLB
109	cheerful	20/Apr/03	easter friday saturday mom gave	miss	miss
107	grateful	25-26/Dec/01	gift gets dad opened bought	hit	hit
102	contemplative	11-12/Sep/02	remember world watched watching hit	hit	hit
102	sad	30/Nov/01	george harrison beatles died beatle	miss	miss
90	thankful	25-26/Dec/02	christmas family gifts lots wants	hit	hit
88	nervous	11/Sep/02	died lives bombings pray afew	miss	miss
84	cheerful	24-26/Dec/03	christmas merry family mom house	hit	hit
70	jubilant	24-25/Dec/04	christmas merry santa holidays holiday	hit	hit
69	hungry	27/Nov/03	thanksgiving eat food family turkey	hit	hit
65	grateful	25-26/Dec/04	christmas family merry mom house	hit	hit
64	grateful	25-26/Dec/03	christmas merry family mom presents	hit	hit
59	thankful	25-26/Dec/04	christmas merry family house mom	hit	hit
56	thankful	24-26/Nov/04	turkey thanksgiving pumpkin potatoes mashed	hit	hit
54	thankful	22-23/Nov/01	thanksgiving family big watching lose	hit	hit
49	sad	11-12/Sep/02	remember world september lives bless	hit	hit
48	hungry	25/Nov/04	turkey thanksgiving eat gobble food	hit	hit
44	cheerful	24-26/Dec/02	christmas merry family presents mom	hit	hit
44	hungry	22/Nov/01	turkey thanksgiving food dinner christmas	miss	hit
40	jubilant	25/Dec/03	letter write personal lyrics christmas	hit	hit
32	sad	12/Sep/03	ritter simple television actor star	hit	hit
30	sad	09/Dec/04	dimebag darrell pantera shot metal	hit	miss
28	jubilant	24-25/Dec/02	started xmas real big beat	hit	miss
28	sad	27/Feb/03	rogers fred died mister neighborhood	hit	hit
25	cheerful	25/Nov/04	thanksgiving turkey family eat gobble	miss	miss
25	full	28-29/Nov/02	thanksgiving holiday holidays celebrated christmas	hit	hit
23	thankful	25/Dec/01	christmas merry family school big	miss	hit
21	shocked	22/Jun/01	hooker lee songs albums rolling	miss	hit
19	full	22-23/Nov/01	turkey thanksgiving eat house dinner	hit	hit
18	grateful	22-23/Nov/01	thanksgiving family words james feels	hit	hit
17	full	26/Dec/04	christmas family merry mom dad	hit	hit

Table 5.5: The capacity of KLB and iKLB to detect the top 70 events (continued from Table 5.4).

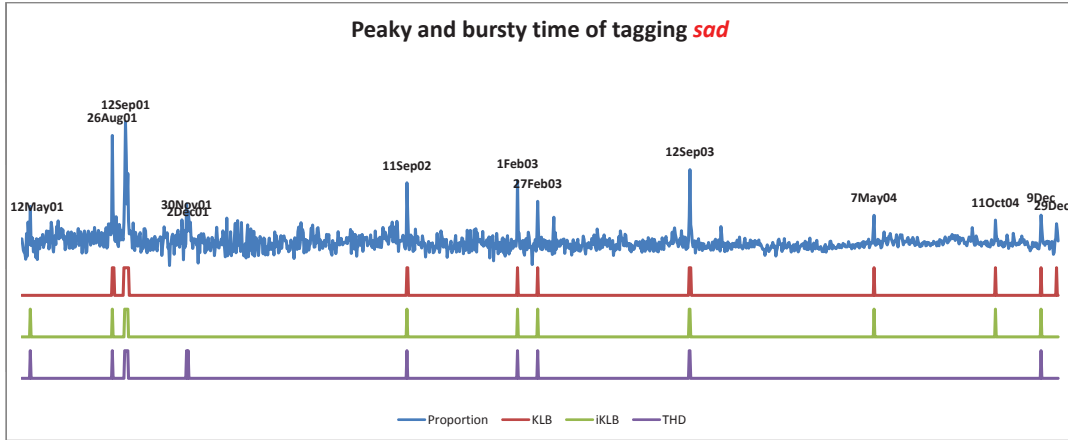


Figure 5.11: Highest points (from the normalised signal of the count of blogs tagged with mood *sad*) and burst detection results for mood *sad* from three burst detection algorithms. For THD, we manually adjust the threshold so that 10 bursts are detected.

Bursty time	Score	CNN	Topics	Related events
26–28Aug01	28	30	aaliyah plane died family young death bar dead <small>remember music heard huge star brother ... hope day ... ..</small>	Aaliyah's death
10–16Sep01	4420	1	world trade plane center planes crashed york <small>pentagon heard hit buildings second united school building ... .. american attack tower towers</small>	September 11
11–12Sep02	130	10	lost remember world september <small>lives bless watch watching ones forget hit united stand american real nation died america country heart</small>	9/11 anniversary
1Feb03	527	2	space shuttle columbia lost challenger crew ... <small>astronaut train ... ..</small>	Shuttle Columbia
27Feb03	30	30	rogers fred died neighborhood master passed childhood <small>beautiful watching ... ..</small>	Mr. Rogers dies
11–13Sep03	39	30	johnny cash died music ritar back ... <small>... ..</small>	Johnny Cash dies
7May04	11	30	episode end miss rachel watching ross watch watched favorite <small>ended finale goodbye bye cried left joey final cry series season</small>	End of "Friends"
11Oct04	38	30	reeve movie ... <small>... ..</small>	Christopher Reeves dies
9Dec04	33	nil	dimebag darrell pantera shot metal rip <small>killed guitar band galena denzepam music ... ..</small>	Dimebag Darrell dies
29Dec04	69	4	tsunami death lost toll red <small>disaster jeg relief asia lives sri earthquake cross world donate thailand thousands aid dead bodies</small>	Natural disasters

Table 5.6: The bursty periods associated with the mood *sad* detected by KLB and the corresponding events, accompanied by the Google Timeline results and annual ranking by CNN. Only the top-ranked topics (by the Google Timeline results) are shown.

### 5.4.2.3 Extreme moods and burst detection algorithm comparison

As shown in Section 5.4.2.1 (see Figure 5.9b), *sad* represents an extreme form of emotion and is effective for extracting events that are significant. To gain further understanding of the burst structure for this mood label, we present the results from three burst detection algorithms: THD, KLB [Kleinberg, 2003] and iKLB. The results are shown in Figure 5.11.

We note that KLB performs well on the detection of bursty events. Based on 10 bursty periods of *sad* moods detected by KLB, we extract 10 corresponding events, extract topics for them and rank these according to the method outlined in Section 5.2. The results are shown in Table 5.6 and we note the significance of the performance given that many of these events are listed in the CNN top stories.

For THD, over 10 bursty periods of *sad*, seven events also found by KLB are detected. Three events are missed: the end of the TV series Friends, Christopher Reeve's death and the Indian Ocean earthquake. This method finds an event on 30 November 2001 (the death of George Harrison, lead guitarist of The Beatles) and presents two false alarms on 12 May 2001 and 2 December 2001 when no significant events are found to have occurred. This simple approach works reasonably well on the task. However, it requires advance knowledge of the whole corpus to determine the threshold.

For the iKLB algorithm, over 10 bursty periods of *sad*, only one event found by KLB (the Indian Ocean earthquake) is missed. This shows that Kleinberg's model can be efficiently adapted as presented in this chapter to handle streaming data to detect salient burst structures in blog data in real-time without compromising the performance. A demonstration system of the bursts and events detected together with their topics and named entities is available (see Appendix A.2).

## 5.5 Conclusion

This chapter presented another global view on sentiment movement in the blogosphere. Real-world events often involve shared emotional responses from a large population. We have investigated the novel problem of leveraging the blogosphere as a sentiment ‘sensor’ to extract events. We have experimented on a large-scale dataset, using millions of blog posts ground truthed with mood labels, and presented a novel method for evaluating the extracted topics and events. We formulated a high-order emotional measure, termed SI, and presented methods for extracting events based on these time-series signals. The results extracted are consistently well matched with lists of top stories as voted by CNN. This demonstrates the usefulness of this sentiment-based approach.

Next, we proposed methods for extracting bursty events by adapting a well-known burst detection algorithm for use on a large-scale dataset. In this setting, data is incrementally added, introducing an on-line implementation for the algorithm. It complements the SI-based approach and the events extracted from the burst detection are specific to individual mood labels at finer levels of analysis of real-world events. After ranking by Google, several of the detected events were found to coincide with the list of CNN top stories. Our proposed approach is a generic framework for sentiment-based event extraction and can be applied to other forms of textual datasets.

## Chapter 6

# Egocentric Aspect in Social Media

In this chapter we investigate another important component of user-generated social media: the users, also known as the egocentric aspect. In particular, we analyse the impact of demographics and personality, such as age and social connectivity, on a number of manifest properties of textual messages of a large corpus of blog posts. Sentiment information, in terms of author mood, is introduced to the analysis of the egocentric aspect of social media. It is compared with two conventional facets of content, topics and psycholinguistics, in the role of features for a classification. We build binary classifiers for old versus young and social versus solo bloggers using these properties as features. They are found to be significantly different across the examined cohorts, suggesting that the content people publish in their on-line diaries is adequate to predict aspects of their profiles. Analysis of discriminative features reveals that age turns upon choice of topics whereas good prediction is achieved for social connectivity using linguistic features. On the other hand, the result obtained by the sentiment approach in all classifications is encouraging at a lighter computational cost, suggesting a potential method for egocentric analysis in social media.

## 6.1 Digital Presence in Social Media

No longer merely readers, users in social media have taken on the role of broadcasters. This inspires people in the new media to publish their identities (or digital presence) for many reasons, including attracting others to their channels. Social media messages can be massive, diverse in topic and style and enriched by user- and system-supplied metadata; all of which allow for statistically significant analysis of linguistic style, demographic properties, sentiment and social connectivity. Insights gained from such analysis have wide application, such as in search applications, business and government intelligence and sociology—in which the web is viewed as a very large community sensor.

A number of approaches have been proposed in the literature, especially in psychology, for analysing the relationship between the textual content of blog posts and the demographic properties and personality of the authors. Studies conducted from the perspective of psychology tend to be on a very small scale, due to the time and cost required to administer them. In contrast, data-driven analyses tend to be restricted to simplistic methodologies, such as word counting. Therefore, there is a need for large-scale analyses that make use of, among other things, advances in probabilistic topic modelling for characterising the content of social media messages. Textual content can also be represented using richer models of language, such as LIWC [Pennebaker et al., 2007b], which categorises words according to a number of linguistic and psychological processes (see Section 2.5.4). Another aspect of text capable of predicting demographics or personality is its sentiment information, which is known to correlate with personality type. For example, *extraverts* have been found to experience more pleasant and less unpleasant moods than do *introverts* [Rusting and Larsen, 1995]; extraverts in a good mood have been found to be more creative than introverts [Stafford et al., 2010].

A good prediction of the demographic properties of on-line users, such as age (old/young) or gender (male/female), has potential application for viral marketing, where companies aim at specific customers, for example, young women, for their products.

On the other hand, a person's social linkage can be used as a proxy to estimate their

capacity to influence others, as well as their personality. Specifically, we take a user's in-degree (henceforth termed *followers*) as a proxy for influence, and out-degree (henceforth termed *friends*) and community membership as proxies for extraversion-introversion.

The notion of a person's *influence* has received much attention with the rise of on-line social networks, and the ready data they provide. Intuitively, someone who is potentially influential should have many followers (that is, others who monitor their actions or pronouncements). Bakshy et al. [2011] have demonstrated that Twitter users having many followers have a large influence in terms of forcing the diffusion of a URL, hence driving a greater volume of attention to the given site. Finding influentials is a key task in viral marketing, which co-opts social networks in marketing a brand or product [Phelps et al., 2004]. Influentials likely cause information about new products to spread more quickly than do non-influentials.

While influence is concerned with the effect of a person on a network or its information flow, *personality* is related to users themselves. Personality might also be inferred from social-linkage. For example, Schrammel et al. [2009] found that individuals scoring high on traits of extraversion have more friends than those with low scores. Ross et al. [2009] found that extraverts are members in significantly more groups than are introverts.

Estimation of those who are most influential in a social network is very useful in viral marketing, in which marketers are concerned with placing the maximum investment in consumers likely to drive the diffusion of positive opinion further. There is a role for personality prediction in the implementation of personalised information retrieval applications, or user-targeted advertising, in which the most suited advertisement from a set of potential advertisements can be delivered to a particular user.

The remainder of this chapter is organised as follows. The hypotheses being evaluated in this chapter are described in Section 6.2. Experimental setup is explained in Section 6.3, including the corpus and various sub-corpora created for this study. Section 6.4 presents prediction results for the age and social connectivity of bloggers. Section 6.5 presents an analysis of the correlation between topic and linguistic style and the age and social connectivity of bloggers. Section 6.6 concludes the chapter and notes prospects for future work.

## 6.2 Hypotheses

We examine how the age and social connectivity of bloggers correlate with the affective information conveyed in their texts, through the current mood tags. The topical and psycholinguistic features of the blogger-authored text is also analysed in relation to their demographic and connection properties. Specifically, we test the following hypotheses:

***Hypothesis 1:*** Posts written by old versus young bloggers will exhibit significant difference in topic, psycholinguistic features (LIWC) and mood.

***Hypothesis 2:*** Posts written by bloggers having small versus large degrees of social connectivity will exhibit significant difference in topic, psycholinguistic features (LIWC) and mood.

We test these hypotheses by examining the performance of classifiers built for each of the two dichotomies. We then report the statistically different features between the two classes, using the *Mann-Whitney U* test.

## 6.3 Experimental Setup

### 6.3.1 Dataset

We use the CHI06 dataset, described in Section 3.3.2, for these experiments. We augment this dataset with the profile information of bloggers and the blog posts for bloggers included in these experiments for the period May 2005 to the end of 2010. Only English posts are included in the dataset. As can be seen in Figure 6.2, most bloggers are Americans. To evaluate our hypotheses we construct the following two sub-corpora from this dataset.





Figure 6.1: An example of a user's profile related to his networking. This user has 79 friends, 2,140 followers and belongs to four communities.

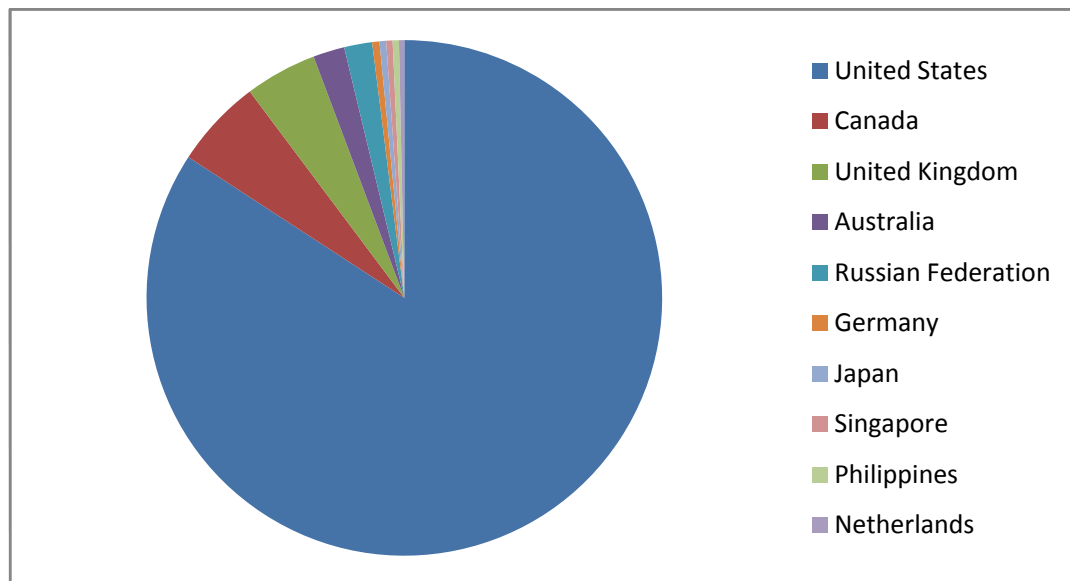


Figure 6.2: Top 10 countries Livejournal bloggers live in.

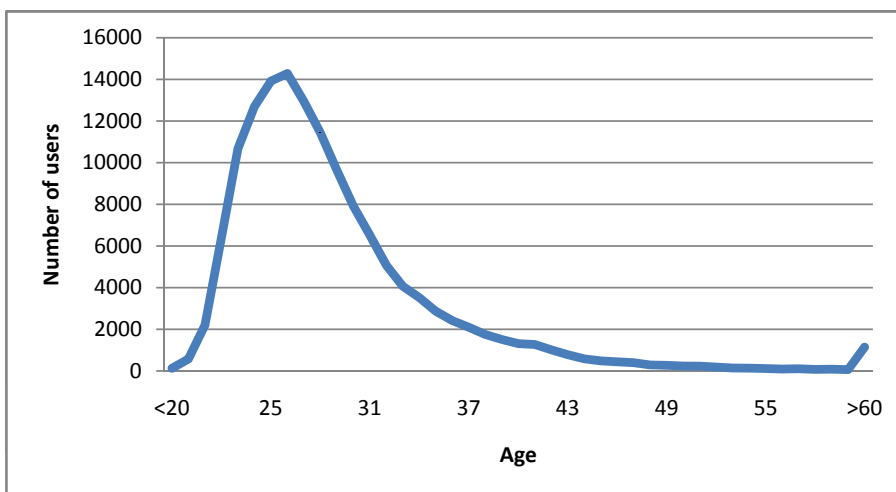


Figure 6.3: Age distribution of bloggers.

#### 6.3.1.1 Old and young blogger sub-corpus

The first corpus targets age difference. From Figure 6.3, it can be seen that almost all bloggers are aged between 20 and 60. We conjecture that differences in age translate to different choices of discussion topic and differing linguistic styles. This sub-corpus is created by selecting users at the extremes of the age spectrum. The *old* category is formed by successively adding bloggers aged 60 and down, to a cut-off of 5,000 bloggers, resulting in an age range of 39 to 60, while the *young* category is created by adding users aged 20 and up, to a cut-off of 5,000 bloggers, resulting in an age range of 20 to 22. Only users posting not less than 20 posts are included.

#### 6.3.1.2 Social connectivity corpora

The second sub-corpora targets social connectivity. Livejournal supports one directed person–person link type. For a given user, we term incoming links *followers* and outgoing links *friends*. Livejournal also allows users to join communities that discuss topics of interest. Figure 6.4 contains plots for the per-user distribution of friends, followers and groups. On average, each blogger has 10 followers, 23 friends and joins seven communities. An example of these participation networks is shown in Figure 6.1. Three sub-corpora are built from bloggers with extreme numbers of followers, friends and/or community membership. For each of the following sub-

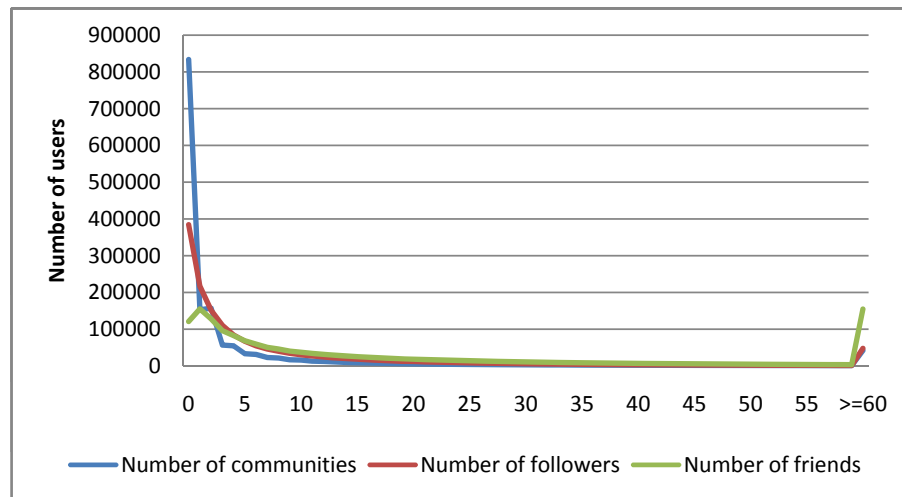


Figure 6.4: Distributions of the number of communities a blogger joins and the number of followers and friends a blogger has.

corpora, bloggers with many friends, followers or communities are categorised as *social* and those with few friends, followers or communities are termed *solo*.

