Mohamed Medhat Gaber
Ayse Goker
Nirmalie Wiratunga
Mihaela Cocea (Eds.)

# BCS SGAI SMA 2013

# The BCS SGAI Workshop on Social Media Analysis

**Workshop co-located with AI-2013 Thirty-third SGAI International Conference on Artificial Intelligence (BCS SGAI)**

**Cambridge, UK, December 10, 2013**

**Proceedings**

*Editors' addresses:*
Robert Gordon University
School of Computing Science and Digital Media
Riverside East
Garthdee Road
Aberdeen
AB10 7GJ, UK

{m.gaber1 | n.wiratunga | a.s.goker}@rgu.ac.uk

University of Portsmouth
School of Computing
Buckingham Building
Lion Terrace
Portsmouth
PO1 3HE, UK

mihaela.cocea@port.ac.uk

## Organizing Committee

Mohamed Medhat Gaber, Robert Gordon University, UK
Nirmalie Wiratunga, Robert Gordon University, UK
Ayse Goker, Robert Gordon University, UK
Mihaela Cocea, University of Portsmouth, UK

## Program Committee

Alexandra Balahur, European Commission Joint Research Centre, Italy
Carlos Martin Dancausa, Robert Gordon University, UK
Samhaa El-Beltagy, Nile University, Egypt
Rosta Farza, University of Pittsburgh, US
Joao Gomes, I2R - Institute for Infocomm Research, Singapore
Jelena Jovanovic, University of Belgrade, Serbia
Frederic Stahl, University of Reading, UK
Gulden Uchyigit, University of Brighton, UK
Berrin Yanikoglu, Sabanci University, Turkey

## Contents

# Introduction to the proceedings of the BCS SGAI Workshop on Social Media Analysis 2013

Mihaela Cocea[1], Mohamed Gaber[2], Nirmalie Wiratunga[2], and Ayse Göker[2]

[1] School of Computing, University of Portsmouth, UK
mihaela.cocea@port.ac.uk
[2] School of Computing & Digital Media,
Robert Gordon University, Aberdeen, UK
{m.gaber1, n.wiratunga, a.s.goker}@rgu.ac.uk

Social media websites such as Twitter, Facebook, Instagram, and YouTube continue to share user-generated content on a massive scale. Users attempting to find relevant information within such vast and dynamic volumes risk being overwhelmed. In response, efforts are being made to develop new tools and methods that help users make sense of and make use of social media sites. In this workshop we will bring together commercial and academic researchers to discuss these issues, and explore the challenges for social media mining.

The current expansion of social media leads to masses of affective data related to peoples emotions, sentiments and opinions. Knowledge discovery from such data is an emerging area of research in the past few years, with a potential number of applications of paramount importance to business organisations, individual users and governments. Data mining and machine learning techniques are used to discover knowledge from various types of affective data such as ratings, text or browsing data. Sentiment analysis techniques have grown tremendously over the last few years, addressing applications of paramount importance. Obama's presidential election campaign and Gap logo change are two of these examples. Business organisations, individuals and governments are keen on extracting what people think of a particular product, a newly introduced governmental policy, etc. Applications are growing rapidly and so are the techniques. However, the gap between techniques and applications is still an issue that needs to be addressed.

All submitted papers received two or three review reports from Program Committee members. Based on the recommendations of the reviewers, 4 full papers have been selected for publication and presentation at BCS SGAI 2013. The selected papers address a variety of research themes, ranging from the importance of domain-specific lexicons when analysing social media text to theme extraction and combining text with multimedia sources for opinion mining.

The paper "Domain-Based Lexicon Enhancement for Sentiment Analysis" by Aminu Muhammad, Nirmalie Wiratunga, Robert Lothian and Richard Glassey propose an approach for learning a domain-focused sentiment lexicon. They show that by combining a general lexicon with a domain-focused one better results are obtained for sentiment analysis on Twitter text.

The paper "Towards Passive Political Opinion Polling using Twitter" by Nicholas A. Thapen and Moustafa M. Ghanem investigated automatic analysis of political tweets.

They looked at sentiment analysis of tweets from UK members of parliament towards the main political parties, as well as voters' tweets analysis for inferring voting intentions. In addition, they conducted an automatic identification of key topics discussed by members of parliament and voters.