- ***Based on number of followers***

- For this sub-corpus, the social category was created by collecting bloggers having 150 followers and less, to a cut-off of 5,000 bloggers. In addition, bloggers in this category had to have been active for less than one year. The resulting range of followers was 78 to 150. The solo category was created from 5,000 bloggers with only one or two followers, who had to have been active for more than one year.

- ***Based on number of friends***

- This sub-corpus was created similarly to the followers-based sub-corpus. The social category includes 5,000 bloggers, active for less than one year, having a number of friends ranging from 108 to 150. The solo category includes 5,000 users, active for more than one year, having between one and three friends.

- ***Based on number of communities joined***

- For this sub-corpus, the social category was created with a cut-off of 5,000 bloggers, active for less than one year, having joined a number

of communities ranging from 56 to 150. The solo category was created from 5,000 users, active for more than one year, having joined one to two communities.

To avoid spam and noisy data, the upper bound for the number of links is 150 (according to Dunbar’s number [Dunbar, 1993]: a quasi-limit on the number of meaningful relationships a person can have). In addition, only users posting not less than 20 posts are included.

	<i>Social</i>	<i>Solo</i>
<b>Number of followers based</b>		
Number of friends	144.4 (0.9)	39.1 (0.5)
Number of communities	68.5 (0.9)	9.2 (0.2)
<b>Number of friends based</b>		
Number of followers	37.1 (0.5)	9.5 (0.3)
Number of communities	49.9 (0.7)	4.9 (0.3)
<b>Number of communities based</b>		
Number of friends	88.1 (0.8)	17.1 (0.2)
Number of followers	29.2 (0.4)	8.7 (0.1)

Table 6.1: The mean and standard deviation (in brackets) of connectedness variables between *social* and *solo* bloggers categorised by another connectedness variable.

There are high correlations between the in-degree (number of *followers*), out-degree (number of *friends*) and community membership. Table 6.1 shows the means of these variables. If a user has many followers, he or she will also have many friends and join many communities. Alternatively, a blogger with few followers will have few friends and join few communities.

### 6.3.2 Feature sets and classifiers

Three types of feature are used to perform the binary classifications corresponding to the hypotheses listed above.

- The first type is based on the topics discussed in blog post content. We use LDA [Blei et al., 2003] to learn the latent topics from post text (see Section

2.5.5) and Gibbs sampling is used to learn 50 topics for each sub-corpus, with 1,000 samples for burn-in, followed by 5,000 iterations, ultimately yielding topic mixtures for each blogger.

- The second type of feature is drawn from the LIWC (see Section 2.5.4). The LIWC 2007 package and its standard dictionaries [Pennebaker et al., 2007b, Tausczik and Pennebaker, 2010] are used to extract statistics for the 68 language groups.
- The third type of feature is mood, as manually annotated by the blog authors.

As shown in Table 6.1, the features of social connectivity are highly correlated. This potentially helps to predict the connectivity of bloggers when other networked information is available. Then, the remaining two connectedness variables are used as features to predict whether a user will be social or solo, which is categorised by another connectedness variable.

Classification is performed by SVM and logistic regression to predict age and the degree of social connectivity of bloggers (see Section 2.4.1). 10-fold cross-validation is conducted for each SVM classifier and each logistic regressor. The average F-measure of the predictions is reported.

Feature sets that perform well in classification are potentially useful for defining and interpreting the differences between the categories. Statistical tests are carried out to evaluate the hypotheses. Since all the features in all polar categories are found to be not normally distributed (Kolmogorov–Smirnov tests with Lilliefors significance correction,  $ps < .001$ ), the differences are evaluated using a non-parametric statistical test—Mann–Whitney U. All statistical tests are conducted using IBM SPSS Statistics 17.0.

## 6.4 Prediction Results

Since *number of friends*, *number of groups*, and *number of followers* are highly correlated, *solo* versus *social* bloggers are well predicted using the other connectivity features. To be precise, *number of followers* and *number of groups* are powerful for

*none* or *many friends* prediction, gaining an F-measure of 78.6 per cent; *number of friends* and *number of groups* are good indices for predicting *none* or *many followers* bloggers (81.7 per cent F-measure); and *number of friends* and *number of followers* achieve an F-measure of 80.4 per cent when predicting *none* or *many groups* bloggers.

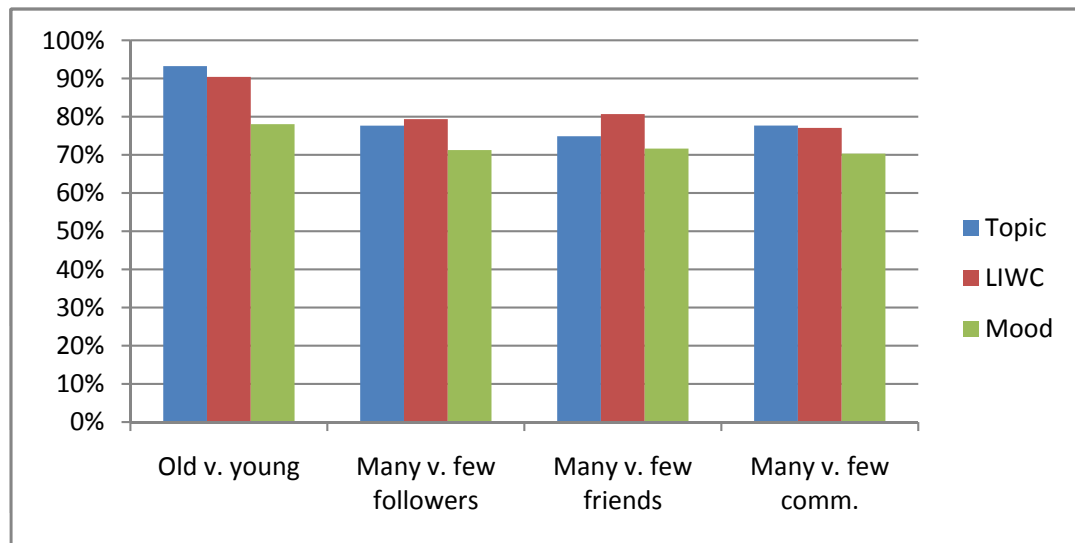
However, connectivity information is not always available in social media. In these cases, linguistic style, topic and mood are employed for prediction. Classification results (in F-measure) for both SVM and logistic regression using these features are presented in Figure 6.5. It can be seen that performance was not affected by classifier type. In short, classification of blogger age (old versus young) achieved an F-measure above 90 per cent, while classifier F-measure for all other categories was close to 80 per cent.

The topics discussed in the content are good features for classifying old and young bloggers, achieving an F-measure of more than 90 per cent. The linguistic styles, through LIWC groups, are also effective in predicting users by age group (90 per cent). Using moods as features in age-group predictions gains an F-measure of nearly 80 per cent, a promising result given that a supervised feature selection stage is not required as it is in the topic case,.

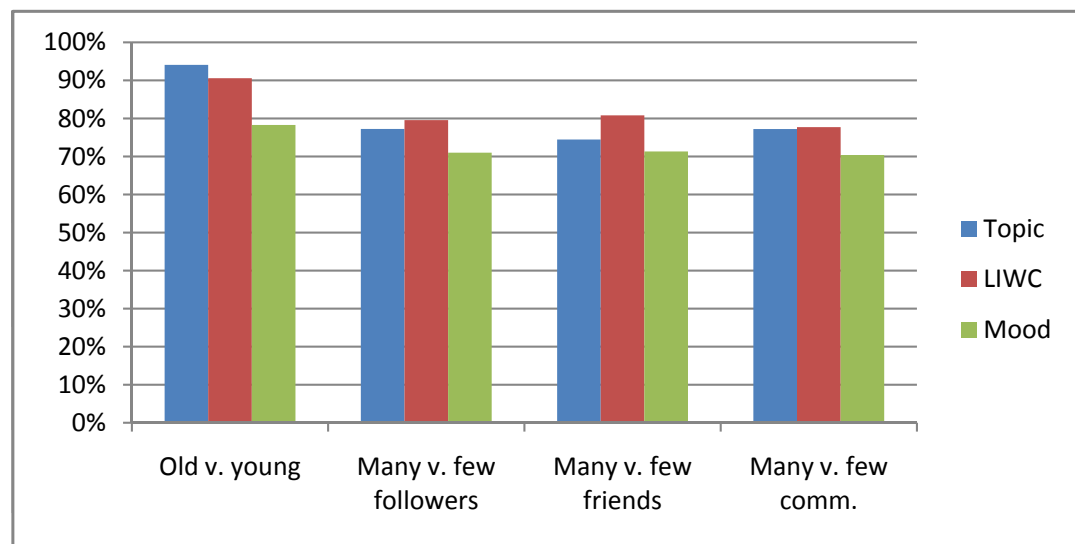
In contrast, *solo* versus *social* bloggers are well predicted using LIWC features. For example, prediction of *none* or *many friends* bloggers using LIWC features alone achieves an F-measure of 81.7 per cent. The best result (in F-measure) of predicting *solo* and *social* bloggers by topics is approximately 77 per cent, while that by moods is higher than 70 per cent—an encouraging outcome.

## 6.5 Demographic and Personality Impact

As shown in Section 6.4, topic, linguistic style (as represented by the LIWC's categories) and mood are useful features for discriminating age and social connectivity.



(a) SVM classifiers.



(b) Logistic regressions.

Figure 6.5: Prediction performance (F-measure).

Topics in favour of the old					Topics in favour of the young					
<b>house</b>	<b>stuff</b>	<b>bit</b>	<b>job</b>	<b>car</b>	<b>years</b>	<b>place</b>				
working	big	couple	room	bed	started	left	money			
start	weeks	dinner	10	saturday						
<b>bush</b>	<b>president</b>	<b>war</b>	<b>state</b>	<b>american</b>						
country	states	government	vote	gay	world	america				
church	law	iraq	court	rights	political	henry	popo			
<b>movie</b>	<b>book</b>	<b>story</b>	<b>film</b>	<b>character</b>						
read	series	characters	books	episode	star	season	fic			
movies	game	stories	writing	favorite	scene	reading				
<b>chicken</b>	<b>cheese</b>	<b>chocolate</b>	<b>sugar</b>							
sauce	cup	bread	cream	butter	water	flavoured				
lemon	cake	garlic	eggs	salt	milk	rice	add	pan		
<b>world</b>	<b>away</b>	<b>mind</b>	<b>years</b>	<b>read</b>	<b>end</b>					
kind	heart	point	care	fact	wrong	course				
place	real	words	feeling	hell	far	use				
<b>hearts</b>										
<b>favorite</b>	<b>sex</b>	<b>color</b>	<b>hair</b>	<b>nope</b>	<b>crush</b>	<b>black</b>	<b>movie</b>			
music	blue	current	food	sung	never	hesitated	to	phone	update	air
<b>anime</b>	<b>japanese</b>	<b>manga</b>	<b>chan</b>	<b>character</b>	<b>yo</b>	<b>quartzite</b>				
art	draw	add	gray	glock	en	x3	drawing	japan	reaction	uh
<b>school</b>	<b>mom</b>	<b>stuff</b>	<b>house</b>	<b>guess</b>	<b>cool</b>	<b>cause</b>	<b>dad</b>			
funny	left	class	awesome	okay	wanted	stared	miss	kinda	ridic	watched

Table 6.2: Differences in topic consideration by old and young bloggers ( $ps < 0.001$ ).

This indicates that the features in the blog posts are influenced by age and personality. Below we further examine differences in these features for the various sub-corpora with reference to U tests.

### 6.5.1 Difference of topic across age and social connectivity

Figure 6.5 reports that topic features enable classification of blogger age (old versus young) above 90 per cent. Table 6.2 includes ‘term clouds’ of topics that are used with significantly different frequency between the two groups, allowing further insight into the result.

Older bloggers write more about political matters (for example, Bush and the world), confirming the finding of Quintelier [2007] that political participation is comparatively lower for young people. The remaining topics provide an interesting picture of the contrast: for the older bloggers, house ‘stuff’, cooking, movies and books are of central concern, while for the young, it is love, school and Japanese *anime*.

Interestingly, the topics noted above as being characteristic of older bloggers are also associated with social bloggers, and those of interest to the young correlate with solo bloggers. From the term clouds in Tables 6.3a, 6.3b and 6.3c, it can be seen that social bloggers write more about entertainment (for example, books, songs, movies and animation), whereas solo bloggers focus particularly on school-related topics.





Further, those having more followers and friends write more about politics than do solo bloggers, which is intuitive when the social component of political participation is considered.

## 6.5.2 Difference of linguistic style across age and social connectivity

### 6.5.2.1 Old versus young bloggers

All LIWC features are found to be significantly different between old and young categories (Mann–Whitney U tests,  $ps < .05$  two-tailed,  $n_1 = n_2 = 5,000$ ), barring three exceptions: *future* ( $p < .882$ ), *wc* ( $p < .393$ ) and *funct* ( $p < .118$ ).

The differences in LIWC categories can be seen in more detail in Figure 6.6. We observe that older bloggers favour *greater-than-six-letter* words, which supports the finding of Pennebaker and Stone [2003] that the aged do not decline in using more complex language, whereas the young prefer to use *spoken*, informal language. The young appear to be more *self-focused* (using more first person singular pronouns), whereas the old favour third person plurals, which also accords with the finding of Pennebaker and Stone [2003] that the old are less self-focused. The young also use more *affective* and *negation* words, and less *article* and *preposition* words, which supports the finding of Tausczik and Pennebaker [2010] that *emotion* words are positively correlated with *negation* use and negatively correlated with *articles* and *prepositions*.

### 6.5.2.2 Social versus solo bloggers

Many LIWC features are found to be significantly different for the different social connectivity categories. Figures 6.7a, 6.7b and 6.7c graph the distinctive use of the LIWC categories for the three social connectivity sub-categories.

First, we make some observations that apply to the overarching *social* versus *solo* blogger categories. Social bloggers make more use of words of greater than six letters,

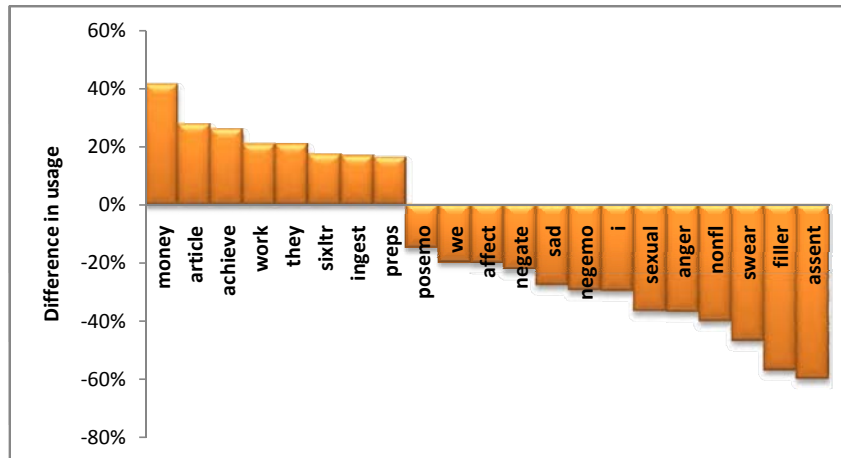


Figure 6.6: The LIWC features above the zero line are favoured by the *old*; otherwise, they are favoured by the *young* ( $ps < 0.001$ ).

which is referred to as complex language and a marker of presidential language [Slatcher et al., 2007]. Social bloggers use fewer words per sentence (*wps*) than solo bloggers. If we assume social bloggers are likely *extraverts* and solo bloggers are likely *introverts*,<sup>1</sup> this property accords with the finding of Mairesse et al. [2007] that *extraverts* use fewer words per sentence than *introverts*. In contrast, solo bloggers tend to use the *spoken* category (that is, assent, non-fluency and fillers). In addition, solo bloggers prefer to use *self-focused* (*i* group: for example, I, me and mine), *affective* and *swear* words, which together have implications for how acceptable their language is to others.

Social bloggers use more *articles* than solo bloggers. Normally, an article is followed by a concrete noun. Social bloggers also use less *pronouns* and *verbs*. Following the assumption above, this finding partly contrasts with that of Dewaele and Furnham [2000] who found that, in oral languages, *extroverts* use more *pronouns* and *verbs* and fewer *nouns* and *prepositions* than do *introverts*.

Bloggers that have many friends use more ‘good’ LIWC words, such as *leisure* or *achievement*, whereas friendless bloggers use more words of *negative* emotion. Similarly, users with few followers use more ‘bad’ LIWC words, including *sad*, *anger* and *anxious*.

<sup>1</sup>As defined in Freyd [1924], *introverts* are those who tend to withdraw from social contacts, whereas *extraverts* tend to make social contacts.

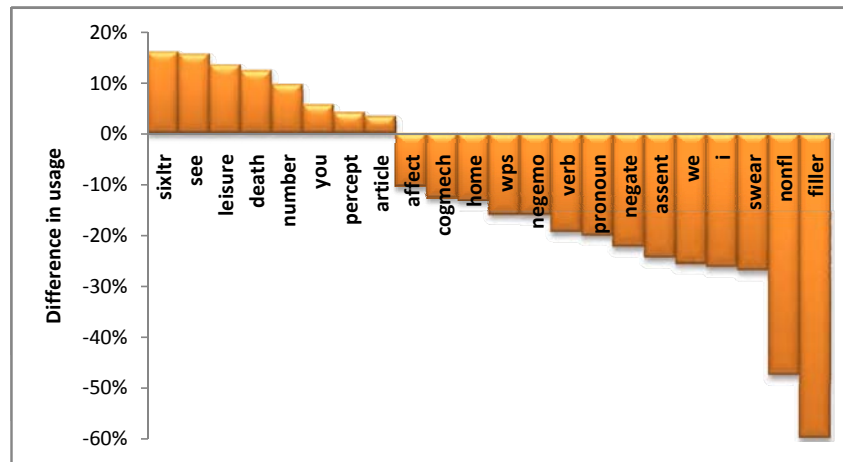
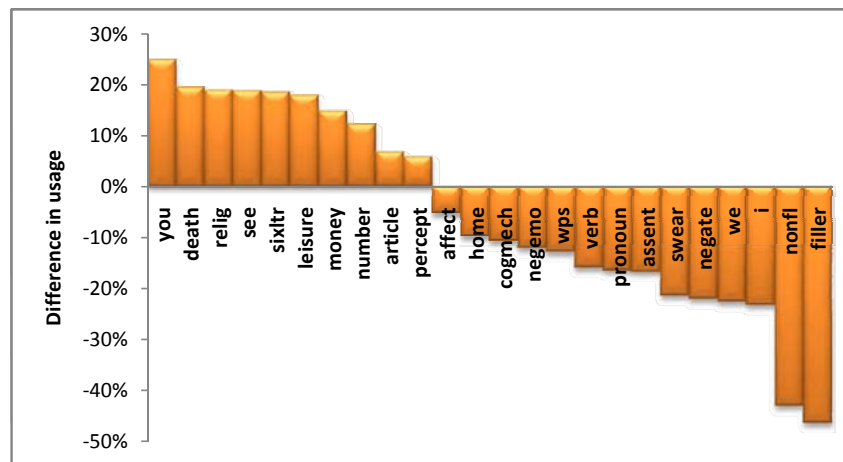
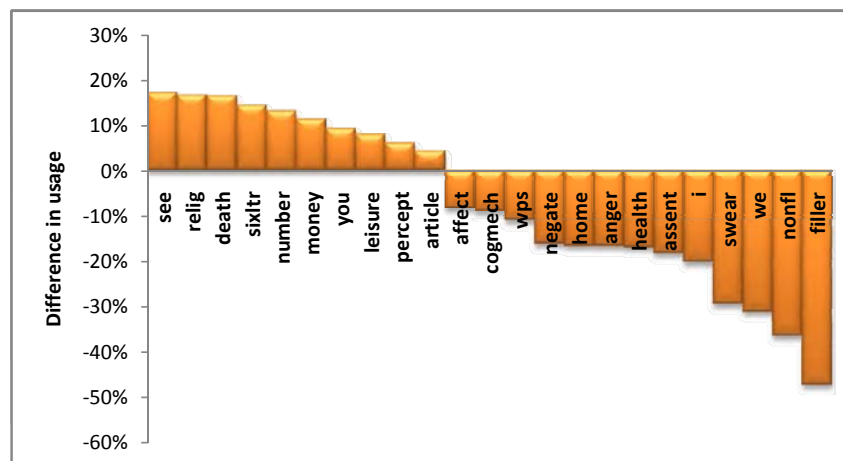
(a) *Many followers versus few.*(b) *Many friends versus few.*(c) *Many communities versus few.*

Figure 6.7: The LIWC features above the zero line are in favour of the *social*; otherwise, they are in favour of the *solo* ( $ps < 0.001$ ).