The paper "Mining Newsworthy Topics from Social Media" by Carlos Martin, David Corney, Ayse Göker, and Andrew MacFarlane explore the real-time detection of newsworthy stories by looking at "bursts" of phrases. This allows the identification of emerging topics. An interesting evaluation methods is used, where the ground truth is established from news stories that were published in the mainstream media, thus ensuring their newsworthiness.

The paper "Multimodal Sentiment Analysis of Social Media" by Diana Maynard, David Dupplaw, and Jonathon Hare combines mining of text and of multimedia sources such as images and videos for opinion mining. The analysis of multimedia content complements the opinion extraction from text by resolving ambiguity and providing contextual information.

The papers in these proceedings addressed various aspects of social media analysis, covering different techniques for analysis, as well as different applications. They illustrate the advancement of research in this field and the refinement of techniques to suit the application domains.

# Domain-Based Lexicon Enhancement for Sentiment Analysis

Aminu Muhammad, Nirmalie Wiratunga, Robert Lothian and Richard Glassey

IDEAS Research Institute, Robert Gordon University, Aberdeen UK
{a.b.muhammad1, n.wiratunga, r.m.lothian, r.j.glassey}@rgu.ac.uk

**Abstract.** General knowledge sentiment lexicons have the advantage of wider term coverage. However, such lexicons typically have inferior performance for sentiment classification compared to using domain focused lexicons or machine learning classifiers. Such poor performance can be attributed to the fact that some domain-specific sentiment-bearing terms may not be available from a general knowledge lexicon. Similarly, there is difference in usage of the same term between domain and general knowledge lexicons in some cases. In this paper, we propose a technique that uses distant-supervision to learn a domain focused sentiment lexicon. The technique further combines general knowledge lexicon with the domain focused lexicon for sentiment analysis. Implementation and evaluation of the technique on Twitter text show that sentiment analysis benefits from the combination of the two knowledge sources. The technique also performs better than state-of-the-art machine learning classifiers trained with distant-supervision dataset.

## 1 Introduction

Sentiment analysis concerns the study of opinions expressed in text. Typically, an opinion comprises of its polarity (positive or negative), the target (and aspects) to which the opinion was expressed and the time at which the opinion was expressed [14]. Sentiment analysis has a wide range of applications for businesses, organisations, governments and individuals. For instance, a business would want to know customer's opinion about its products/services and that of its competitors. Likewise, governments would want to know how their policies and decisions are received by the people. Similarly, individuals would want make use of other people's opinion (reviews or comments) to make decisions [14]. Also, applications of sentiment analysis have been established in the areas of politics [3], stock markets [1], economic systems [15] and security concerns [13] among others.

Typically, sentiment analysis is performed using machine learning or lexicon-based methods; or a combination of the two (hybrid). With machine learning, an algorithm is trained with sentiment labelled data and the learnt model is used to classify new documents. This method requires labelled data typically generated through labour-intensive human annotation. An alternative approach to generating labelled data called distant-supervision has been proposed [9, 23]. This approach relies on the appearance of certain emoticons that are deemed to signify positive (or negative) sentiment to tentatively labelled documents as positive (or negative). Although, training data generated through

distant-supervision have been shown to do well in sentiment classification [9], it is hard to integrate into a machine learning algorithm, knowledge which is not available from its training data. Similarly, it is hard to explain the actual evidence on which a machine learning algorithm based its decision.

The lexicon-based, on the other hand, involves the extraction and aggregation of terms' sentiment scores offered by a lexicon (i.e prior polarities) to make sentiment prediction. Sentiment lexicons are language resources that associate terms with sentiment polarity (positive, negative or neutral) usually by means of numerical score that indicate sentiment dimension and strength. Although sentiment lexicon is necessary for lexicon-based sentiment analysis, it is far from enough to achieve good results [14]. This is because the polarity with which a sentiment-bearing term appears in text (i.e. contextual polarity) could be different from its prior polarity. For example in the text "the movie sucks", although the term 'sucks' seems highly sentiment-bearing, this may not be reflected by a sentiment lexicon. Another problem with sentiment lexicons is that they do not contain domain-specific, sentiment-bearing terms. This is especially more common when a lexicon generated from standard formal text is applied in sentiment analysis of informal text.