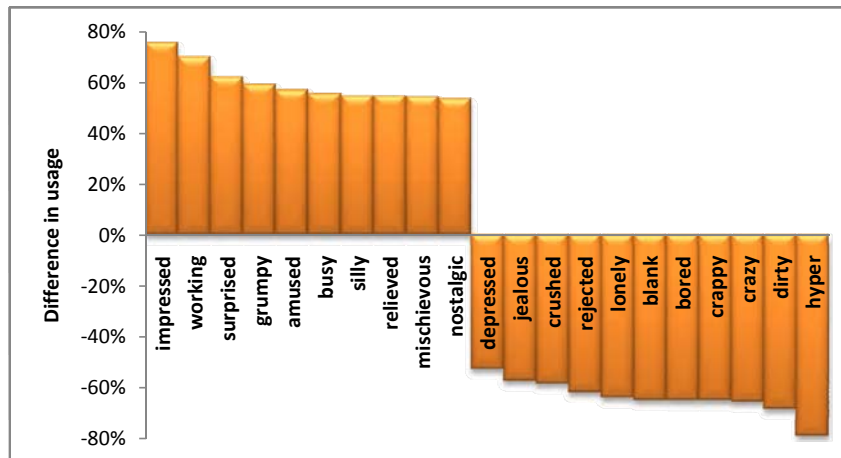


Figure 6.8: Moods above the zero line are in favour of the *old*; otherwise, they are in favour of the *young*.

Bloggers who join many communities favour using more *personal concerns* words, including *work*, *leisure*, *religion* and *death*, than those who join few communities. Those joining many communities favour *perceptual* words (for example, *hear* and *see*), whereas the less social bloggers prefer *cognitive* words (such as *think* and *because*).

### 6.5.3 Difference of mood across age and social connectivity

As shown in Figure 6.5, though worse than topic and linguistic style in the role of features, moods perform reasonably well in prediction for age (approximately 80 per cent) and social connectivity (approximately 70 per cent) of bloggers. As can be seen from Figure 6.8, the young use more low valence moods than the old, such as *rejected* (valence: 1.5), *depressed* (1.83), *lonely* (2.17), *jealous* (2.51) and *bored* (2.95). On the other hand, the old tag more high valence moods to their blog posts, such as *impressed* (7.33), *silly* (7.41) and *surprised* (7.47).

The differences in manifestation of mood usage in the model by social and solo bloggers are shown in Figure 6.9. All moods (except *shocked*) used by social bloggers are positive (based on LIWC) and high valence (based on ANEW). In contrast, solo bloggers use language dominated by negative and low valence emotions. This is in accordance with the finding in Rusting and Larsen [1995] that extraverts experience

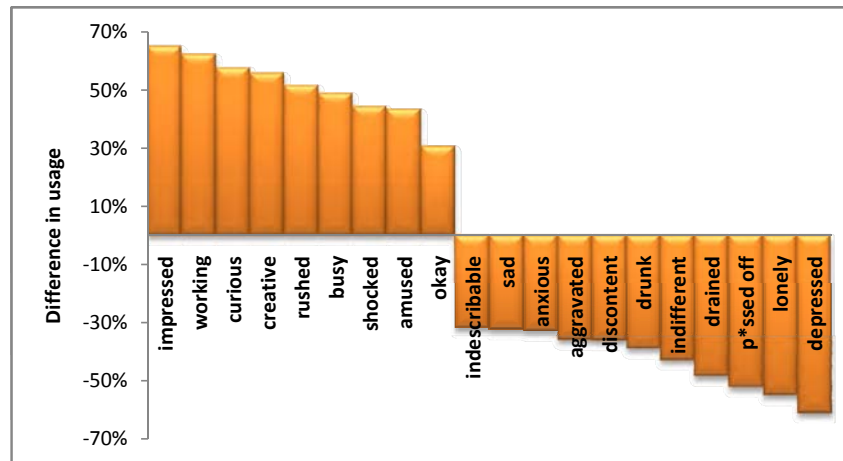
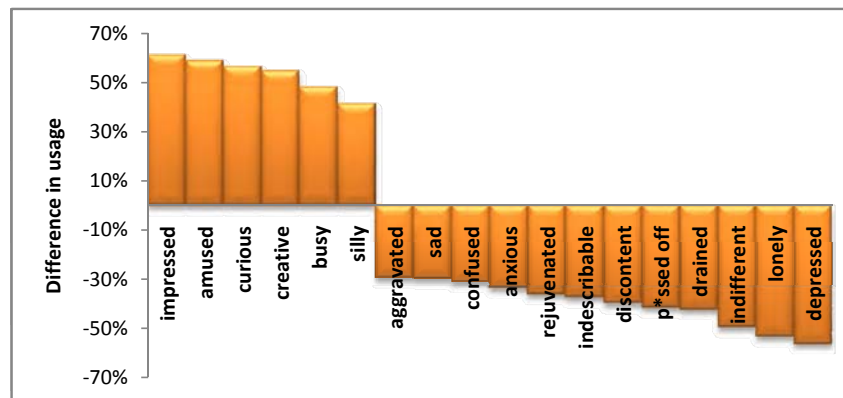
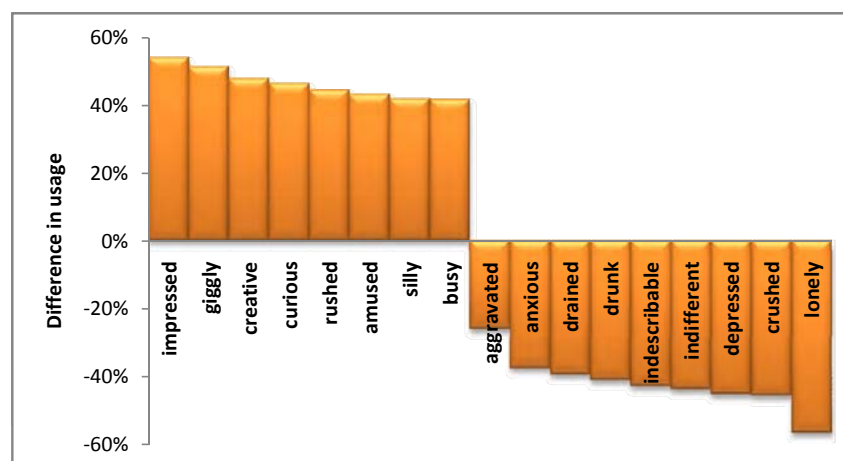
(a) *Many followers versus few.*(b) *Many friends versus few.*(c) *Many communities versus few.*

Figure 6.9: Moods above the zero line are in favour of the *social*; otherwise, they are in favour of the *solo* ( $ps < 0.001$ ).

more pleasant and less unpleasant moods than introverts.

*Influentials* or *non-influentials* here are defined by the number of followers they have in Livejournal. Following this assumption, influentials use more high valence moods (all are larger than 6.0) while non-influentials use more low valence moods (all are smaller than 5.0). Furthermore, influentials prefer using *curious* or *shocked*, which potentially attract people to read the accompanying content.

## 6.6 Conclusion and Future Work

We have investigated the potential for using textual social media, in our case, blog posts, to infer properties of their authors, including age and social connectivity. We have found significant differences among these cohorts when characterised by latent topics of discussion and psycholinguistic features. Latent topics are found to have greater predictive power than linguistic features when classifying bloggers as either old or young. Meanwhile, the degree of blogger social connectivity is effectively predicted by both content-based feature sets, with linguistic features marginally outperforming topic.

On the other hand, manually tagged moods, which do not require a supervised feature selection stage, perform well as features in the prediction of the age and social connectivity of bloggers. This reveals potential for a sentiment approach in egocentric analysis in social media.

The results presented here have application for personalised information retrieval, which relies on knowledge of user attributes such as age and personality traits to re-rank results. On-line advertising can also make use of estimated user profile attributes to further target advertisements based on age and user personality.

# Chapter 7

## Networking Aspect of Social Media

Networking and interaction play a central part in any social media system and the ‘community’ is an example of this aspect of social media. To recommend users suitable communities to join demands the clustering of communities into groups of homogeneous communities, that is, hyper-communities. To this end, the *sentiment* information conveyed in the content can provide additional insight into the process of community formation. However, this sentiment-based perspective has not been explored in community structure studies to date.

In this chapter sentiment information is incorporated into studies of community organisation. It is compared with two other aspects of the content: topics and linguistic styles. We use non-parametric clustering to automatically discover the unknown number of hidden hyper-communities and present the results obtained from a large dataset. The results show that a sentiment-based approach can yield useful insights into community formation.

In addition, this chapter offers a comprehensive comparison of using different features for the problem of community prediction in social networks. A range of conventional features is employed for prediction, including topics, linguistic styles, sentiment information and profiles. Good prediction is achieved when using topic, linguistic and sentiment information as features, leaving the others a modest role. The result is the introduction of a new approach, incorporating sentiment information, for analysis of the networking properties of social media.



## 7.1 Networking in Social Media: Communities

Unlike conventional broadcasting media, users in social media can exchange content and interact with others. A means enabling this practice is *communities*, through which people of common interest can join to discuss their preferred topics. To enable individuals to locate suitable communities to join requires the categorisation of those communities. While one level of such categorisation is provided by the community name itself, a higher level of structure, hidden from the user, can exist across such communities. Such hyper-groups constitute communities of communities and are useful to uncover automatically because they enable the user to find ‘communities like the ones I am interested in’.

Learning hyper-groups of communities in social media has been attempted by learning the link structures among communities [Backstrom et al., 2006, Girvan and Newman, 2002, Kumar et al., 2006]. The obvious drawback of such an approach lies in the dynamic nature of media, which allows users to freely join and leave communities, leading to link structures that are not stable over time. Further, in many instances, these links are not explicit. Other methods have used metadata such as tag and membership information to discover hyper-groups in Flickr [Negoescu et al., 2009].

An alternative to discovering hyper-groups is to use the content itself. Here, we explore the role of sentiment conveyed in the content. We will collate the sentiments for the blogs belonging to one community and perform clustering across these ‘*bag-of-sentiments*’ to uncover meta-groups. These meta-groups produce similar mood or sentiment and thus serve as a barometer of meta-group mood. For comparison, two other conventional aspects of content are used. First, we consider the topics discussed by users in each community by performing a topic-based analysis of the blogs (using LDA) and then cluster these ‘*bag-of-topics*’ to discover meta-groups. The expectation is that these meta-groups will share similar topics. Second, we use psycholinguistic features to categorise communities alike in linguistic style into a cluster.

In addition, this chapter uses different features for community membership prediction in social networks. The results from this investigation provide insight into the

question ‘Why do people join virtual communities?’ The reasons given are not expressed in terms of friendship, exchange of information, social support or recreation, as they are in Ridings and Gefen [2004], but rather with regard to the topics, linguistic styles and moods that people express in their on-line writings. These are used as features to predict community membership for bloggers.

Our contribution lies in producing the first work to identify latent hyper-groups in social communities through an understanding of sentiment and content. In particular, using sentiment to uncover groupings in this way is novel. The significance lies in the use of the characteristics of meta-communities, such as in targeted support for communities or members whose moods are regularly low in valence, and targeted marketing to focus on groups that have similar sentiment and topics.


The rest of this chapter is organised as follows. Section 7.2 presents a scheme for hyper-community detection and community prediction in the blogosphere. Sections 7.3 and 7.4 present the results of the scheme applied to a blog dataset and are followed by some concluding remarks.

## 7.2 Hyper-community Detection Framework

In this section, we present content-based, sentiment-based and psycholinguistic-based approaches to cluster blog communities into groups automatically—a problem we term *hyper-community detection*. The aim is to group communities that are related in either content, sentiment or both. In the content-based method, we extract topics from blog content using topic-modelling tools and measure content similarity using topic-based representations for clustering. For the sentiment-based case, we investigate the usefulness of including sentiment information in the clustering task. To achieve this, a mood or emotion-bearing lexicon (ANEW) is extracted from blog content and used as features. In the psycholinguistic-based approach, we use psycholinguistic features provided from psychological studies to cluster communities. In all cases we use data extracted from Livejournal<sup>1</sup> for our investigations. However we note that the proposed method is directly applicable for similar datasets. An example of Livejournal blog posts in communities is shown in Figure 7.2.


---

<sup>1</sup><http://www.livejournal.com>.




**Cooking**  
Created on 2001-01-10 17:14:13 (#40870), last updated 2010-06-07 334,304 comments received  
**Members (9853)**  
32,732 Journal Entries, 218 Tags, 3,121 Memories, 0 Virtual Gifts, 2 Userpics

A community for anyone who likes to cook (in the kitchen, that is!) Feel free to post all things culinary here, be it your favorite recipes that you want to share with others, cooking advice, questions, suggestions... anything that's on your mind.




**Bentolunch**  
Created on 2005-08-25 09:48:27 (#8115356), last updated 2010-06-07 112,642 comments received  
**Members (7436)**  
16,444 Journal Entries, 375 Tags, 7 Memories, 0 Virtual Gifts, 6 Userpics

This community is for people who pack bento boxes or similar lunches and want to share their lunch ideas and get lunch ideas from others. You don't need to be packing your lunch in a bento box to participate or share ideas.





**Bookish**  
" Quoth the raven, 'Nevermore.' "  
Created on 2001-06-26 18:06:32 (#207338), last updated 2010-06-07 79,418 comments received  
**Members (7411)**  
9,503 Journal Entries, 23 Tags, 0 Memories, 0 Virtual Gifts, 1 Userpic

Are you bookish? Is it one of the great thrills in your life to have a good conversation about literature? Then you are a dork... But who isn't? This community is for a specific type of dork, those of us who have a passion for great books, and no other place to talk about them without our non-bookish friends being bored to tears.




**50 Book Challenge**  
Created on 2003-01-05 07:31:53 (#840538), last updated 2010-06-07 81,240 comments received  
**Members (7126)**  
44,507 Journal Entries, 191 Tags, 0 Memories, 0 Virtual Gifts, 1 Userpic

This community is for people who have joined me (  jadis ) in my challenge to read 50 books each year. It is a place to discuss progress, get ideas for further reading, chat about books you've completed, etc.



**POKÉMON**  
Gotta catch 'em all!  
Created on 2001-02-03 04:56:18 (#51808), last updated 2010-06-07 251,718 comments received  
**Members (5844)**  
13,758 Journal Entries, 341 Tags, 14 Memories, 4 Virtual Gifts, 2 Userpics


Discussion of all things Pokémon





**pkmncollectors**  
pkmncollectors's Journal  
Created on 2007-04-12 23:55:18 (#12712834), last updated 2010-06-07 619,845 comments received  
**Members (2055)**  
22,340 Journal Entries, 544 Tags, 0 Memories, 0 Virtual Gifts, 2 Userpics

Buy, sell, trade, collect and show-off your Pokémon merchandise. :D

Figure 7.1: Illustration of profiles for six Livejournal communities: *Cooking* and *Bentolunch*, consider similar topics; as do *Bookish* and *50bookchallenge*; and *Pokemon* and *Pkmncollectors*.



Teeluh (  teeluh ) wrote in **obama\_2008** on 2008-08-05 00:56:00

Current mood:  good

**User: Teeluh**

**Community: Obama-2008**

**Mood: Good**

Change;

My **mother**, age 47, has finally registered to vote. In the past, she's always been so apathetic about anything political; claiming she felt she didn't understand politics enough to feel the right to vote, (when in reality she knows a **hell** of a lot more than the average voter). Now, she's so enthralled with this election that she has finally taken that step of registering to vote! I'd be **proud** of her even if it was McCain that captured her faith but because she happens to believe so strongly in Obama, I'm absolutely ecstatic! I just think it's amazing that Barack Obama's politics has made her change her **mind** about voting~  
:D  
But on another note, I **lost** my precious Obama pin, (that took me over a **month** to get)!

Figure 7.2: An example of blog posts posted in *obama\_2008* community. It is tagged with the 'current mood' *good*. ANEW words used in the blog content are highlighted, including *mother*, *hell*, *proud*, *mind*, *lost* and *month*.

Category	Community
Advice-Support	add_a_writer, addme25_and_up, baristas, boys_and_girls, i_am_thankful, iworkatborders, thenicestthings, todayirealized, walmart_employe, weddingplans
Creative-Expression	__quotexwhore, 20sknitters, adayinmylife, aesthetes, amateur_artists, behind_the_lens, charloft, color_theory, naturesbeauty, sew_hip
Entertainment-Music	beatlepics, bjorkish, broadway, just_good_music, news_jpop, patd, relaxmusic, rilokiley, theater_icons, thecure
Fandom	chuunin, house_cameron, miracle_____, ncisfcfind, patdslashseek, rpattz_kstew, sgagenrefinders, sgastoryfinders, sheldon_penny, time_and_chips
Fashion-Style	beauty101, corsetmakers, curlyhair, dyed_hair, egl, egl_comm_sales, madradstalkers, ourbedrooms, ru_glamour, vintagehair
Food-Travel	bentolunch, davis_square, eurotravel, filmmsg, newyorkers, ofmornings, picturing_food, seattle, trashy_eats, world_tourist
Gaming-Technology	computer_help, computerhelp, gamers, htmlhelp, ipod, macintosh, thesims2, webdesign, worldofwarcraft, wow_ladies
Parenting-Pets	altparent, baby_names, breastfeeding, cat_lovers, clucky, dog_lovers, dogsintraining, naturalbirth, note_to_cat, parenting101
Politics-Culture	birds, blackfolk, classics, ftm, nonfluffypagans, ontd_political, poor_skills, prolife, queer_rage, transnews
Television	_we_are_lost, battlestar_blog, calmallamadown, doctorwho, glee_tv, gleeclub, gundam00, lword, theoffice_us, topmodel

Table 7.1: Communities from 10 Livejournal directories used in experiments.

We crawled the communities listed in the Livejournal directory.<sup>2</sup> These communities are categorised by Livejournal into 13 groups: Advice-Support, Creative-Expression, Entertainment-Music, Everything Else, Fandom, Fashion-Style, Food-Travel, Gaming-Technology, Parenting-Pets, Politics-Culture, Sports-Fitness, Television and Thread-based RP. From the 579 communities obtained (consisting of 1,090,408 posts and 10,081,215 comments by 182,197 members), we extracted a subset consisting of the top 100 communities having the most members across 10 categories, resulting in a dataset of 211,740 posts by 59,496 users. Table 7.1 lists the communities used in the following experiments.

### 7.2.1 Community representation

There has been extensive work on characterising collections of text documents by topic, including blog sub-communities [Adams et al., 2010] and tagged media [Negoescu et al., 2009]. However, the role of sentiment and mood has not been studied. Arguably, mood is often an integral feature of a text, particularly for social media forums; two communities might discuss precisely the same topics, yet within an entirely different atmosphere. For example, where one forum might host conversations about politics in a cerebral, serious-minded and friendly fashion; another will discuss the same issues adversarially, with zest and tolerance of profanity. Such mood-related distinctions are important for many kinds of analysis and application, not the least of which is community recommendation.

On-line communities come in many shapes and sizes and are affected by many factors, including the demographics of their members, reason for existence and facilities afforded by the hosting application. The Livejournal blog site, referred to in the previous sections, includes a community feature. Each community is defined by the scope of topics it aims to host and comprises, among other things, members and posts and comments made in response to posts. Figure 7.1 shows example profiles for six Livejournal communities.

---

<sup>2</sup><http://www.livejournal.com/browse/>, retrieved July 2011.

### 7.2.1.1 Topic-based representation

To represent what community members talk about, we apply LDA—a Bayesian probabilistic topic model (see Section 2.5.5)—to the corpus of blog posts. All posts for each community are aggregated to form the corpus input to LDA, wherein each post is considered as one document. LDA learns the probabilities  $p(\text{vocabulary} \mid \text{topic})$ , that are used to describe a topic and assigns a topic to each word in every document. Each post can then be represented as a mixture of topics using the probability  $p(\text{topic} \mid \text{document})$ . A topic-based representation for each community is then constructed based on topic mixtures for those blog posts belonging to that community. We expect similar communities to discuss a similar mixture of topics, and hence to have a similar mixture of  $p(\text{topic} \mid \text{document})$  aggregated from their posts.

Formally, let  $J$  be the number of communities, denote by  $\mathbf{x}_j = \{x_{1j}, x_{2j}, \dots, x_{n_j j}\}$  the set of blog posts in the  $j^{\text{th}}$  community where  $n_j$  is the total number of posts by this community. Thus, the corpus to be modelled consists of  $N = \sum_{j=1}^J n_j$  documents aggregated from all communities  $D = \cup_{j=1}^J \mathbf{x}_j$ . Finally, if  $\theta_{ij}$  denotes the topic mixture for blog post  $x_{ij}$ , the  $j^{\text{th}}$  community can be represented by  $\theta_j = (1/n_j) \sum_{i=1}^{n_j} \theta_{ij}$ .  $\theta_j$  is a  $K$ -dimensional vector, where  $K$  is the number of topics used by LDA and the  $k^{\text{th}}$  element represents the mixture proportion of topic  $k$  for community  $j$ . Figure 7.3 shows the topic mixtures for the 100 communities. These mixtures are used to perform topic-based community clustering.

The topic distributions are well separated among some groups of communities. As can be seen in Figure 7.4,  $\{\text{dog\_lovers}, \text{dogs\_in\_training}\}$  could be inferred as a group of communities mainly talking about the character **Dog** (topic 9); similarly,  $\{\text{cat\_lovers}, \text{note\_to\_cat}\}$  about **cat** (topic 20);  $\{\text{macintosh}, \text{computer\_help}, \text{computerhelp}, \text{ipod}\}$  about **computer/ipod** (topic 23); and  $\{\text{webdesign}, \text{htmlhelp}\}$  about **web design** (topic 29).

### 7.2.1.2 Sentiment-based representation

Instead of grouping communities based on their topics as in the previous section, we group communities based on sentiment. Sentiment extracted from blog posts

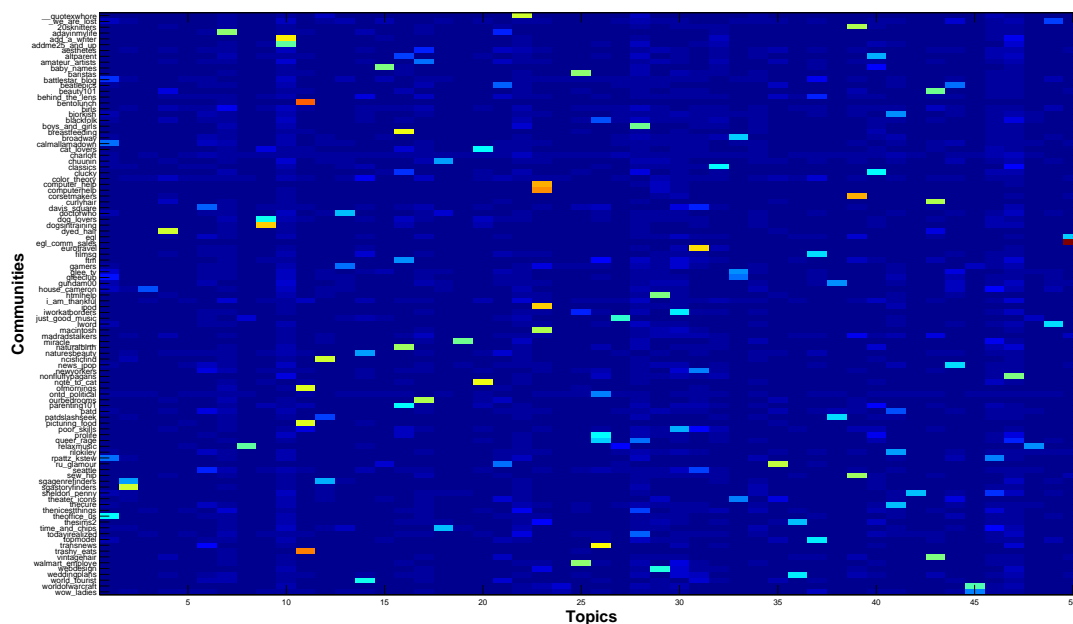
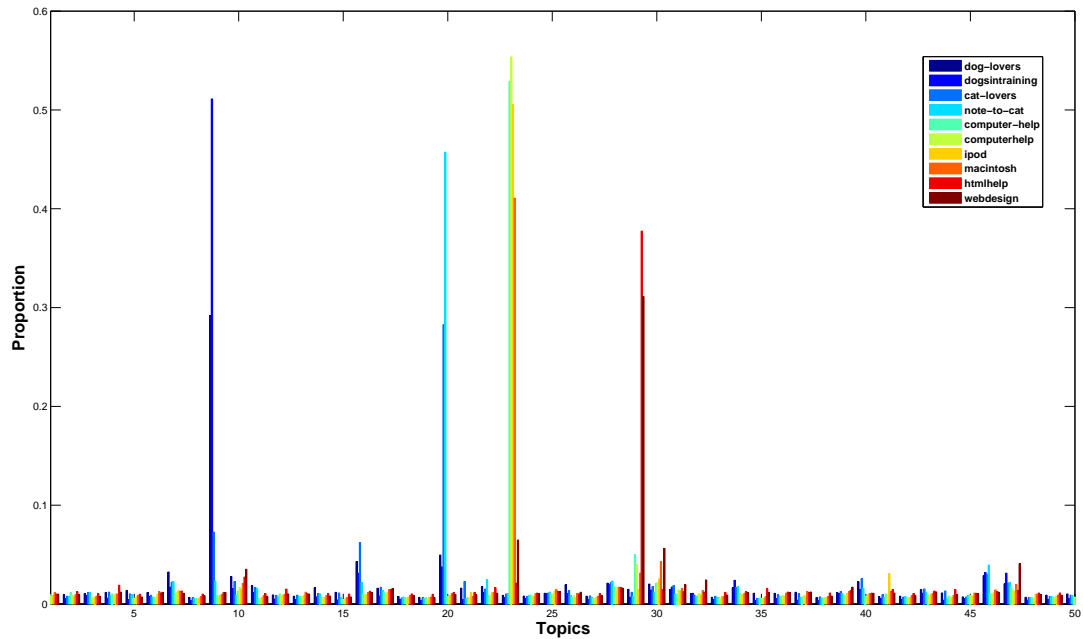


Figure 7.3: Topic proportions of 100 communities.

is analysed without considering topic. Two methods to extract sentiment from a community are used in this study. If a blog post was tagged with a mood when it was composed, we can use this information to compute an overall sentiment for the community based on moods aggregated from its set of blog posts. Otherwise, when mood is not available, we propose to use sentiment ratings using ANEW, a sentiment-bearing lexicon described in Section 2.5.2.

**Using mood** Recall that Livejournal offers 132 moods for users to tag their posts. We assume that there exists a difference in tagged moods among communities, supporting the intuition that such communities can be grouped by mood.

Let  $\mathcal{M} = \{sad, happy, \dots\}$  again be the predefined set of moods where  $|\mathcal{M}| = 132$  is the total number of moods provided by Livejournal. Using the notation in the previous section, each blog post  $x_{ij}$  in the  $j^{th}$  community is further tagged with a mood  $m_{ij} \in \mathcal{M}$ . For each community, a 132-dimension mood usage vector  $\mathbf{m}_j$  is constructed whose  $k^{th}$  element is the number of times the  $k^{th}$  mood in  $\mathcal{M}$  was tagged within this community. Figure 7.5a shows the mood proportions for the 100 communities. These proportions are used to perform mood-based community clustering.



Topic	Top Topic Terms
9	<b>dog</b> dogs puppy training animal gets started tried vet outside pet months house problem walk away pretty run big loves
16	<b>baby</b> months sleep weeks month birth started start milk question thanks tried hospital doctor eat daughter breast couple pain weight
20	<b>dear cat</b> cats food stop mommy kitty thank bed vet water sleep mom big eat kitten litter glad room clean
23	<b>computer ipod</b> tried problem windows itunes using apple mac thanks drive files screen file running internet music fix open download
29	<b>link table page thanks site</b> code links click text journal change post website layout background thank picture entries box entry

Figure 7.4: Above: topic proportions of 10 communities. Below: example topics and most likely words sized by  $p(\text{word} | \text{topic})$ .



Figure 7.5b shows a plot of the mood usage by eight different communities in Livejournal. It can be seen that the mood usage in one group of communities (*computer\_help*, *computerhelp*, *htmlhelp*, *ipod*, *webdesign*) is well separated from another group (*ncisficfind*, *sgagenrefinders*, *sgastoryfinders*). The first group favours using moods having low valence (such as *p\*ssed off*, *worried* and *confused*) while the second prefers high valence moods (for instance, *hopeful*), empirically suggesting that it is sensible to study grouping behaviour based on mood.

**Using an emotion bearing lexicon** When mood ground truth is not available (for example, the social media does not implement a mood feature or it is present but not used), sentiment-based hyper-community detection can be performed using ANEW vectors (as described in Chapter 3).

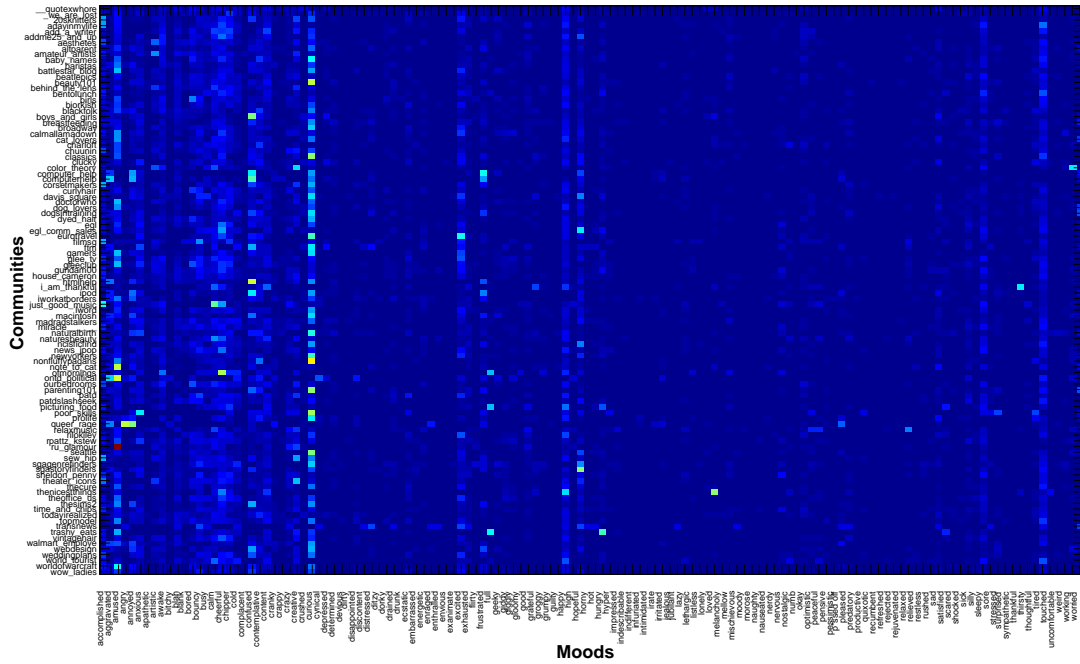
Using the notation in Section 7.2.1.1, each  $j^{th}$  community is now represented with a 1,034-dimension ANEW feature vector  $\mathbf{a}_j$ , whose  $k^{th}$  element is the number of times the  $k^{th}$  ANEW word is used in the content of the blog posts made by users belonging to the community.

### 7.2.1.3 Psycholinguistic-based representation

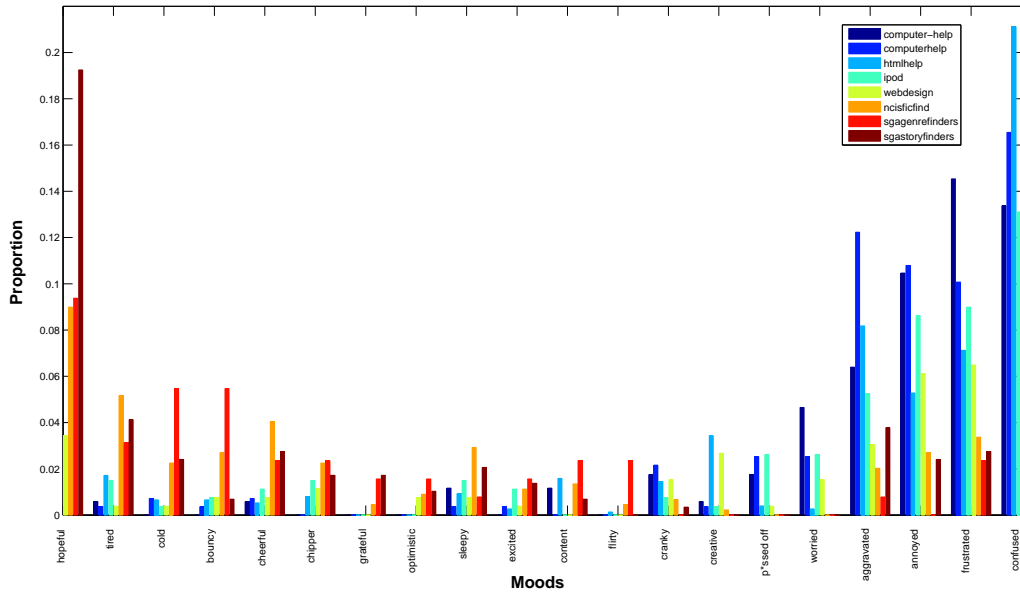
As a final point of comparison that bridges pure topical and sentiment-based representation of communities, we use psycholinguistic features as classified by the LIWC taxonomy (see Section 2.5.4). Recall that LIWC assigns terms to 68 categories, ranging across topics, emotions and other processes such as cognition and tense. In all, 68 LIWC features are used to build a vector to provide a psycholinguistic representation of each community.

## 7.2.2 Community clustering

To group communities into hyper-communities we again use AP—a non-parametric clustering algorithm (see Section 2.4.2). Recall that AP can automatically discover the number of clusters as well as the cluster exemplars. This is crucial in our setting as the number of hyper-communities can be extremely difficult to know in advance.



(a) The mood usage proportions of 100 communities used in hyper-community detection.



(b) An illustration of mood usage proportions in two groups of communities: {*computer\_help*, *computerhelp*, *htmlhelp*, *ipod*, *webdesign*} and {*ncisficfind*, *sgagenrefinders*, *sgastoryfinders*}.

Figure 7.5: Communities and mood usage.

The algorithm requires the pairwise similarities between data points. In our case, it is the similarity computed between  $\theta_j$  and  $\theta_l$  for the  $(j, l)$ -pair of communities.

For topical hyper-communities, we note that each  $\theta_j$  is a proper probability mass function over topics, summing up to 1. For sentiment and psycholinguistic-based hyper-communities the feature vectors are normalised. Thus, any suitable probability distance functions can be employed to compute the similarities. In this work, we use the negative KL and the BC, described in Section 2.4.2.1.

## 7.3 Hyper-community Detection Results

Overall clustering performance for the different community representations—topic, mood, mood-proxy and psycholinguistic—is shown in Table 7.6. We report CP and NMI (defined in Section 2.4.2.2) using the Livejournal community classification (shown in Table 7.1), which is a topical classification, as the groundtruth. Therefore, it is expected that these metrics will be highest for the topic-based community representation, 70 per cent purity and 62 per cent NMI, versus the result using mood information, 46 per cent purity and 43 per cent NMI. We are chiefly concerned with new knowledge discovered using mood-related representations, which will be analysed in more detail below for each type of representation.

### 7.3.1 Topic-based hyper-groups

Using LDA with 50 topics yielded 20 hyper-communities, which are listed in Table 7.2. Figure 7.6 shows community assignments to clusters together with Livejournal category. Clustering appears to have gathered topically similar communities together in a number of cases (for example,  $\{ofmornings, bentolunch, picturing\_food, trashy\_eats\}$ ), but also elucidated finer distinctions (such as in the cases of  $\{cat\_lovers, note\_to\_cat\}$  and  $\{dog\_lovers, dogsintraining\}$ ). Hyper-community VIII has the lowest purity. On further inspection, a number of its communities have a significant romance or relationships component. For example, in addition to those communities with obvious topics, three are about particular fictional relationships: *house\_cameron*, *sheldon\_penny* and *time\_and\_chips*.

No.	Members	No.	Members
I	20sknitters, corsetmakers, sew_hip	XI	egl, egl_comm_sales <b>(Fashion-Style2)</b>
II	addme25_and_up, add_a_writer <b>(Advice-Support1)</b>	XII	glee_tv, broadway, gleeclub, theater_icons
III	beauty101, curlyhair, dyed_hair, vintagehair <b>(Fashion-Style1)</b>	XIII	macintosh, computer_help, computerhelp, ipod <b>(Gaming-Technology1)</b>
IV	bjorkish, patd, rilokiley, thecure <b>(Entertainment-Music)</b>	XIV	newyorkers, davis_square, eurotravel, poor_skills, seattle
V	blackfolk, classics, nonfluffypagans, ontd_political, prolife, queer_rage, transnews <b>(Politics-Culture)</b>	XV	ofmornings, bentolunch, picturing_food, trashy_eats <b>(Food-Travel)</b>
VI	calmallamadown, battlestar_blog, news_jpop, rpattz_kstew, theoffice_us	XVI	parenting101, altparent, breastfeeding, clucky, ftm, naturalbirth
VII	cat_lovers, note_to_cat <b>(Parenting-Pets1)</b>	XVII	sgagenrefinders, ncisficfind, sgastoryfinders <b>(Fandom)</b>
VIII	charloft, _we_are_lost, baby_names, birls, chuunin, doctorwho, gamers, gundam00, house_cameron, i_am_thankful, just_good_music, lword, miracle_-, patdslashseek, relaxmusic, sheldon_penny, thesims2, time_and_chips, weddingplans, worldofwarcraft, wow_ladies	XVIII	todayirealized, __quotexwhore, boys_and_girls, thenicestthings
IX	color_theory, adayinmylife, aesthetes, amateur_artists, beatlepics, behind_the_lens, filmsg, madradstalkers, naturesbeauty, ourbedrooms, ru_glamour, topmodel, world_tourist	XIX	walmart_employe, baristas, iworkatborders <b>(Advice-Support2)</b>
X	dog_lovers, dogsintraining <b>(Parenting-Pets2)</b>	XX	webdesign, htmlhelp <b>(Gaming-Technology2)</b>

Table 7.2: 20 topic-based hyper-communities (exemplar listed first for each).