In this paper, we introduce lexicon enhancement technique (LET) to address the the afore-mentioned problems of lexicon-based sentiment analysis. LET leverages the success of distant-supervision to mine sentiment knowledge from a target domain and further combines such knowledge with the one obtained from a generic lexicon. Evaluation of the technique on sentiment classification of Twitter text shows performance gain over using either of the knowledge sources in isolation. Similarly, the techniques performs better than three standard machine learning algorithms namely Support Vector Machine, Naive Bayes and Logistic Regression. The main contribution of this paper is two-fold. First, we introduce a new fully automated approach of generating social media focused sentiment lexicon. Second, we propose a strategy to effectively combine the developed lexicon with a general knowledge lexicon for sentiment classification.

The remainder of this paper is organised as follows. Section 2 describes related work. The proposed technique is presented in Section 3. Evaluation and discussions appear in Section 4, followed by conclusions and future work in Section 5.

## 2 Related Work

Typically, three methods have been employed for sentiment analysis namely machine learning, lexicon based and hybrid. For machine learning, supervised classifiers are trained with sentiment labelled data commonly generated through labour-intensive human annotation. The trained classifiers are then used to classify new documents for sentiment. Prior work using machine learning include the work of Pang et al [20], where three classifiers namely, Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVMs) were used for the task. Their results show that, like topic-based text classification, SVMs perform better than NB and ME. However, performance of all the three classifiers in sentiment classification is lower than in topic-based text classification. Document representation for machine learning is an unordered list of terms that appear in the documents (i.e. bag-of-words). A binary representation based on term

presence or absence attained up to 87.2% accuracy on a movie review dataset [18]. The addition of phrases that are used to express sentiment (i.e. appraisal groups) as additional features in the binary representation resulted in further improvement of 90.6% [32] while best result of 96.9% was achieved using term-frequency/inverse-document-frequency (tf/idf) weighting [19]. Further sentiment analysis research using machine learning attempt to improve classification accuracy with feature selection mechanisms. An approach for selecting bi-gram features was introduced in [16]. Similarly, feature space reduction based on subsumption hierarchy was introduced in [24]. The aforementioned works concentrate on sentiment analysis of reviews, therefore, they used star-rating supplied with reviews to label training and test data instead of hand-labelling. This is typical with reviews, however, with other forms of social media (e.g. discussion forums, blogs, tweets e.t.c.), star-rating is typically unavailable. Distant-supervision has been employed to generate training data for sentiment classification of tweets [9, 23]. Here, emoticons supplied by authors of the tweets were used as noisy sentiment labels. Evaluation results on NB, ME and SVMs trained with distant-supervision data but tested on hand-labelled data show the approach to be effective with ME attaining the highest accuracy of 83.0% on a combination of unigram and bigram features. The limitation of machine learning for sentiment analysis is that it is difficult to integrate into a classifier, general knowledge which may not be acquired from training data. Furthermore, learnt models often have poor adaptability between domains or different text genres because they often rely on domain specific features from their training data. Also, with the dynamic nature of social media, language evolves rapidly which may render a previous learning less useful.

The lexicon based method excludes the need for labelled training data but requires sentiment lexicon which several are readily available. Sentiment lexicons are dictionaries that associate terms with sentiment values. Such lexicons are either manually generated or semi-automatically generated from generic knowledge sources. With manually generated lexicons such as General Inquirer [25] and Opinion Lexicon [12], sentiment polarity values are assigned purely by humans and typically have limited coverage. As for the semi-automatically generated lexicons, two methods are common, *corpus-based* and *dictionary-based*. Both methods begin with a small set of seed terms. For example, a positive seed set such as 'good', 'nice' and 'excellent' and a negative seed set could contain terms such as 'bad', 'awful' and 'horrible'. The methods leverage on language resources and exploit relationships between terms to expand the sets. The two methods differ in that corpus-based uses collection of documents while the dictionary-based uses machine-readable dictionaries as the lexical resource. Corpus-based was used to generate sentiment lexicon [11]. Here, 657 and 679 adjectives were manually annotated as positive and negative seed sets respectively. Thereafter, the sets were expanded to conjoining adjectives in a document collection based on the connectives 'and' and 'but' where 'and' indicates similar and 'but' indicates contrasting polarities between the conjoining adjectives. Similarly, a sentiment lexicon for phrases generated using the web as a corpus was introduced in [29, 30]. Dictionary-based was used to generate sentiment lexicon in [2, 31]. Here, relationships between terms in WordNet [8] were explored to expand positive and negative seed sets. Both corpus-based and dictionary-based lex-

icons seem to rely on standard spelling and/or grammar which are often not preserved in social media [27].