Members	Categories	Mood Cloud
altparent, boys_and_girls, breastfeeding, cat_lovers, clucky, dog_lovers, dogsintraining, macintosh, naturalbirth, parenting101, todayirealized	Advice-Support, Gaming-Technology, Parenting-Pets	<b>curious</b> confused amused worried contemplative happy cheerful anxious annoyed frustrated calm tired hopeful aggravated sad excited shyover accomplished awake aware
baristas, blackfolk, iworkatborders, nonfluffpagans, note_to_cat, ontd_political, prolife, queer_rage, thesims2, transnews, walmart_employe, worldofwarcraft	Advice-Support, Gaming-Technology, Parenting-Pets, Politics-Culture	<b>amused</b> curious annoyed aggravated confused angry contemplative tired frustrated happy awake passedof cheerful excited sad sorry hopeful ban worried accomplished
beatlepics, __quotexwhore, addme25_and_up, birls, charloft, egl_comm_sales, gundam00, house_cameron, miracle_____, news_jpop, patdslashseek, rpattz_kstew, sheldon_penny, thenicestthings, time_and_chips	Advice-Support, Creative-Expression, Entertainment-Music, Fandom, Fashion-Style, Politics-Culture, Television	<b>cheerful</b> amused bouncy accomplished happy calm tired hopeful sleepy chipper bored curious content awake anxious creative blah excited busy loved
behind_the_lens, add_a_writer, aesthetes, amateur_artists, color_theory, filmsg, just_good_music, naturesbeauty, ofmornings, ourbedrooms, relaxmusic	Advice-Support, Creative-Expression, Entertainment-Music, Fashion-Style, Food-Travel	<b>calm</b> cheerful accomplished creative artistic awake chipper amused happy curious tired bored bouncy content sleepy cold busy contemplative working anxious
bentolunch, adayinmylife, i_am_thankful, picturing_food, trashy_eats, world_tourist	Advice-Support, Creative-Expression, Food-Travel	<b>happy</b> hungry full accomplished cheerful tired calm content amused sleepy chipper creative thankful satisfied busy cold awake bouncy curious exhausted
broadway, _we_are_lost, baby_names, beauty101, bjorkish, classics, curlyhair, davis_square, doctorwho, egl, eurotravel, ftm, gamers, lword, madradstalkers, newyorkers, poor_skills, rilokiley, seattle, thecure, theoffice_us, topmodel, weddingplans, wow_ladies	Advice-Support, Entertainment-Music, Fashion-Style, Food-Travel, Gaming-Technology, Parenting-Pets, Politics-Culture, Television	<b>curious</b> amused excited cheerful chipper confused anxious tired hopeful happy bouncy bored calm contemplative content accomplished awake annoyed aware
chuunin, 20sknitters, battlestar_blog, calmallamadown, corsetmakers, dyed_hair, glee_tv, gleeclub, patd, ru_glamour, sew_hip, theater_icons, vintagehair	Creative-Expression, Entertainment-Music, Fandom, Fashion-Style, Television	<b>amused</b> curious accomplished creative cheerful excited bouncy happy chipper confused bored artistic calm tired anxious content awake hopeful sleepy crazy
htmlhelp, computer_help, computerhelp, ipod, webdesign	Gaming-Technology	<b>confused</b> curious frustrated annoyed aggravated hopeful anxious worried sad annoyed aware
ncisficfind, sgagenrefinders, sgastoryfinders	Fandom	<b>hopeful</b> curious tired cold cheerful bouncy frustrated anxious bored aggravated confused chipper calm sleepy ban worried accomplished aware

Table 7.3: Nine mood-based hyper-communities.

Members	Categories	ANEW Cloud
addme25_and_up, add_a_writer, htmlhelp, nonfluffypagans	Advice-Support, Gaming-Technology, Politics-Culture	<b>journal</b> people love time good free life music name pretty thought hope sake part head name see back party ...
altparent, baby_names, breastfeeding, clucky, ftm, naturalbirth, parenting101, prolife	Parenting-Pets, Politics-Culture	<b>baby name</b> time people good thought love month pretty abortion child kids home surgery girl doctor milk hope life couple
beatlepics, amateur_artists, birls, calmallamadown, chuunin, classics, gundam00, news_jpop, ourbedrooms, patd, theater_icons, thecure, theoffice_us	Creative-Expression, Entertainment-Music, Fandom, Fashion-Style, Politics-Culture, Television	<b>love time</b> thought good hope people news pretty friend office free happy name song art part kind panic journal cut
behind_the_lens, aesthetes, color_theory, filmmsg, naturesbeauty, ru_glamour, world_tourist	Creative-Expression, Fashion-Style, Food-Travel	<b>free time</b> love hope thought good spring art part glamour people color white black pretty fall sunset city red happy
bentolunch, ofmornings, picturing_food, trashy_eats	Food-Travel	<b>chocolate green egg good</b> salad food dinner time butter love eat red white pancakes milk pretty black cake sugar cut
curlyhair, beauty101, dyed_hair, vintagehair	Fashion-Style	<b>good</b> cut time love color pretty red black thought kind people face dark blue hope natural couple size new trend
dog_lovers, cat_lovers, dogsintraining, note_to_cat	Parenting-Pets	<b>dog cat</b> love time good home food thought puppy people house pretty cute happy best pet our name size car
egl, 20sknitters, corsetmakers, egl_comm_sales, madradstalkers, sew_hip, weddingplans	Advice-Support, Creative-Expression, Fashion-Style	<b>good dress time</b> love pretty black people fabric thought wedding white kind idea friend free hope cute cut couple name
lword, _we_are_lost, battlestar_blog, bjorkish, blackfolk, broadway, glee_tv, gleeclub, rilokiley, topmodel	Entertainment-Music, Politics-Culture, Television	<b>love people</b> thought good time lost song music hope part watch free pretty girl name friend black kind cut happy
macintosh, computer_help, computerhelp, ipod, webdesign	Gaming-Technology	<b>computer</b> time good music free hard people thought size party name music trend name car name name name name name
miracle_-, charloft, house_cameron, just_good_music, ncisficfind, patdslashseek, relaxmusic, rpattz_kstew, sgastoryfinders, sheldon_penny, thesims2	Creative-Expression, Entertainment-Music, Fandom, Gaming-Technology	<b>love time part</b> house good hope thought people pretty music life song rock kind happy family home world name size
newyorkers, adayinmylife, baristas, davis_square, eurotravel, iworkatborders, ontd_political, poor_skills, seattle, transnews, walmart_employe	Advice-Support, Creative-Expression, Food-Travel, Politics-Culture	<b>time good people</b> free love thought pretty city money home friend couple book part hope kind car idea travel nice
time_and_chips, doctorwho	Fandom, Television	<b>doctor</b> time thought ... .. .
todayirealized, __quotexwhore, boys_and_girls, i_am_thankful, queer_rage, sgagenrefinders, thenicestthings	Advice-Support, Creative-Expression, Fandom, Politics-Culture	<b>love people</b> good thankful time friend life thought happy person pretty kind girl hope taste size best best name size
worldofwarcraft, gamers, wow_ladies	Gaming-Technology	<b>game time</b> good people love priest pretty fun thought quick world happy name hit free task trend couple pet size

Table 7.4: 15 ANEW-based hyper-communities.













		
Food-Travel	Creative-Expression, Fashion-Style	Fashion-Style
<i>trashy_eats, bentolunch, ofmornings, picturing_food</i>	<i>color_theory, naturesbeauty, ru_glamour</i>	<i>vintagehair, beauty101, curlyhair, dyed_hair</i>
		
Creative-Expression, Entertainment-Music, Politics-Culture, Television	Parenting-Pets	Advice-Support, Parenting-Pets
<i>just_good_music, battlestar_blog, charloft, ontd_political, relaxmusic</i>	<i>altparent, baby_names, breastfeeding, clucky, naturalbirth, note_to_cat, parenting101</i>	<i>dog_lovers, boys_and_girls, cat_lovers, dogsintraining, thenicestthings</i>
		
Creative-Expression, Fandom	Fandom, Fashion-Style, Food-Travel, Politics-Culture	Advice-Support, Creative-Expression, Food-Travel, Gaming-Technology, Television
<i>sheldon_penny, adayinmylife, house_cameron, miracle_ , rpattz_kstew, time_and_chips</i>	<i>newyorkers, davis_square, egl_comm_sales, eurotravel, madradstalkers, ourbedrooms, poor_skills, seattle, sgagenrefinders, world_tourist</i>	<i>aesthetes, amateur_artists, behind_the_lens, doctorwho, filmmsg, i_am_thankful, wow_ladies</i>
		
Advice-Support, Entertainment-Music, Fandom, Fashion-Style, Gaming-Technology, Politics-Culture, Television	Advice-Support, Creative-Expression, Fashion-Style, Gaming-Technology, Politics-Culture	Advice-Support, Creative-Expression, Fandom, Gaming-Technology, Politics-Culture, Television
<i>calmallamadown, addme25_and_up, baristas, beatlepics, birls, bjorkish, broadway, chuunin, egl, gamers, glee_tv, gleeclub, gundam00, lword, ncisficfind, news_jpop, rilokiley, patd, patdslashseek, theater_icons, thecure, theoffice_us, topmodel, weddingplans, worldofwarcraft</i>	<i>macintosh, 20sknitters, add_a_writer, classics, computer_help, computerhelp, corsetmakers, ftm, htmlhelp, ipod, iworkatborders, sew_hip, webdesign</i>	<i>blackfolk, __quotexwhore, _we_are_lost, nonfluffypagans, prolife, queer_rage, sgastoryfinders, thesims2, todayirealized, transnews, walmart_employe</i>

Table 7.5: 12 LIWC-based hyper-communities, showing pie charts of favourite LIWC features; Livejournal categories; and grouped communities.

	Topic	Mood	ANEW	LIWC
No. Clusters	20	9	15	12
CP	70%	46%	63%	54%
NMI	62%	43%	59%	51%

Table 7.6: CP and NMI of the clusterings based on different community representations.

It is evident from Table 7.7 that, in the feature space, the intra-category distances are much smaller than inter-category distances. The smallest distance is between hyper-community VII ( $\{cat\_lovers, note\_tocat\}$ ) and hyper-community X ( $\{dog\_lovers, dogsintraining\}$ ), which indicates the high degree of topical commonality of discussion about pets, despite their being different animals. It is also interesting to note how close the two hyper-communities IV (Entertainment-Music) and V (Politics-Culture) are. The farthest distance is found between hyper-community XV (Food-Travel) and hyper-community XIII (Gaming-Technology) or hyper-community XVII (Fandom).

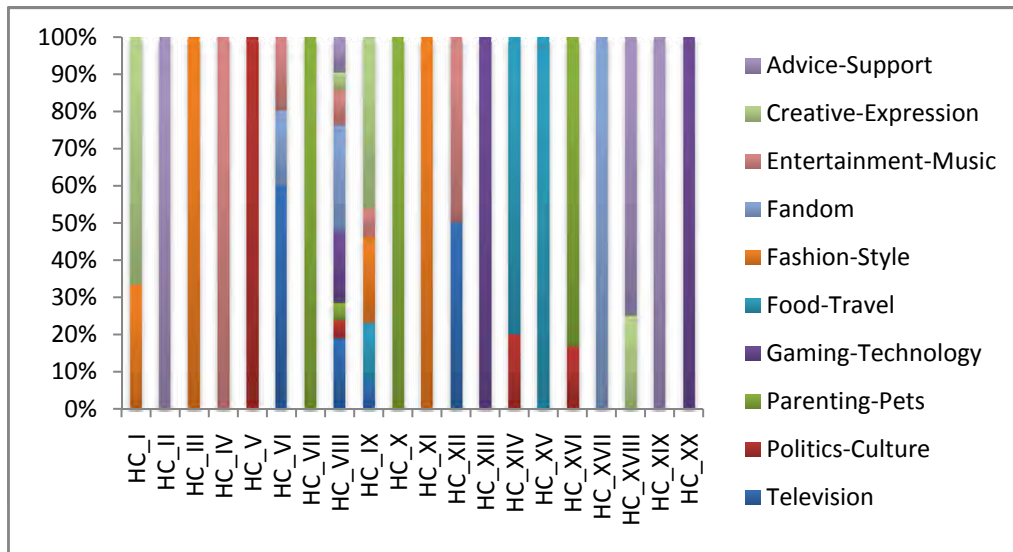


Figure 7.6: Topic-based hyper-communities with Livejournal category; multi-coloured clusters are less pure.



	II	III	IV	V	VII	X	XI	XIII	XV	XVII	XIX	XX
II	0.03	0.25	0.15	0.11	0.20	0.19	0.19	0.24	0.25	0.22	0.23	0.19
III		0.06	0.20	0.18	0.22	0.21	0.18	0.25	0.26	0.26	0.24	0.23
IV			0.01	0.09	0.17	0.17	0.14	0.17	0.21	0.19	0.19	0.14
V				0.07	0.17	0.15	0.15	0.19	0.21	0.17	0.17	0.15
VII					0.03	0.08	0.20	0.24	0.24	0.24	0.22	0.22
X						0.04	0.19	0.23	0.23	0.23	0.21	0.21
XI							0.15	0.21	0.23	0.23	0.21	0.17
XIII								0.01	0.28	0.26	0.23	0.14
XV									0.01	0.28	0.26	0.26
XVII										0.07	0.26	0.23
XIX											0.04	0.22
XX												0.02

Table 7.7: The intra-hypergroup and inter-hypergroup JS-based distances on the topic distribution.

### 7.3.2 LIWC-based hyper-groups

Clustering based on psycholinguistic features yielded 12 hyper-communities, shown in Table 7.5. Three hyper-communities contain communities with the same LiveJournal category and appear to have been associated topically. The top three LIWC categories for these hyper-communities are illuminating: for Fashion-Style, *feel*, *body* and *percept* (that is, perceptual processes); for Food-Travel, *ingest*, *bio* (that is, biological processes) and *percept* and for Parenting-Pets, *family*, *health* and *humans* (for example, adult, baby and boy).

Other hyper-communities appear to exhibit a characteristic mixture of topic and style of discussion, which is in part captured by the linguistic processes of LIWC. For example, one hyper-community (which includes *sheldon\_penny*) aggregates all of the communities in the dataset about fictional relationships (plus one community about documenting a day in one’s life). These communities are a kind of meta-genre not easily captured by topical features alone. Linguistic features, such as post length and extensive use of the third personal singular (that is, *shehe*), appear to help associate these communities.

Lastly, there is at least one hyper-community for which mood appears to be the discriminating feature. This hyper-community, which includes *blackfolk*, has above

average use of the LIWC categories *sad*, *swear*, *death*, *sexual*, *health*, *humans*, *anger*, *relig* and *they* (that is, third person plural). These communities could be described as some combination of angst-ridden, gritty, adversarial (note the above average use of third person plural) or forthright, and contain much negative emotion and introversion. Contrast this hyper-community with that which includes the community *aesthetes*—which discusses similar topics (for example, *health*, *death* and *relig*) but does so with above average *posemo* (that is, positive emotion).

### 7.3.3 Mood-based hyper-groups

Clustering based on explicit mood labels yielded nine hyper-communities, which are recorded in Table 7.3. In contrast to the topic-based clustering, only two hyper-communities have 100 per cent purity with respect to the topical ground truth, one of which is the only group characterised by negative mood: *htmlhelp*, *computer\_help*, *computerhelp*, *ipod*, and *webdesign*.

Mood-based clustering reveals distinctions not apparent in the topic-based representation. For example, the group including *behind\_the\_lens*, while having significant overlap with Group IX (see Table 7.2) in the topic-based clustering, has some illuminating differences: gone are the communities *beatlepics*, *madradstalkers*, *ru\_glamour*, *topmodel*, and *worldtourist*; replacing them are *add\_a\_writer*, *just\_good\_music*, and *ofmornings*.

From an appraisal of the content of these communities we find the distinctions to be nuanced. The topic-based hyper-community is loosely united by pictures and people, whereas the mood-based hyper-community is united by the desire to create and its outcomes—differences that are best explained by prevailing mood and intent. Indeed, these distinctions are captured by the predominant moods of the different hyper-communities, respectively *curious*, *cheerful* or *happy* versus *calm*, *accomplished* and *creative*.

### 7.3.4 ANEW-based hyper-groups

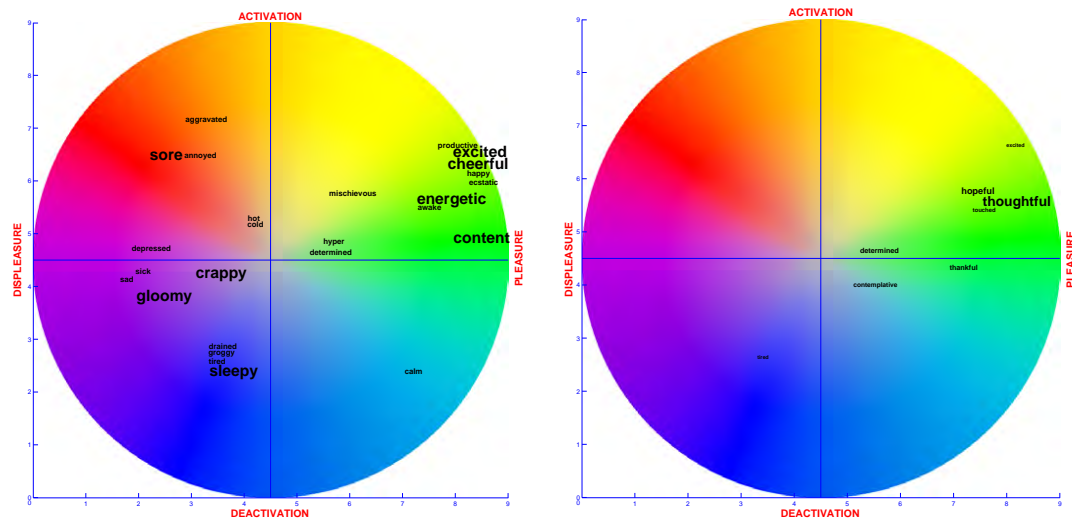
Clustering based on ANEW features as proxy mood yielded 15 hyper-communities, which are listed in Table 7.4. Of these, five consisted of communities with matching Livejournal categories (for example, *curlyhair*, *beauty101*, *dyed\_hair*, and *vintage-hair* all classified as Fashion-Style). Two hyper-communities are examples of the sub-category distinctions returned by the topic-based clustering:  $\{\textit{macintosh}, \textit{computer\_help}, \textit{computerhelp}, \textit{ipod}, \textit{webdesign}\}$  and  $\{\textit{worldofwarcraft}, \textit{gamers}, \textit{wow\_ladies}\}$  are both from Livejournal’s Gaming-Technology category.

### 7.3.5 Discussion

It is not surprising that the different community representations lead to hyper-communities that reflect these varying emphases. Topic-based representation is the method of choice for recovering hierarchy within, and associations across, Livejournal’s canonical topic categories. Likewise, the results for the mood-based representation indicate an ability to recover non-topical features of a community such as prevailing intent and atmosphere of discussion. However, contrary to expectations, ANEW does not appear to be a well-suited and cheap alternative to mood-based representation for the task of hyper-community detection.

The clustering results for LIWC’s psycholinguistic representation are worthy of follow-up. LIWC offers a wide scope of classification—due to including topical, linguistic, stylistic and mood categories—yet is cheap to obtain. Some of the distinctions captured by the hyper-communities arising from LIWC representation are a kind of topic + atmosphere that seems relevant to the Web 2.0 denizen, who is faced with a surfeit of choice and whose decision as to which community they will invest in may turn on the presence of more than one characteristic of the content. Consequently, psycholinguistic analysis demonstrates potential for use in community recommendation (and analysis).

Of course, there are more dimensions than community along which to break-down and re-factor the Livejournal corpus; for example, user and discussion. Figure 7.7 contains plots of mood aggregated from the posts of two different users. Figure 7.7a



(a) High mood variance (user *live4thin0812* from *24\_7\_posting* community).

(b) Low mood variance (user *sparowe* from *devotional365* community).

Figure 7.7: Aggregated mood distribution for two users, plotted by valence (pleasure) x-axis, and arousal (activation) y-axis. Larger mood labels indicate more frequent occurrence.

represents a user with high mood variance—this user’s moods are distributed across three quadrants of the mood circle, with a few instances in the fourth; Figure 7.7b represents a user with low mood variance—this user’s moods tend to be restricted to positive valence with average arousal. User profiling of this kind is interesting in and of itself, but when correlated with communities, it offers additional insight into the user and/or community. For example, we have found users who appear to project different personas conditional on the community of posting—with the difference captured by valence. Table 7.8 records the average valence of moods tagged by those users joining more than one community. In general, it is found that the valence of the same user in different communities is similar, barring the exception of *dark\_weezing*.<sup>3</sup> This user tags more positive moods when posting to *same\_law*,<sup>4</sup> of which he is the administrator, but with neutral moods when posting to *therightfangirl*<sup>5</sup>—a community ‘where right-wing fanfolks could talk politics’.

<sup>3</sup><http://dark-weezing.livejournal.com/profile>, retrieved October 2011.

<sup>4</sup>[http://community.livejournal.com/same\\_law](http://community.livejournal.com/same_law), retrieved October 2011.

<sup>5</sup><http://community.livejournal.com/therightfangirl>, retrieved October 2011.

User	Community	Number of posts	Average valence
sadandangry	__postsecret	12	3.4
	dear_stupid	21	3.4
	dear_you	23	4.4
butterflycell	jantolution	54	4.8
	jackxianto	57	4.8
autobot_339	dragonbrood	17	7.2
	dragonspam	32	7.1
dark_weezing	therightfangirl	69	5.2
	same_law	119	6.3

Table 7.8: Examples of users joining more than one community and their valence.

Community	Members	Posts	Community	Members	Posts
20sknitters	518	2,241	egl	654	1,286
addme25_and_up	1,133	1,800	glee_tv	289	627
beauty101	706	1,472	macintosh	712	2,794
bjorkish	561	2,293	newyorkers	1,384	6,774
blackfolk	257	1,303	ofmornings	174	385
calmallamadown	382	1,246	parenting101	773	3,540
cat_lovers	546	1,344	sgagenrefinders	232	485
charloft	148	283	todayirealized	965	4,684
color_theory	552	1,609	walmart_employe	444	2,124
dog_lovers	572	2,089	webdesign	427	1,001

Table 7.9: Communities used in the community membership prediction: the number of members and posts.

## 7.4 Community Prediction

In the previous sections we have investigated the problem of detecting hyper-communities, which are groups of communities similar in mood, topic and linguistic style. In this section, we examine if these features are predictive of community membership. Twenty communities, considered exemplars of the clusters in Section 7.3, are selected for a community prediction task, as the features are assumed to be discriminative among them. The selected communities consist of 11,429 users and 39,380 blog posts. The number of members and the number of posts made by members in these communities are shown in Table 7.9. Two predictions will be attempted: (1) by which community a post is made and (2) to which community a blogger belongs.

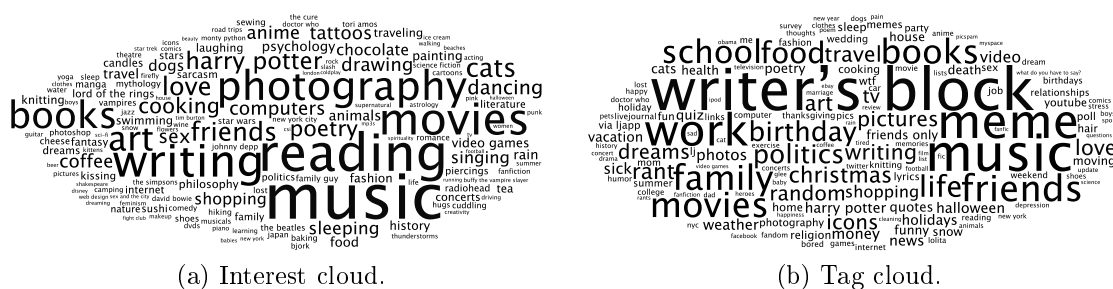
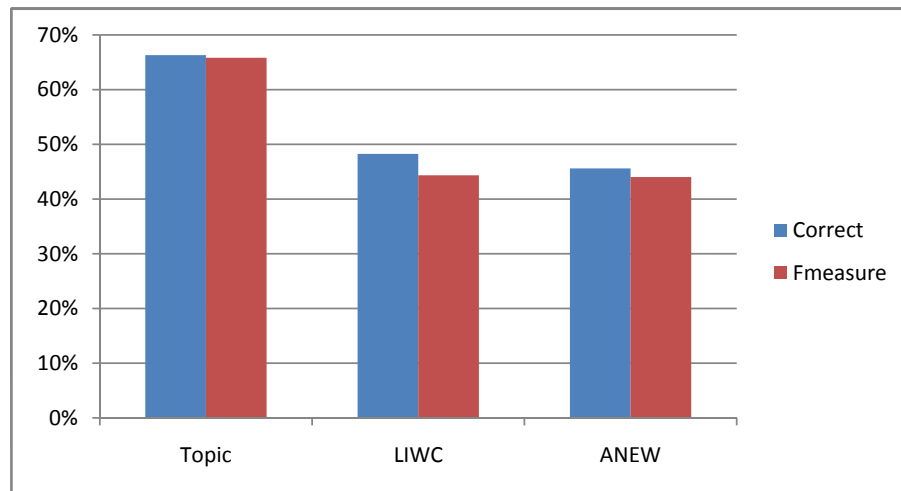


Figure 7.8: Interests and tags as features.

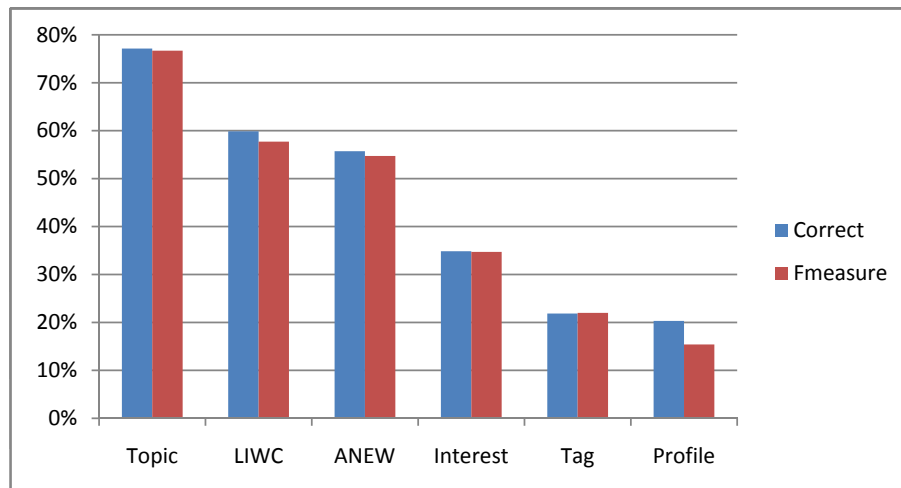
### 7.4.1 Features

Three feature sets are used to predict community for a post or for a user. These are topics (50 topics), psycholinguistic features (LIWC, 68 variables) and sentiment-bearing lexicon (ANEW, 1,034 words). To predict membership of a user, the following three additional feature sets are used:

- Top 1034 interests: For comparison with ANEW, we use the same number (1,034) of top interests users declared in their profiles as a feature set for community prediction. They are shown in Figure 7.8a. Of these, the top 10 are *music*, *reading*, *writing*, *movies*, *photography*, *books*, *art*, *cats*, *friends* and *love*.
- Top 1034 tags: Different from interests, tags are words or short phrases used to organise journal entries, optionally added when users compose a new entry or edit an old entry. They help users browse the posts under a given tag. The same number of ANEW words and interests (1,034) is used to reveal the top tags for use as a feature in community prediction, as shown in Figure 7.8b. The top 10 tags are *writer's block*, *music*, *meme*, *work*, *family*, *friends*, *movies*, *books*, *school* and *food*. These are similar to the interests.
- Profiles (13 variables): The data gleaned from profiles includes the number of days on the web, number of posts, number of comments received, number of comments posted, gender, age, nationality, state (if American), number of friends, number of followers, number of group memberships, number of tags and number of interests.



(a) By posts.



(b) By users.

Figure 7.9: Community classification results.

## 7.4.2 Community prediction

Our experimental design examines the effect of all features mentioned above on community prediction. Therefore, a classifier is used across the feature sets.

The first prediction is to assign a blog post to a community. Denote by  $\mathcal{B}$  the corpus of all  $N$  blog posts. Given a blog post  $d \in \mathcal{B}$ , we are interested in predicting the community of the document based on textual features extracted from  $d$ . The feature vector  $\mathbf{x}^{(d)} = [\dots, \mathbf{x}_i^{(d)}, \dots]$  is a vector of topic, LIWC or ANEW usage. When topics are the features,  $\mathbf{x}_i^{(d)}$  is the probability of topic  $j$  in document  $d$ . If LIWC processes are the features,  $\mathbf{x}_i^{(d)}$  represents the quantity of the process  $i$  in document  $d$ . In the ANEW case,  $\mathbf{x}^{(d)}$  is the distribution of ANEW words in document  $d$ .

The second prediction is to assign a user to a community. Let  $\mathcal{U} = \{u_1, \dots, u_M\}$  be a corpus of  $M$  users ( $M \leq N$ ). When topics, LIWC processes or ANEW lexicon are the features for predicting community membership of a user,  $u_m$  is a collated vector from all documents in  $\mathcal{B}$  made by user  $m$ . A user is also characterised by his interests, tags or profile properties shown in the ‘About Me’ page. In these cases,  $u_m$  is the distribution of interests or tags or the values of his profile characteristics.

An SVM classifier is employed for community prediction. In particular, the Weka implementation of SMO is used for the prediction task. For each run, we use 10-fold cross-validation, repeat 10 runs and report the average result. To evaluate the results we report two commonly used measures: accuracy and F-score (which is measured based on recall and precision, as defined in Section 2.4.1.4).

The results of predicting community for a post are presented in Figure 7.9a. As can be seen, prediction is best achieved when using topics as features, achieving an accuracy of more than 65 per cent. The LIWC and ANEW features also perform well, at a lower cost, gaining an accuracy of approximately 45 per cent in a 20-class prediction.

Figure 7.9b shows the results for predicting the community of a user. Similar to the results of community prediction for a post, topics are the best features for the task (77 per cent). The linguistic styles, through LIWC groups and the ANEW lexicon, are also effective in predicting community membership for users at approximately



55 per cent for the 20-class prediction. Using tags or the information revealed in bloggers' profiles as features in community predictions gains accuracy of less than 35 per cent, making this method less effective than the topic, linguistic or affective lexicon cases.

## 7.5 Conclusion

We have investigated the problem of discovering hyper-groups of communities by using topics, sentiment information and psycholinguistic properties of the posts of members. The sentiment information was extracted in two ways: (1) using mood tags and (2) inferring from emotion-bearing words used in the blog posts. We presented an unsupervised approach based on AP, a non-parametric clustering algorithm that does not require the number of clusters a priori, to detect hyper-groups of communities in the blogosphere and to reveal interesting content-based, sentiment-based and linguistic-based grouping behaviours.

We have proposed a novel approach for addressing hyper-community detection based on users' sentiment. The problem of sentiment-based clustering for community structure discovery is rich with many interesting open aspects to be explored. The grouping of meta-communities based on sentiment information can be a social indicator, having potential applications in, for example, mental health—by targeting support or surveillance to communities with negative mood—or in marketing—by targeting customer communities having the same sentiment on similar topics.

Further, the psycholinguistic hyper-groups detected provide insight into the language styles of people in specific categories. For example, bloggers in Fashion-Style categories favour the use of spoken language in their on-line diaries. In addition to this, topical meta-communities enable users to find suitable communities based on their interests.

We have also investigated the capacity of a range of features to predict community membership for a post or a user. Topics are the best features for the predictions in both cases. Without the need for a feature selection stage, the results for the predictions using sentiment information (through ANEW) are comparable to topics, but

---

at a lighter computational cost. This indicates a potential application of sentiment information for the purpose of analysing the networking properties of social media.

# Chapter 8

## Conclusion

### 8.1 Summary

This thesis has presented a sentiment-based approach to pattern discovery and classification in social media. In seeking the building blocks for sentiment analysis in social media, the thesis has investigated numerous feature spaces and feature-selection methods to learn what can be transferred from generic text categorisation. It has also introduced two feature sets originating from psychology (ANEW and LIWC) to mood classification, which perform comparatively well at a fraction of the computational cost of supervised schemes. Switching to an unsupervised framework, this thesis then used the affective lexicon to reveal intrinsic patterns in the expressed mood of a large corpus of social media documents. The global patterns of moods resemble the core affect model for the structure of human emotion proposed within psychology. This data-driven organisation of moods enables the discovery of mood synonymy. Analysis of the correlation between mood labels and ANEW terms extends the application of the discovered patterns to texts that lack mood ground truth.

This thesis then turned to the use of sentiment information conveyed in social media to detect real-world events. Through comparison with ranked events in official media (CNN), and a novel quantitative measure of event coherence based on on-line archive (Google Timeline), the sentiment time series demonstrated effectiveness in capturing

real-world events by using the shared emotional responses from a large population. This novel use of sentiment as a ‘sensor’ to extract events was then examined using an on-line implementation for burst detection, thus making the event detection problem practicable for large-scale and streaming data. Intuitively appealing patterns were discovered for the aggregate SI at weekly and monthly periodicities, as were major one-off events. When the analysis focused on particular moods in isolation, finer-grained insights were obtained (for example, the inverse weekly relationship between ‘*working*’ and ‘*drunk*’). Bursts of particular moods were found to correlate with particular kinds of events, thus allowing for a finer degree of inference.

Affect information conveyed by social media text was then used to infer about the impact of age and personality on the way bloggers express themselves in their blogs. Mood was analysed in comparison with the topics and language styles of the text. Sentiment information was shown to be a powerful predictor of user age and social connectivity, offering a potential means of investigating related attributes, such as personality type, on influence.

Finally, sentiment information was incorporated into studies of the formation of communities, in conjunction with the topics discussed and linguistic styles used by the members of those communities. Community clusters formed by topical representation were similar to their topical ground truth, as predicted. Interesting insights were gained from the other representations. For example, mood was able to capture similarity in the tone of discussion hosted by a community; and psycholinguistic features—bridging, as they do, topical, stylistic, sentiment and cognitive elements—were able to capture fusions of these features that we interpret as a kind of meta-genre of community, not easily represented by any one feature set. Posts and users were classified into communities based on topic, mood and psycholinguistic features. In addition, user classification was performed with users’ tagging behaviour and self-expressed interests. Again, the cheap mood and psycholinguistic features performed well compared to the much more computationally expensive topic-based representation. The hyper-communities grouped based on the sentiment information can be used in mental health, for example in monitoring those communities in low valence and high arousal moods. The sentiment information has also proven to be a powerful predictor of community membership, suggesting an alternative mode for studying on-line social networking.

## 8.2 Future Directions

The work undertaken in this thesis suggests several potential areas for future research.

First, the feature spaces in sentiment analysis in this thesis use a ‘bag-of-word’ representation, neglecting the context relationships among elements in a blog text. Future work should integrate context information in searching for more effective feature sets for mood classification.

Second, the application of a quantitative measure as a universal standard for sentiment analysis in this thesis, such as using the valences of ANEW words for happiness estimation, is not completely appropriate, as the values were not rendered from the context of social media. Therefore, learning new affective lexicons for this new media automatically is worth consideration in future research.

Third, for event detection and evaluation, since the concept of ‘significant events’ varies across cultures, objective measures for event evaluation should be devised. Given that user profile information such as location, gender, age and interests are often publicly available, or else can be estimated using methods described in this thesis or elsewhere, it would be feasible to perform event extraction on sub-corpora created by partitioning the global corpus along these dimensions. Comparison of which groups respond with what moods to which events would be of interest to researchers from diverse fields. In addition, due to the mixed property of ‘valence’, purer measures of emotions should be taken into the computation of SI.

Fourth, for the egocentric view of social media, a more detailed and longitudinal survey of those whose on-line profiles have been studied here would offer insight into the relationship between their virtual and real-world presences. Two questions worthy of further study are: (1) whether there is a stronger tie among those who are friends in both worlds and (2) whether the life people express in the virtual world is the real-world extended or idealised.

Finally, with regard to the networking properties of social media, we have not considered the temporal dimension of communities, or the adding or removing of links. Communities are dynamic entities, with their own stories and life cycles. The for-

mation and dissolution patterns of friendship in social networks can help researchers understand certain social phenomena, such as the idea that to add someone to your friend list is easier than to remove them, no matter who they are. Therefore, studies into the time series of ‘friending’ and ‘unfriending’ should be conducted in the near future.

# Appendix A

## Appendix

### A.1 Mood Hierarchical Tree

In Livejournal, 132 predefined moods are organised in a hierarchical tree, where those moods conveying similar sentiment are on the same branch of the tree.<sup>1</sup> Figure A.1 shows examples of this tree.

### A.2 Sentiment Burst Detection and Retrieval System

A demonstration system of the bursts and events is built based on the framework proposed in Chapter 5. Figures A.2 and A.3 display screen-shots from the Web user interface. A user can enter a set of moods and/or the timeframe to query (Figure A.2) and the software subsequently returns bursts associated with those moods with timestamp during the bursty duration (Figure A.3). For example, a query of ‘angry, sad’ and duration ‘11 September 2001 to 31 December 2004’ results in 3 and 8 bursty events corresponding to the moods angry and sad respectively. Figure A.3 provides a detailed result including topics and entities for the event ‘9/11

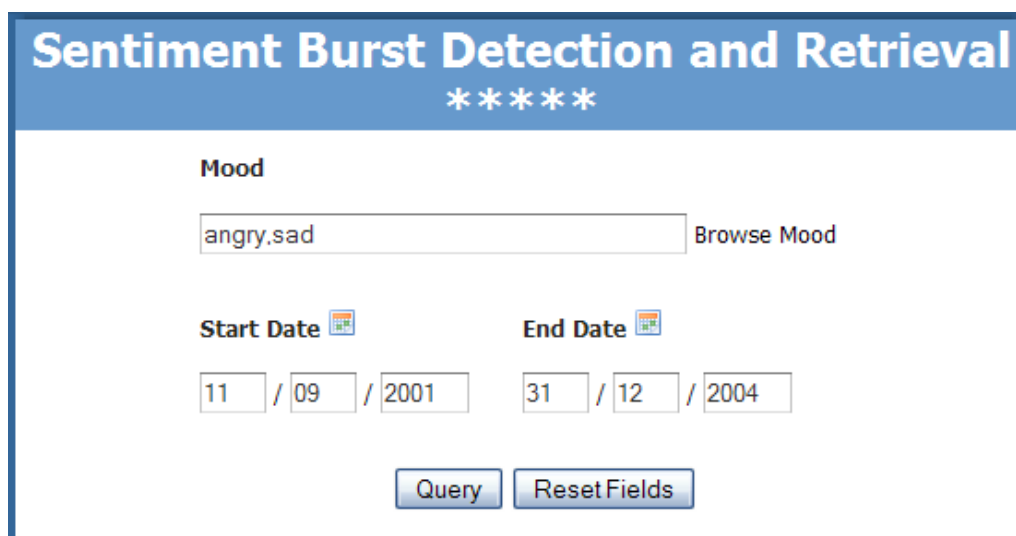
---

<sup>1</sup><http://www.livejournal.com/moodlist.bml>, retrieved November 2011.



Figure A.1: Happy, sad and angry and related moods in the hierarchical layout.





**Sentiment Burst Detection and Retrieval**  
\*\*\*\*\*

**Mood**

angry,sad

**Start Date**

11 / 09 / 2001

**End Date**

31 / 12 / 2004

Figure A.2: The input Web user interface for querying bursty moods and related events.

attacks' and the timeline for two other events associated with the mood angry: 'Iraq war' and 'US election'.