Lexicon-based sentiment analysis begins with the creation of a sentiment lexicon or the adoption of an existing one, from which sentiment scores of terms are extracted and aggregated to predict sentiment of a given piece of text. Term-counting approach has been employed for the aggregation. Here, terms contained in the text to be classified are categorised as positive or negative and the text is classified as the class with highest number of terms [30]. This approach does not account for varying sentiment intensities between terms. An alternative approach is the aggregate-and-average strategy [26]. This classifies a piece of text as the class with highest average sentiment of terms. As lexicon-based sentiment analysis often rely on generic knowledge sources, it tends to perform poorly compared to machine learning.

Hybrid method, in which some elements from machine learning and lexicon based are combined, has been used in sentiment analysis. For instance, sentiment polarities of terms obtained from lexicon were used as additional features to train machine learning classifiers [5, 17]. Similarly, improvement was observed when multiple classifiers formed from different methods are used to classify a document [22]. Also, machine learning was employed to optimize sentiment scores in a lexicon [28]. Here, initial score for terms, assigned manually are increased or decreased based on observed classification accuracies.

## 3 Lexicon Enhancement Technique

Lexicon enhancement technique (LET) addresses the semantic gap between generic and domain knowledge sources. As illustrated in Fig. 1, the technique involves obtaining scores from a *generic lexicon*, automated domain *data labelling using distant-supervision*, *domain lexicon generation* and *aggregation strategy for classification*. Details of these components is presented in the following sub sections.

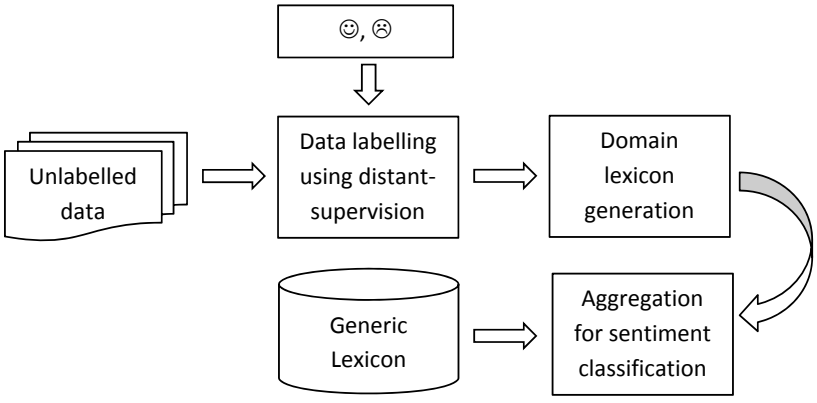



**Fig. 1.** Diagram showing the architectural components of the proposed technique (LET)

### 3.1 Generic Lexicon

We use SentiWordNet [7] as the source of generic sentiment scores for terms. SentiWordNet is a general knowledge lexicon generated from WordNet [8]. Each synset (i.e. a group of synonymous terms based on meaning) in WordNet is associated with three numerical scores indicating the degree of association of the synset with positive, negative and objective text. In generating the lexicon, seed (positive and negative) synsets were expanded by exploiting *synonymy* and *antonymy* relations in WordNet, whereby synonymy preserves while antonymy reverses the polarity with a given synset. As there is no direct synonym relation between synsets in WordNet, the relations: see_also, similar_to, pertains_to, derived_from and attribute were used to represent synonymy relation while direct antonym relation was used for the antonymy. Glosses (i.e. textual definitions) of the expanded sets of synsets along with that of another set assumed to be composed of objective synsets were used to train eight ternary classifiers. The classifiers are used to classify every synset and the proportion of classification for each class (positive, negative and objective) were deemed as initial scores for the synsets. The scores were optimised by a random walk using the PageRank [4] approach. This starts with manually selected synsets and then propagates sentiment polarity (positive or negative) to a target synset by assessing the synsets that connect to the target synset through the appearance of their terms in the gloss of the target synset. SentiWordNet can be seen to have a tree structure as shown in Fig. 2. The root node of the tree is a term whose child nodes are the four basic PoS tags in WordNet (i.e. noun, verb, adjective and adverb). Each PoS can have multiple word senses as child nodes. Sentiment scores illustrated by a point within the triangular space in the diagram are attached to word-senses. Subjectivity increases (while objectivity decreases) from lower to upper, and positivity increases (while negativity decreases) from right to the left part of the triangle.