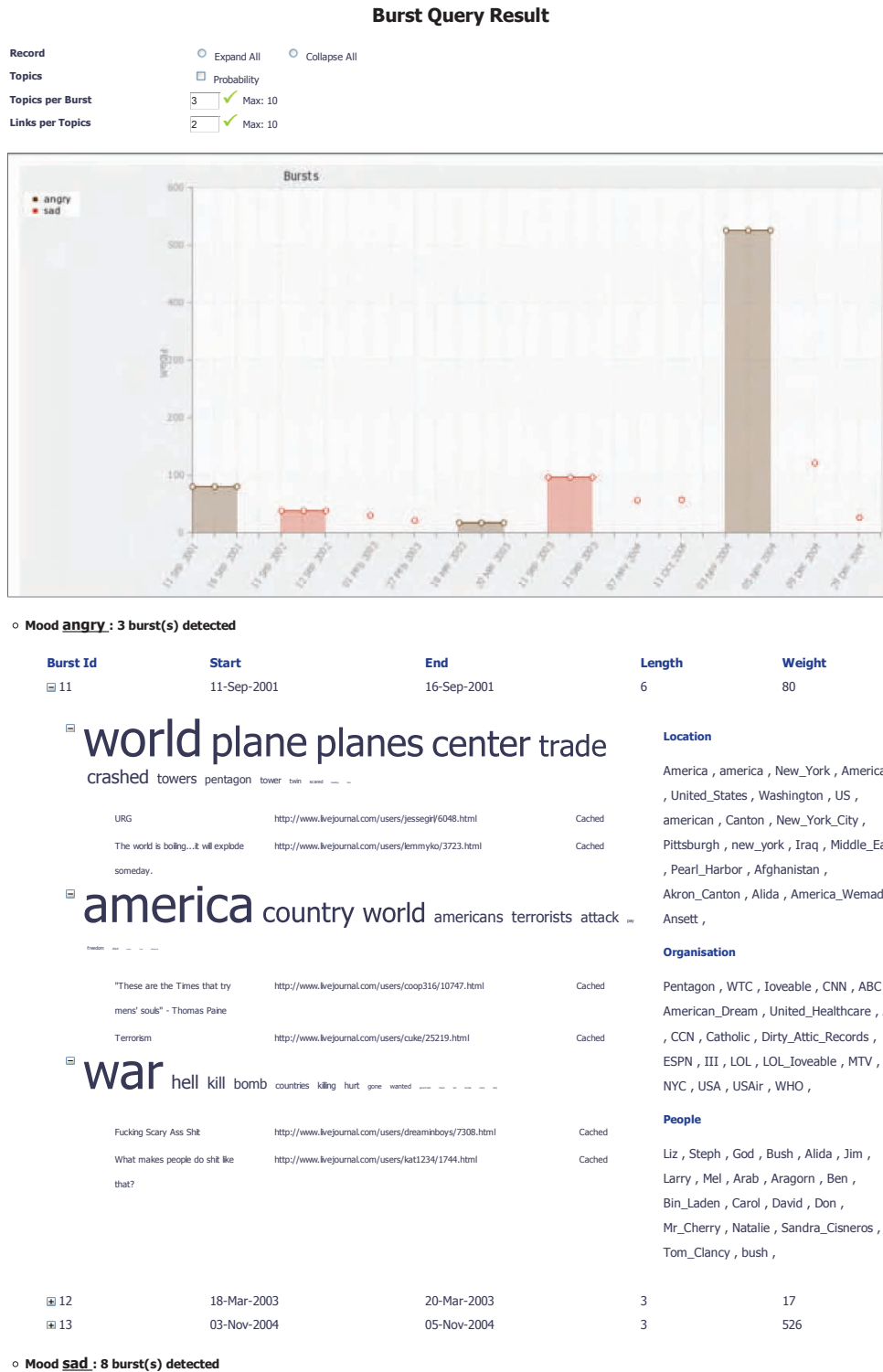


Figure A.3: Partial result of the bursty moods, related events, topics and named entities returned for the query shown in Figure A.2.

# Bibliography

- B. Adams, D. Phung, and S. Venkatesh. Discovery of latent subcommunities in a blog's readership. *ACM Transactions on the Web*, 4(3):1–30, 2010.
- N. Agarwal and H. Liu. Blogosphere: Research issues, applications, and tools. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2008.
- J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 1998.
- A. Andreevskaia and S. Bergler. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.
- M.D. Back, A.C.P. Küfner, and B. Egloff. The emotional timeline of September 11, 2001. *Psychological Science*, 21(10):1417, 2010.
- L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2006.
- E. Bakshy, J.M. Hofman, W.A. Mason, and D.J. Watts. Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.
- K. Balog, G. Mishne, and M. de Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

- B. Berendt and C. Hanser. Tags are not metadata, but ‘just more content’—to some people. *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2007.
- D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- E. Boiy and M.F. Moens. A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*, 12:526–558, 2009.
- I. Borg and P.J.F. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Verlag, 2005.
- P.J. Bracken, J.E. Giller, and D. Summerfield. Psychological responses to war and atrocity: The limitations of current concepts. *Social Science & Medicine*, 40(8): 1073–1082, 1995.
- M.M. Bradley and P.J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. *University of Florida*, 1999.
- C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
- G. Chastain, P.S. Seibert, and F.R. Ferraro. Mood and lexical access of positive, negative, and neutral words. *Journal of General Psychology*, 122(2):137–157, 1995.
- N.A. Christakis and J.H. Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown and Company, 2009.
- A.Y.K. Chua, K. Razikin, and D.H. Goh. Social tags as news event detectors. *Journal of Information Science*, 37(1):3, 2011.
- T. Church, M. S. Katigbak, J.A.S. Reyes, and S.M. Jensen. Language and organisation of Filipino emotion concepts: Comparing emotion concepts and dimensions across cultures. *Cognition & Emotion*, 12(1):63–92, 1998.
- G. Colombetti. Appraising valence. *Journal of Consciousness Studies*, 12, 8(10): 103–126, 2005.
- R. Coontz. Blogs: Happiness barometers? *Science*, 325:5941, 2009.

- S.R. Das and M.Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.
- A. Das Sarma, A. Jain, and C. Yu. Dynamic relationship and event discovery. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.
- J.M. Dewaele and A. Furnham. Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences*, 28(2):355–365, 2000.
- X. Ding, B. Liu, and P.S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2008.
- P.S. Dodds and C.M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.
- R. I. M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.
- P. Ekman and W.V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124, 1971.
- E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(90001):5220–5227, 2004.
- J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050, 2007.
- B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.
- M. Freyd. Introverts and extroverts. *Psychological Review*, 31(1):74–87, 1924.
- T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura. Identification of bursts in a document stream. In *Proceedings of the First International Workshop on Knowledge Discovery in Data Streams*, 2004.

- D. Galati, B. Sini, C. Tinti, and S. Testa. The lexicon of emotion in the neo-Latin languages. *Social Science Information*, 47(2):205, 2008.
- T. Gehm and K. R. Scherer. *Factors determining the dimensions of subjective emotional space*, chapter 5. Lawrence Erlbaum Associates, 1988.
- E. Gilbert and K. Karahalios. Widespread worry and the stock market. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.
- J. Giles. Blogs and tweets could predict the future. *The New Scientist*, 206(2765):20–21, 2010.
- M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.
- N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *Proceedings of the WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.
- T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(90001):5228–5235, 2004.
- D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2004.
- D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2005.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The Weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- C. Hayes and P. Avesani. Using tags and clustering to identify topic-relevant blogs. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2007.

- Q. He, K. Chang, and E.P. Lim. Using burstiness to improve clustering of topics in news streams. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2007a.
- Q. He, K. Chang, E.P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2007b.
- M. A. Hearst, M. Hurst, and S. T. Dumais. What should blog search look like? In *Proceedings of the ACM Workshop on Search in Social Media*, 2008.
- S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu. Conversations in the blogosphere: An analysis “from the bottom up”. In *Proceedings of the Hawaii International Conference on System Sciences*, 2005.
- T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.
- T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.
- M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2004.
- M. Hu and B. Liu. Opinion extraction and summarization on the web. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2006.
- X. Hu and J.S. Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the International Conference on Music Information Retrieval*, 2007.
- V. Jijkoun, M. de Rijke, and W. Weerkamp. Generating focused topic-specific sentiment lexicons. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.
- T. Joachims. Text categorization with support vector machines: Learning with many relevant. In *Proceedings of the European Conference on Machine Learning*, 1998.

- A. M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68, 2010.
- F. Keshtkar and D. Inkpen. Using sentiment orientation features for mood classification in blogs. In *Proceedings of the International IEEE Conference on Natural Language Processing and Knowledge Engineering*, 2009.
- J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre. Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3):241 – 251, 2011.
- S.M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2004.
- J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.
- T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
- T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. SOM PAK: The self-organizing map program package. Technical report, Helsinki University of Technology, 1996.
- A.D.I. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (SIGCHI)*, 2010.
- L.W. Ku, L.Y. Lee, T.H. Wu, and H.H. Chen. Major topic detection and its application to opinion summarization. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2005.
- L.W. Ku, Y.T. Liang, and H.H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
- S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.



- R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2003.
- R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2006.
- G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2004.
- C.A.C. Lampe, N. Ellison, and C. Steinfield. A familiar face (book): Profile elements as signals in an online social network. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (SIGCHI)*, 2007.
- G. Leshed and J.J. Kaye. Understanding how bloggers feel: Recognizing affect in blog posts. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (SIGCHI)*, 2006.
- D.D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the European Conference on Machine Learning*, 1998.
- C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2009.
- B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2005.
- D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, 2012.
- F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.

- J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Topic detection and tracking with spatio-temporal evidence. *Advances in Information Retrieval*, 2003.
- H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.
- C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.
- C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the International Conference on Hypertext and Hypermedia*, 2006.
- I.B. Mauss and M.D. Robinson. Measures of emotion: A review. *Cognition & Emotion*, 23(2):209–237, 2009.
- A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007a.
- A. McCallum, X. Wang, and N. Mohanty. Joint group and topic discovery from relations and text. *Lecture Notes in Computer Science*, 4503:28, 2007b.
- G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM Workshop on Stylistic Analysis of Text for Information Access*, 2005.
- G. Mishne. Information access challenges in the blogspace. In *Proceedings of the International Workshop on Intelligent Information Access*, 2006.
- G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
- T. Mitchell. *Machine Learning*. McGraw Hill, 1997.
- G. Murray and G. Carenini. Subjectivity detection in spoken and written conversations. *Natural Language Engineering*, 17(03):397–418, 2011.
- R. Nallapati and W. Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2008.

- R.A. Negoescu, B. Adams, D. Phung, S. Venkatesh, and D. Gatica-Perez. Flickr hypergroups. In *Proceedings of the ACM International Conference on Multimedia*, 2009.
- M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45:211–236, 2008.
- S. Nowson. *The language of weblogs: A study of genre and individual differences*. PhD thesis, University of Edinburgh, 2006.
- I. Ounis, M. De Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 blog track. In *Proceedings of the Text Retrieval Conference (TREC)*, 2006.
- B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.
- B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.
- B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
- B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, 2002.
- J.W. Pennebaker and L.D. Stone. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291–301, 2003.
- J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth. The development and psychometric properties of LIWC2007. *Austin, Texas: LIWC Inc*, 2007a.
- J.W. Pennebaker, M.E. Francis, and R.J. Booth. Linguistic inquiry and word count (LIWC) [computer software]. *Austin, Texas: LIWC Inc*, 2007b.

- J.E. Phelps, R. Lewis, L. Mobilio, D. Perry, and N. Raman. Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research*, 44(4):333–348, 2004.
- E. Quintelier. Differences in political participation between young and old people. *Contemporary Politics*, 13(2):165, 2007.
- L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- D.C.V. Ramesh and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2000.
- C.M. Ridings and D. Gefen. Virtual community attraction: Why people hang out online. *Journal of Computer-Mediated Communication*, 10(1), 2004.
- E. Riloff, J. Wiebe, and W. Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference On Artificial Intelligence*, 2005.
- C. Ross, E.S. Orr, M. Sisic, J.M. Arseneault, M.G. Simmering, and R.R. Orr. Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2):578–586, 2009.
- S. Rude, E.M. Gortner, and J. Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.
- J.A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.
- J.A. Russell. Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, 45(6):1281, 1983.
- J.A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145, 2003.
- J.A. Russell. Emotion, core affect, and psychological construction. *Cognition & Emotion*, 23(7):1259–1283, 2009.

- C.L. Rusting and R.J. Larsen. Moods as sources of stimulation: Relationships between personality and desired mood states. *Personality and Individual Differences*, 18(3):321–329, 1995.
- T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2010.
- S. Sara and V. Lucy. SentiSearch: Exploring mood on the web. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.
- J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
- J. Schrammel, C. Köffel, and M. Tscheligi. Personality traits, usage patterns and information disclosure in online communities. In *Proceedings of the British Computer Society Conference on Human-Computer Interaction*, 2009.
- F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
- P.R. Shaver, U. Murdaya, and R.C. Fraley. Structure of the Indonesian emotion lexicon. *Asian Journal of Social Psychology*, 4(3):201–224, 2001.
- R.C. Silver, E.A. Holman, D.N. McIntosh, M. Poulin, and V. Gil-Rivas. Nationwide longitudinal study of psychological responses to September 11. *Journal of the American Medical Association*, 288(10):1235, 2002.
- R.B. Slatcher, C.K. Chung, J.W. Pennebaker, and L.D. Stone. Winning words: Individual differences in linguistic style among US presidential and vice presidential candidates. *Journal of Research in Personality*, 41(1):63–75, 2007.
- C.A. Smith and P.C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813, 1985.
- R. C. Solomon. Against valence (‘positive and negative emotions’). *Not Passion’s Slave*, 1(9):162–178, 2003.

- S. Somasundaran, J. Ruppenhofer, and J. Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 2007.
- X. Song, C.Y. Lin, B.L. Tseng, and M.T. Sun. Modeling and predicting personal information dissemination behavior. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2005.
- S.O. Sood and L. Vasserman. ESSE: Exploring mood on the web. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.
- L.D. Stafford, W. Ng, R.A. Moore, and K.A. Bard. Bolder, happier, smarter: The role of extraversion in positive mood and cognition. *Personality and Individual Differences*, 48(7):827–832, 2010.
- S.W. Stirman and J.W. Pennebaker. Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine*, 63(4):517, 2001.
- Y.R. Tausczik and J.W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24, 2010.
- M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.
- Y. Tsuruoka and J. Tsujii. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.
- A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Weppe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.
- J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. SOM toolbox for Matlab. Technical report, Helsinki University of Technology, 2000.

- L. Wasserman. *All of statistics: A concise course in statistical inference*. Springer Verlag, 2004.
- J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.
- T.A. Wilson. *Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity, and attitudes of private states*. PhD thesis, University of Pittsburgh, 2008.
- Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1997.
- M. Yoshida, R. Kinase, J. Kurokawa, and S. Yashiro. Multi-dimensional scaling of emotion. *Japanese Psychological Research*, 12(2):45–61, 1970.
- H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, 2003.
- K. Zhang, J. Zi, and L.G. Wu. New event detection based on indexing-tree and named entity. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2007.
- Q. Zhao, P. Mitra, and B. Chen. Temporal and information flow based event detection from social text streams. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2007.
- 

**Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.**

**ASSOCIATION FOR COMPUTING MACHINERY, INC. LICENSE  
TERMS AND CONDITIONS**

Sep 26, 2011

---

---

This is a License Agreement between Thin Nguyen ("You") and Association for Computing Machinery, Inc. ("Association for Computing Machinery, Inc.") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Association for Computing Machinery, Inc., and the payment terms and conditions.

License Number	2756250138790
License date	Sep 25, 2011
Licensed content publisher	Association for Computing Machinery, Inc.
Licensed content publication	Proceedings of second ACM SIGMM workshop on Social media
Licensed content title	Hyper-community detection in the blogosphere
Licensed content author	Thin Nguyen, et al
Licensed content date	Oct 25, 2010
Type of Use	Thesis/Dissertation
Requestor type	Author of this ACM article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	A Sentiment Based Approach to Pattern Discovery and Classification in Social Media
Expected completion date	Dec 2011
Estimated size (pages)	150
Billing Type	Credit Card
Credit card info	Visa ending in 5970
Credit card expiration	02/2013
Total	18.75 USD
Terms and Conditions	

**Rightslink Terms and Conditions for ACM Material**

1. The publisher of this copyrighted material is Association for Computing Machinery, Inc. (ACM). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at ).
2. ACM reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
3. ACM hereby grants to licensee a non-exclusive license to use or republish this ACM-copyrighted material\* in secondary works (especially for commercial distribution) with the stipulation that consent of the lead author has been obtained independently. Licenses are for one-time use in a single edition of the work, only with a maximum distribution equal to the number that you identified in the licensing process. Any additional



form of republication must be specified according to the terms included at the time of licensing.

\*Please note that ACM cannot grant republication or distribution licenses for embedded third-party material. You must confirm the ownership of figures, drawings and artwork prior to use.

4. Any form of republication or redistribution must be used within 180 days from the date stated on the license and any electronic posting is limited to a period of six months unless an extended term is selected during the licensing process. Separate subsidiary and subsequent republication licenses must be purchased to redistribute copyrighted material on an extranet. These licenses may be exercised anywhere in the world.

5. Licensee may not alter or modify the material in any manner (except that you may use, within the scope of the license granted, one or more excerpts from the copyrighted material, provided that the process of excerpting does not alter the meaning of the material or in any way reflect negatively on the publisher or any writer of the material).

6. Licensee must include the following copyright and permission notice in connection with any reproduction of the licensed material: "[Citation] © YEAR Association for Computing Machinery, Inc. Reprinted by permission." Include the article DOI as a link to the definitive version in the ACM Digital Library. Example: Charles, L. "How to Improve Digital Rights Management," Communications of the ACM, Vol. 51:12, © 2008 ACM, Inc. <http://doi.acm.org/10.1145/nnnnnn.nnnnnn> (where nnnnnn.nnnnnn is replaced by the actual number).

7. Translation of the material in any language requires an explicit license identified during the licensing process. Due to the error-prone nature of language translations, Licensee must include the following copyright and permission notice and disclaimer in connection with any reproduction of the licensed material in translation: "This translation is a derivative of ACM-copyrighted material. ACM did not prepare this translation and does not guarantee that it is an accurate copy of the originally published work. The original intellectual property contained in this work remains the property of ACM."

8. You may exercise the rights licensed immediately upon issuance of the license at the end of the licensing transaction, provided that you have disclosed complete and accurate details of your proposed use. No license is finally effective unless and until full payment is received from you (either by CCC or ACM) as provided in CCC's Billing and Payment terms and conditions.

9. If full payment is not received within 90 days from the grant of license transaction, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted.

10. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

11. ACM makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

12. You hereby indemnify and agree to hold harmless ACM and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

13. This license is personal to the requestor and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

14. This license may not be amended except in a writing signed by both parties (or, in the case of ACM, by CCC on its behalf).

15. ACM hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and ACM (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

16. This license transaction shall be governed by and construed in accordance with the laws of New York State. You hereby agree to submit to the jurisdiction of the federal and state courts located in New York for purposes of resolving any disputes that may arise in connection with this licensing transaction.

17. There are additional terms and conditions, established by Copyright Clearance Center, Inc. ("CCC") as the administrator of this licensing service that relate to billing and payment for licenses provided through this service. Those terms and conditions apply to each transaction as if they were restated here. As a user of this service, you agreed to those terms and conditions at the time that you established your account, and you may see them again at any time at <http://myaccount.copyright.com>

18. Thesis/Dissertation: This type of use requires only the minimum administrative fee. It is not a fee for permission. Further reuse of ACM content, by ProQuest/UMI or other document delivery providers, or in republication requires a separate permission license and fee. Commercial resellers of your dissertation containing this article must acquire a separate license.

#### Special Terms:

**If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500633980.**

**Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.**

**Make Payment To:**  
Copyright Clearance Center  
Dept 001  
P.O. Box 843006  
Boston, MA 02284-3006

**For suggestions or comments regarding this order, contact RightsLink Customer Support: [customercare@copyright.com](mailto:customercare@copyright.com) or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.**

**Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.**

---

---

**SPRINGER LICENSE  
TERMS AND CONDITIONS**

Sep 26, 2011

---

---

This is a License Agreement between Thin Nguyen ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2756240873883
License date	Sep 25, 2011
Licensed content publisher	Springer
Licensed content publication	Springer eBook
Licensed content title	Classification and Pattern Discovery of Mood in Weblogs
Licensed content author	Thin Nguyen
Licensed content date	May 29, 2010
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	3
Author of this Springer article	Yes and you are a contributor of the new work
Order reference number	
Title of your thesis / dissertation	A Sentiment Based Approach to Pattern Discovery and Classification in Social Media
Expected completion date	Dec 2011
Estimated size(pages)	150
Total	0.00 USD
Terms and Conditions	

### Introduction

The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

### Limited License

With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry. Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided it is password protected or on the university's intranet, destined to microfilming by UMI and University repository. For any other electronic use, please contact Springer at ([permissions.dordrecht@springer.com](mailto:permissions.dordrecht@springer.com) or [permissions.heidelberg@springer.com](mailto:permissions.heidelberg@springer.com))

The material can only be used for the purpose of defending your thesis, and with a maximum of 100 extra copies in paper.