We extract scores from SentiWordNet as follows. First, input text is broken into unit tokens (tokenization) and each token is assigned a lemma (i.e. corresponding dictionary entry) and PoS using Stanford CoreNLP library[1]. Although in SentiWordNet scores are associated with word-senses, disambiguation is usually not performed as it does not seem to yield better results than using either the average score across all senses of a term-PoS or the score attached to the most frequent sense of the term (e.g. in [21], [17], [6]). In this work, we use average positive (or negative) score at PoS level as the positive (or negative) for terms as shown in Equation 1.

$$gs(t)_{dim} = \frac{\sum_{i=1}^{|senses(t,PoS)|} ScoreSense_i(t,PoS)_{dim}}{|senses(t,PoS)|} \quad (1)$$

Where $gs(t)_{dim}$ is the score of term $t$ (given its part-of-speech, $PoS$) in the sentiment dimension of $dim$ ($dim$ is either positive or negative). $ScoreSense_i(t,PoS)_{dim}$ is the sentiment score of the term $t$ for the part-of-speech ($PoS$) at sense $i$. Finally, $|senses(t,PoS)|$ is number of word senses for the part-of-speech ($PoS$) of term $t$.

[1] nlp.stanford.edu/software/corenlp.shtml

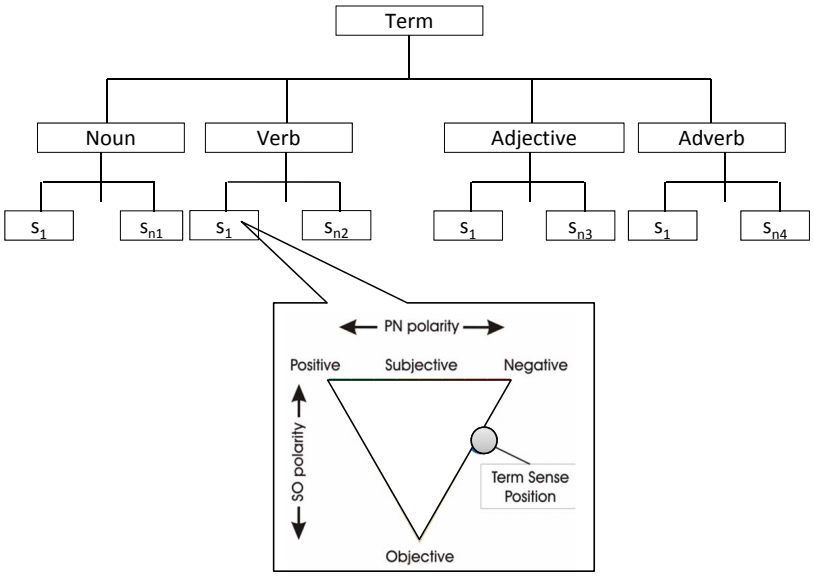



**Fig. 2.** Diagram showing the structure of SentiWordNet

### 3.2 Data Labelling Using Distant-Supervision

Distant-supervision offers an automated approach to assigning sentiment class labels to documents. It uses emoticons as noisy labels for documents. It is imperative to have as many data as possible at this stage as this affects the reliability of scores to be generated at the subsequent stage. Considering that our domain of focus is social media, we assume there will be many documents containing such emoticons and, therefore, large dataset can be formed using the approach. Specifically, in this work we use Twitter as a case study. We use a publicly available distant-supervision dataset for this stage [9][2]. This dataset contains 1,600,000 tweets balanced for positive and negative sentiment classes. We selected first 10,000 tweets from each class for this work. This is because the full dataset is too big to conveniently work with. For instance, building a single machine learning model on the full dataset took several days on a machine with 8GB RAM, 3.2GHZ Processor and 64bit Operating System. However, we aim to employ "big data" handling techniques to experiment with larger datasets in the future. The dataset is preprocessed to reduce feature space using the approach introduced in [9]. That is, all user names (i.e. words that starts with the @ symbol) are replaced with the token 'USERNAME'. Similarly all URLs (e.g. "http://tinyurl.com/cvvg9a") are replaced with the token 'URL'. Finally, words consisting of sequence of three or more repeated character (e.g. "haaaaapy") are normalised to contain only two of such repeated character in sequence.