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, provided permission is also obtained from the (co) author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well). Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

#### Altering/Modifying Material: Not Permitted

However figures and illustrations may be altered minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at [permissions.dordrecht@springer.com](mailto:permissions.dordrecht@springer.com) or [permissions.heidelberg@springer.com](mailto:permissions.heidelberg@springer.com))

#### Reservation of Rights

Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

#### Copyright Notice:

Please include the following copyright citation referencing the publication in which the material was originally published. Where wording is within brackets, please include verbatim.

"With kind permission from Springer Science+Business Media: <book/journal title, chapter/article title, volume, year of publication, page, name(s) of author(s), figure number(s), and any original (first) copyright notice displayed with material>."

**Warranties:** Springer Science + Business Media makes no representations or warranties with respect to the licensed material.

#### Indemnity

You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

#### No Transfer of License

This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.

#### No Amendment Except in Writing

This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).

#### Objection to Contrary Terms

Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer Science + Business Media (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

#### Jurisdiction

All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by the country's law in which the work was originally published.

Other terms and conditions:

v1.2

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500633977.

Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

**Make Payment To:**  
Copyright Clearance Center  
Dept 001  
P.O. Box 843006  
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: [customercare@copyright.com](mailto:customercare@copyright.com) or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

---

---

## SPRINGER LICENSE TERMS AND CONDITIONS

May 23, 2012

---

---

This is a License Agreement between Thin Nguyen ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2915010396377
License date	May 23, 2012
Licensed content publisher	Springer
Licensed content publication	Springer eBook
Licensed content title	Emotional Reactions to Real-World Events in Social Networks
Licensed content author	Thin Nguyen
Licensed content date	Feb 21, 2012
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	3
Author of this Springer article	Yes and you are a contributor of the new work
Order reference number	
Title of your thesis / dissertation	A Sentiment Based Approach to Pattern Discovery and Classification in Social Media
Expected completion date	May 2012
Estimated size(pages)	200
Total	0.00 USD
Terms and Conditions	

### Introduction

The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

### Limited License

With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry.

Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided its password protected or on the university's intranet or repository, including UMI (according to the definition at the Sherpa website: <http://www.sherpa.ac.uk/romeo/>). For any other electronic use, please contact Springer at ([permissions.dordrecht@springer.com](mailto:permissions.dordrecht@springer.com) or [permissions.heidelberg@springer.com](mailto:permissions.heidelberg@springer.com)).

The material can only be used for the purpose of defending your thesis, and with a maximum of 100 extra copies in paper.

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, provided permission is also obtained from the (co) author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well).

Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

#### Altering/Modifying Material: Not Permitted

You may not alter or modify the material in any manner. Abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at ([permissions.dordrecht@springer.com](mailto:permissions.dordrecht@springer.com) or [permissions.heidelberg@springer.com](mailto:permissions.heidelberg@springer.com)))

#### Reservation of Rights

Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

#### Copyright Notice:Disclaimer

You must include the following copyright and permission notice in connection with any reproduction of the licensed material: "Springer and the original publisher /journal title, volume, year of publication, page, chapter/article title, name(s) of author(s), figure number(s), original copyright notice) is given to the publication in which the material was originally published, by adding; with kind permission from Springer Science and Business Media"

#### Warranties: None

Example 1: Springer Science + Business Media makes no representations or warranties with respect to the licensed material.

Example 2: Springer Science + Business Media makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

#### Indemnity

You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

#### No Transfer of License

This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.

#### No Amendment Except in Writing

This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).

#### Objection to Contrary Terms

Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer Science + Business Media (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

#### Jurisdiction

All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in The Netherlands, in accordance with Dutch law, and to be conducted under the Rules of the 'Netherlands Arbitrage Instituut' (Netherlands Institute of Arbitration). **OR:**

**All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in the Federal Republic of Germany, in accordance with German law.**

#### Other terms and conditions:

##### v1.3

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500785557.

Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

#### Make Payment To:

Copyright Clearance Center  
Dept 001  
P.O. Box 843006  
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: [customercare@copyright.com](mailto:customercare@copyright.com) or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.





## SPRINGER LICENSE TERMS AND CONDITIONS

Oct 16, 2011

---

---

This is a License Agreement between Thin Nguyen ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2770680355427
License date	Oct 16, 2011
Licensed content publisher	Springer
Licensed content publication	Springer eBook
Licensed content title	Prediction of Age, Sentiment, and Connectivity from Social Media Text
Licensed content author	Thin Nguyen
Licensed content date	Oct 6, 2011
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	3
Author of this Springer article	Yes and you are a contributor of the new work
Order reference number	
Title of your thesis / dissertation	A Sentiment Based Approach to Pattern Discovery and Classification in Social Media
Expected completion date	Dec 2011
Estimated size(pages)	150
Total	0.00 USD
Terms and Conditions	

### Introduction

The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

### Limited License

With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry. Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided it is password protected or on the university's intranet, destined to microfilming by UMI and University repository. For any other electronic use, please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

The material can only be used for the purpose of defending your thesis, and with a maximum of 100 extra copies in paper.

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, provided permission is also obtained from the (co) author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another

source, authorization from that source is required as well). Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

#### Altering/Modifying Material: Not Permitted

However figures and illustrations may be altered minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at [permissions.dordrecht@springer.com](mailto:permissions.dordrecht@springer.com) or [permissions.heidelberg@springer.com](mailto:permissions.heidelberg@springer.com))

#### Reservation of Rights

Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

#### Copyright Notice:

Please include the following copyright citation referencing the publication in which the material was originally published. Where wording is within brackets, please include verbatim.

"With kind permission from Springer Science+Business Media: <book/journal title, chapter/article title, volume, year of publication, page, name(s) of author(s), figure number(s), and any original (first) copyright notice displayed with material>."

**Warranties:** Springer Science + Business Media makes no representations or warranties with respect to the licensed material.

#### Indemnity

You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

#### No Transfer of License

This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.

#### No Amendment Except in Writing

This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).

#### Objection to Contrary Terms

Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer Science + Business Media (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

#### Jurisdiction

All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by the country's law in which the work was originally published.

Other terms and conditions:

v1.2

**If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500646279.**

**Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.**

**Make Payment To:  
Copyright Clearance Center  
Dept 001  
P.O. Box 843006  
Boston, MA 02284-3006**

**For suggestions or comments regarding this order, contact RightsLink Customer Support: [customercare@copyright.com](mailto:customercare@copyright.com) or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.**

**Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.**

---

---

# The Student Research Workshop at the 48th Annual Meeting of the Association for Computational Linguistics

## SRW @ ACL 2010

### Final Submission Page

**Title of Submission:** Mood Patterns and Affective Lexicon Access in Weblogs

**Authors:** Thin Nguyen

#### Part A. Information for Proceedings

1. Enter final paper title as it appears on the paper:

Mood Patterns and Affective Lexicon Access in Weblogs

2. Enter short version of title for conference schedule:

Mood Patterns and Affective Lexicon Access in Weblogs

3. Enter final author list, in the order in which the authors appear on the paper.

	First name(s)	Last name	Email	Affiliation
#1	Thin	Nguyen	thin.nguyen@postgrad.c	Curtin University of Technolog

[More Authors](#)   [Fewer Authors](#)

4. Review your abstract in the box below, and make any changes you deem to be appropriate.

The emergence of social media brings chances, but also challenges, to linguistic analysis. In this paper we investigate a novel problem of discovering patterns based on emotion and the association of moods and affective lexicon usage in blogosphere, a representative for social media. We propose the use of normative emotional scores for English words in combination with a psychological model of emotion measurement and a nonparametric clustering process for inferring meaningful emotion patterns automatically from data. Our results on a dataset consisting of more than 17 million mood-groundtruthed blogposts have shown interesting evidence of the emotion patterns automatically discovered that match well with the core-affect emotion model theorized by psychologists. We then present a method based on information theory to discover the association of moods and affective lexicon usage in the new media.

5. Enter length of paper in pages (enter a single integer, which is the total number of pages, including references):

6

---

## Part B. Copyright

Please read the following copyright form. Then type your name below, which signifies your agreement to the terms herein. Authors of exceptional cases should type the string "NA" to the "signature" field and forward an authorized "License to Publish" or equivalent forms to the PC chairs at [program@acl2010.org](mailto:program@acl2010.org). (*Authors of workshop papers, if requested, should send the "license" to their respective workshop PC chairs, not the main conference PC chairs.*)

Copyright to the above work (including, without limitation, the right to publish the work in whole or in part in any and all forms and media, now or hereafter known) is hereby transferred to the Association for Computational Linguistics (ACL), effective as of the date of this agreement, on the understanding that the work has been accepted for presentation at a meeting sponsored by the ACL and for publication in the proceedings of that meeting. However, each of the authors and the employers for whom the work was performed reserve all other rights, specifically including the following: (1) All proprietary rights other than copyright and publication rights transferred to ACL; (2) The right to publish in a journal or collection or to be used in future works of the author's own (such as articles or books) all or part of this work, provided that acknowledgment is given to the ACL and a full citation to its publication in the particular proceedings is included; (3) The right to make oral presentation of the material in any forum; (4) The right to make copies of the work for internal distribution within the author's organization and for external distribution as a preprint, reprint, technical report, or related class of document.

Exceptions: Some of the foregoing conditions will not apply if the work is not copyrighted or has a non-transferrable copyright, because it has been produced by government employees acting within the scope of their employment. The exceptions are:

- The work has been produced by government employees acting within the scope of their employment and is not copyrighted.
- The work has been produced by government employees acting within the scope of their employment and has non-transferrable copyright.
- The work has been produced by government employees acting within the scope of their employment and the government reserves the right to reproduce the work for government purposes.

By typing my signature below, I confirm that all authors of the work have agreed to the above and that I am authorized to execute this agreement on their behalf.

Signature (type your name or "NA" if non-transferrable):

Thin Nguyen

Job title (if not author):

Name and address of organization:

Curtin University of Technology  
Bentley, WA 6102, Australia

---

## Part C. Upload Camera-Ready Copy

Please provide a pdf copy of the final version of your paper. Please check the following:

- All fonts are embedded in the PDF.
- The pdf file is formatted using ACL 2010 style files. In particular,

- The paper size should be A4 (NOT Letter).
- The margins should follow the camera-ready instructions.
- The paper title should not exceed the left and right margins.
- Page numbers should NOT be given.
- Authors are listed on the paper.
- You have properly filled out parts A and B above, including the copyright transfer form.

See [http://acl2010.org/authors\\_final.html](http://acl2010.org/authors_final.html) for complete instructions and answers to common questions.

After you successfully submit your paper, please come back to this page. **A test button will appear below** to let you check your submission and fix the margins.

A version of the camera-ready copy has been successfully uploaded. Below is a list of possible actions you may take. Please do not forget to check margins of the submission, by clicking on the `Test` button below.

**Note:** Given the correct A4 paper size when generating the final PDF, you should almost always get the correct margins if you also used the Latex/dot templates provided on the official ACL conference site.

Here are more specific steps to better identify the margin problems if you find any problems using the `Test` button:

1. Check whether your PDF is formatted for A4 paper size. If you are not using the correct paper size, the margins will almost always be incorrect. Using the Acrobat PDF reader, open your PDF and point your cursor to the bottom left corner; the current page size should appear there.
2. Print your PDF file *and* the sample camera-ready PDF from the conference site. See if the text in the two documents occupy roughly the same amount of space, and are well-aligned. (You can also measure the margins using a ruler. But the printer may produce some minor deviation due to mechanical problems. If your PDF and the sample PDF align well, you can safely ignore the potential mechanical problems.)
3. If the text area does not align well, check if the settings for margins are correct in your MS Word or Latex template. If you find a big mismatch due to accidental changes to the parameters for the page layout, simply correct them and re-generate the PDF file.
4. If the above steps do not work, you can try to correct the margins using the margin correction parameters in the final submission page. You should also send a copy of the PDF to your workshop chairs (or book chair) for double checking. For main conference papers, send the copy to the publication chairs.

---

[Download the current version of the submission.](#)

---

Read the instructions for submission test and margin correction  
([click here to show/hide instructions](#))

Margin correction:

[Click here to enter a margin correction](#)

---

Test the correctness of the current version of the submission

Click the "Test" button to test the submission

---

(Note: Javascript must be enabled to make a submission.)

START Conference Manager (V2.56.8 - Rev. 1655)



# Association for the Advancement of Artificial Intelligence

445 BURGESS DRIVE  
MENLO PARK, CA 94025

## AAAI COPYRIGHT FORM

Title of Article/Paper: Towards Discovery of Influence and Personality Traits through Social Link Prediction

Publication in Which Article Is to Appear: Proc. of the AAAI Conf. on Weblogs and Social Media (ICWSM 2011)

Author's Name(s): Thin Nguyen, Dinh Phung, Brett Adams and Svetha Venkatesh

Please type or print your name(s) as you wish it (them) to appear in print

*(Please read and sign Part A only, unless you are a government employee and created your article/paper as part of your employment. If your work was performed under Government contract, but you are not a Government employee, sign Part A and see item 5 under returned rights.)*

### PART A—Copyright Transfer Form

The undersigned, desiring to publish the above article/paper in a publication of the Association for the Advancement of Artificial Intelligence, (AAAI), hereby transfer their copyrights in the above article/paper to the Association for the Advancement of Artificial Intelligence (AAAI), in order to deal with future requests for reprints, translations, anthologies, reproductions, excerpts, and other publications.

This grant will include, without limitation, the entire copyright in the article/paper in all countries of the world, including all renewals, extensions, and reversions thereof, whether such rights current exist or hereafter come into effect, and also the exclusive right to create electronic versions of the article/paper, to the extent that such right is not subsumed under copyright.

The undersigned warrants that they are the sole author and owner of the copyright in the above article/paper, except for those portions shown to be in quotations; that the article/paper is original throughout; and that the undersigned right to make the grants set forth above is complete and unencumbered.

If anyone brings any claim or action alleging facts that, if true, constitute a breach of any of the foregoing warranties, the undersigned will hold harmless and indemnify AAAI, their grantees, their licensees, and their distributors against any liability, whether under judgment, decree, or compromise, and any legal fees and expenses arising out of that claim or actions, and the undersigned will cooperate fully in any defense AAAI may make to such claim or action. Moreover, the undersigned agrees to cooperate in any claim or other action seeking to protect or enforce any right the undersigned has granted to AAAI in the article/paper. If any such claim or action fails because of facts that constitute a breach of any of the foregoing warranties, the undersigned agrees to reimburse whomever brings such claim or action for expenses and attorneys' fees incurred therein.

### Returned Rights

In return for these rights, AAAI hereby grants to the above authors, and the employers for whom the work was performed, royalty-free permission to:

1. Retain all proprietary rights other than copyright (such as patent rights).
2. Personal reuse of all or portions of the above article/paper in other works of their own authorship.
3. Reproduce, or have reproduced, the above article/paper for the author's personal use, or for company use provided that AAAI copyright and the source are indicated, and that the copies are not used in a way that implies AAAI endorsement of a product or service of an employer, and that the copies per se are not offered for sale. The foregoing right shall not permit the posting of the article/paper in electronic or digital form on any computer network, except by the author or the author's employer, and then only on the author's or the employer's own web page or ftp site. Such web page or ftp site, in addition to the aforementioned requirements of this Paragraph, must provide an electronic reference or link back to the AAAI electronic server, and shall not post other AAAI copyrighted materials not of the author's or the employer's creation (including tables of contents with links to other papers) without AAAI's written permission.
4. Make limited distribution of all or portions of the above article/paper prior to publication.
5. In the case of work performed under U.S. Government contract, AAAI grants the U.S. Government royalty-free permission to reproduce all or portions of the above article/paper, and to authorize others to do so, for U.S. Government purposes.

In the event the above article/paper is not accepted and published by AAAI, or is withdrawn by the author(s) before acceptance by AAAI, this agreement becomes null and void.

\_\_\_\_\_  
Author's Signature Thin NGUYEN

\_\_\_\_\_  
Date 24 March 2011

\_\_\_\_\_  
Employer for whom work was performed

\_\_\_\_\_  
Title (if not author)





**Association for the Advancement of Artificial Intelligence**  
 2275 East Bayshore Road, Suite 160  
 Palo Alto, California 94303 USA

**AAAI COPYRIGHT FORM**

Title of Article/Paper: A Sentiment-Aware Approach to Community Formation in Social Media

Publication in Which Article/Paper Is to Appear: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM 2012)

Author's Name(s): Thin Nguyen, Dinh Phung, Brett Adams and Svetha Venkatesh

Please type or print your name(s) as you wish it (them) to appear in print

**PART A – COPYRIGHT TRANSFER FORM**

The undersigned, desiring to publish the above article/paper in a publication of the Association for the Advancement of Artificial Intelligence, (AAAI), hereby transfer their copyrights in the above article/paper to the Association for the Advancement of Artificial Intelligence (AAAI), in order to deal with future requests for reprints, translations, anthologies, reproductions, excerpts, and other publications.

This grant will include, without limitation, the entire copyright in the article/paper in all countries of the world, including all renewals, extensions, and reversion thereof, whether such rights current exist or hereafter come into effect, and also the exclusive right to create electronic versions of the article/paper to the extent that such right is not subsumed under copyright.

The undersigned warrants that they are the sole author and owner of the copyright in the above article/paper, except for those portions shown to be in quotations; that the article/paper is original throughout; and that the undersigned right to make the grants set forth above is complete and unencumbered.

If anyone brings any claim or action alleging facts that, if true, constitute a breach of any of the foregoing warranties, the undersigned will hold harmless and indemnify AAAI, their grantees, their licensees, and their distributors against any liability, whether under judgment, decree, or compromise, and any legal fees and expenses arising out of that claim or actions, and the undersigned will cooperate fully in any defense AAAI may make to such claim or action. Moreover, the undersigned agrees to cooperate in any claim or other action seeking to protect or enforce any right the undersigned has granted to AAAI in the article/paper. If any such claim or action fails because of facts that constitute a breach of any of the foregoing warranties, the undersigned agrees to reimburse whomever brings such claim or action for expenses and attorneys' fees incurred therein.

**Returned Rights**

In return for these rights, AAAI hereby grants to the above author(s), and the employer(s) for whom the work was performed, royalty-free permission to:

1. Retain all proprietary rights other than copyright (such as patent rights).
2. Personal reuse of all or portions of the above article/paper in other works of their own authorship. This does not include granting third-party requests for reprinting, republishing, or other types of reuse. AAAI must handle all such third-party requests.
3. Reproduce, or have reproduced, the above article/paper for the author's personal use, or for company use provided that AAAI copyright and the source are indicated, and that the copies are not used in a way that implies AAAI endorsement of a product or service of an employer, and that the copies per se are not offered for sale. The foregoing right shall not permit the posting of the article/paper in electronic or digital form on any computer network, except by the author or the author's employer, and then only on the author's or the employer's own web page or ftp site. Such web page or ftp site, in addition to the aforementioned requirements of this Paragraph, shall not post other AAAI copyrighted materials not of the author's or the employer's creation (including tables of contents with links to other papers) without AAAI's written permission.
4. Make limited distribution of all or portions of the above article/paper prior to publication.
5. In the case of work performed under a U.S. Government contract or grant, AAAI recognized that the U.S. Government has royalty-free permission to reproduce all or portions of the above Work, and to authorize others to do so, for official U.S. Government purposes only, if the contract or grant so requires.

In the event the above article/paper is not accepted and published by AAAI, or is withdrawn by the author(s) before acceptance by AAAI, this agreement becomes null and void.

(1) Thin NGUYEN 09/03/2012

Author/Authorized Agent for Joint Author's Signature

Date

Kumarasen

Head, Dept. of Computing  
Curtin University

Employer for whom work was performed

Title (if not author)

*(For jointly authored Works, all joint authors should sign unless one of the authors has been duly authorized to act as agent for the others.)*