[2]The dataset available from Sentiment140.com

### 3.3 Domain Lexicon Generation

Domain sentiment lexicon is generated at this stage. Each term from the distant-supervision dataset is associated with positive and negative scores. Positive (or negative) score for a term is determined as the proportion of the term's appearance in positive (or negative) documents given by equation 2. Separate scores for positive and negative classes are maintain in order to suit integration with the scores obtained from the generic lexicon (SentiWordNet). Table 1 shows example terms extracted from the dataset and their associated positive and negative scores.

$$ds(t)_{dim} = \frac{\sum_{dim} \text{tf}(t)}{\sum_{All documents} \text{tf}(t)} \tag{2}$$

Where $ds(t)_{dim}$ is the sentiment score of term $t$ for the polarity dimension $dim$ (positive or negative) and $\text{tf}(t)$ is document term frequency of $t$.

**Table 1.** Some terms from the domain lexicon

| Term | Sentiment Scores | |
|---|---|---|
| | Positive | Negative |
| ugh | 0.077 | 0.923 |
| sucks | 0.132 | 0.868 |
| hehe | 0.896 | 0.104 |
| damn | 0.241 | 0.759 |
| argh | 0.069 | 0.931 |
| thx | 1 | 0 |
| luv | 0.958 | 0.042 |
| xoxo | 0.792 | 0.208 |

### 3.4 Aggregation Strategy for Sentiment Classification

At this stage, scores from generic and domain lexicons for each term $t$ are combined for sentiment prediction. The scores are combined so as to complement each other according to the following strategy.

$$Score(t)_{dim} = \begin{cases} 0, & \text{if } gs(t)_{dim} = 0 \text{ and } ds(t)_{dim} = 0 \\ gs(t)_{dim}, & \text{if } ds(t)_{dim} = 0 \text{ and } gs(t) > 0 \\ ds(t)_{dim}, & \text{if } gs(t)_{dim} = 0 \text{ and } ds(t) > 0 \\ \alpha \times gs(t)_{dim} + (1-\alpha) \times ds(t)_{dim}, & \text{if } gs(t)_{dim} > 0 \text{ and } ds(t)_{dim} > 0 \end{cases}$$

The parameter, $\alpha$, controls a weighted average of generic and domain scores for $t$ when both scores are non-zero. In this work we set $\alpha$ to 0.5 thereby giving equal weights to both scores. However, we aim to investigate an optimal setting for the parameter in the future. Finally, sentiment class for a document is determined using aggregate-and-average method as outlined in Algorithm 1.

**Algorithm 1** Sentiment Classification

1: **INPUT:** *Document*
2: **OUTPUT:** *class* ▷ document sentiment class
3: Initialise: *posScore*, *negScore*
4: **for all** $t \in Document$ **do**
5: **if** $Score(t)_{pos} > 0$ **then**
6: $posScore \leftarrow posScore + Score(t)_{pos}$
7: $nPos \leftarrow nPos + 1$ ▷ increment number of positive terms
8: **end if**
9: **if** $Score(t)_{neg} > 0$ **then**
10: $negScore \leftarrow negScore + Score(t)_{neg}$
11: $nNeg \leftarrow nNeg + 1$ ▷ increment number of negative terms
12: **end if**
13: **end for**
14: **if** $posScore/nPos > negScore/nNeg$ **then return** *positive*
15: **else return** *negative*
16: **end if**

# 4 Evaluation

We conduct a comparative study to evaluate the proposed technique (LET). The aim of the study is three fold, first, to investigate whether or not combining the two knowledge sources (i.e. LET) is better than using each source alone. Second, to investigate performance of LET compared to that of machine learning algorithms trained with distant-supervision data since that is the state-of-the-art use of distant-supervision for sentiment analysis. Lastly, to study the behaviour of LET on varying dataset sizes. We use hand-labelled Twitter dataset, introduced in [9][3] for the evaluation. The dataset consists of 182 positive and 177 negative tweets.

## 4.1 LET Against Individual Knowledge Sources

Here, the following settings are compared:

1. LET: The proposed technique (see Algorithm 1)
2. Generic: A setting that only utilises scores obtained from the generic lexicon (SentiWorNet). In Algorithm 1, $Score(t)_{pos}$ (line 5) and $Score(t)_{neg}$ (line 9) are replaced with $gs(t)_{pos}$ and $gs(t)_{neg}$ respectively.
3. Domain: A setting that only utilises scores obtained from the domain lexicon. In Algorithm 1, $Score(t)_{pos}$ (line 5) and $Score(t)_{neg}$ (line 9) are replaced with $ds(t)_{neg}$ and $ds(t)_{neg}$ respectively.

Table 2 shows result of the comparison. The LET approach performs better than Generic and Domain. This is not suprising since LET utilises generic knowledge which could have been omitted by Domain and also, domain knowledge which could have

[3]The dataset is available from Sentiment140.com

**Table 2.** Performance accuracy of individual knowledge sources and LET

| Generic | Domain | LET |
|---|---|---|
| 60.33 | 71.26 | 75.27 |

been omitted by Generic. Also the result shows that the generated domain lexicon (Domain) is more effective than the general knowledge lexicon (Generic) for sentiment analysis.

### 4.2 LET Against Machine Learning and Varying Dataset Sizes

Three machine learning classifiers namely Naïve Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR) are trained with the distant-supervision dataset and then evaluated with the human-labelled test dataset. These classifiers are selected because they are the most commonly used for sentiment classification and typically perform better than other classifiers. We use presence and absence (i.e. binary) feature representation for documents and Weka [10] implementation for the classifiers. Furthermore, we use subsets of the distant-supervision dataset (16000, 12000, 8000 and 4000; also balanced for positive and negative classes) in order to test the effect of varying distant-supervision dataset sizes for LET (in domain lexicon generation, see Section 3.3) and the machine learning classifiers.

**Table 3.** LET compared to machine learning methods on varying data sizes

| Classifier / Dataset size | NB | SVM | LR | LET |
|---|---|---|---|---|
| 4,000 | 60.17 | 61.00 | 66.02 | 68.70 |
| 8,000 | 54.04 | 59.61 | 69.64 | 73.10 |
| 12,000 | 54.04 | 62.12 | 71.03 | 73.80 |
| 16,000 | 54.04 | 62.95 | 71.87 | 75.27 |
| 20,000 | 54.60 | 62.40 | 73.26 | 75.27 |

Table 3 shows result of the experiment. LET performs better than any of the machine learning classifiers. This can be attributed to the fact that LET utilises generic knowledge which the machine learning classifiers could not have acquired from the training dataset, especially, as the distant-supervision dataset may contain incorrect labels. As for the behaviour of LET and the classifiers on varying dataset sizes, they all tend to improve in performance with increased dataset size as depicted by Fig. 3, with the exception of SVM for which the performance drops. Interestingly however, the difference between the algorithms appeared to be maintained over the different dataset sizes. This shows that the domain lexicon generated in LET becomes more accurate with increased dataset size in a similar manner that a machine learning classifier becomes more accurate with increased training data.

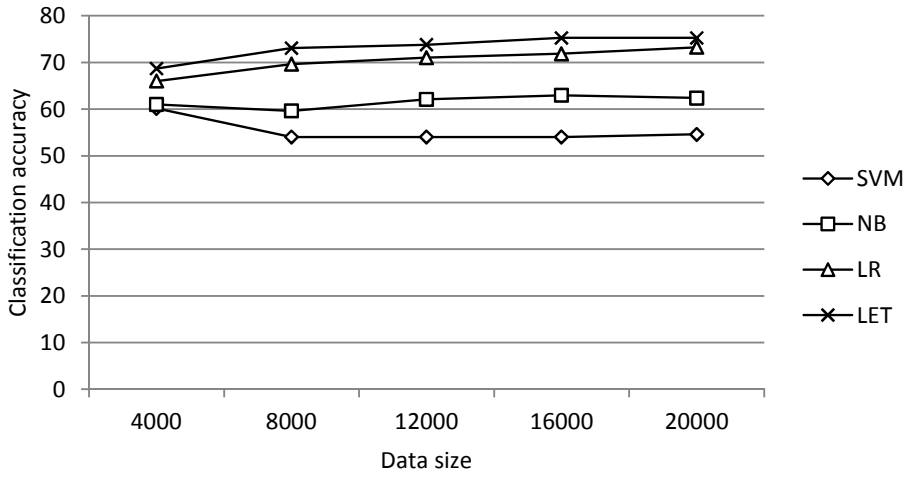



**Fig. 3.** LET compared to machine learning methods on varying data sizes

## 5 Conclusions and Future Work

In this paper, we presented a novel technique for enhancing generic sentiment lexicon with domain knowledge for sentiment classification. The major contributions of the paper are that we introduced a new approach of generating domain-focused lexicon which is devoid of human involvement. Also, we introduced a novel strategy to combine generic and domain lexicons for sentiment classification. Experimental evaluation shows that the technique is effective and better than state-of-the-art machine learning sentiment classification trained the same dataset from which our technique extracts domain knowledge (i.e. distant-supervision data).

As part of future work, we plan to conduct an extensive evaluation of the technique on other social media platforms (e.g. discussion forums) and also, to extend the technique for subjective/objective classification. Similarly, we intend perform experiment to find an optimal setting for $\alpha$ and improve the aggregation strategy presented.

## References

[1] Arnold, I., Vrugt, E.: Fundamental uncertainty and stock market volatility. Applied Financial Economics 18(17), 1425–1440 (2008)

[2] Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Annual Conference on Language Resouces and Evaluation (2010)

[3] Baron, D.: Competing for the public through the news media. Journal of Economics and Management Strategy 14(2), 339–376 (2005)

[4] Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Seventh International World-Wide Web Conference (WWW 1998) (1998)

[5] Dang, Y., Zhang, Y., Chen, H.: A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. IEEE Intelligent Systems 25, 46–53 (2010)

[6] Denecke, K.: Using sentiwordnet for multilingual sentiment analysis. In: ICDE Workshop (2008)
[7] Esuli, A., Baccianella, S., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10) (2010)
[8] Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
[9] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Processing pp. 1–6 (2009)
[10] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009)
[11] Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL. pp. 174–181. New Brunswick, NJ (1997)
[12] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177 (2004)
[13] Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., Hamfors, O.: Usefulness of sentiment analysis. In: 34th European Conference on Information Retrieval (2012)
[14] Liu, B.: Sentiment Analysis and Subjectivity, chap. Handbook of Natural Language Processing, pp. 627–666. Chapman and Francis, second edn. (2010)
[15] Ludvigson, S.: Consumer confidence and consumer spending. The Journal of Economic Perspectives 18(2), 29–50 (2004)
[16] Mukras, R., Wiratunga, N., Lothian, R.: Selecting bi-tags for sentiment analysis of text. In: Proceedings of the Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence (2007)
[17] Ohana, B., Tierney, B.: Sentiment classification of reviews using sentiwordnet. In: 9th IT&T Conference, Dublin, Ireland (2009)
[18] Pang, B., Lee, L.: Polarity dataset v2.0, 2004. online (2004), http://www.cs.cornell.edu/People/pabo/movie-review-data/.
[19] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1), 1–135 (2008)
[20] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods on Natural Language Processing (2002)
[21] Pera, M., Qumsiyeh, R., Ng, Y.K.: An unsupervised sentiment classifier on summarized or full reviews. In: Proceedings of the 11th International Conference on Web Information Systems Engineering. pp. 142–156 (2010)
[22] Prabowo, R., Thelwall, M.: sentiment analysis: A combined approach. Journal of Informetrics 3(2), 143–157 (2009)
[23] Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop. pp. 43–48. ACLstudent '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005)
[24] Riloff, E., Patwardhan, S., Wiebe, J.: Feature subsumption for opinion analysis. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-06) (2006)
[25] Stone, P.J., Dexter, D.C., Marshall, S.S., Daniel, O.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA (1966)
[26] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics 37, 267–307 (2011)

[27] Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 63(1), 163–173 (2012)
[28] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2444–2558 (2010)
[29] Turney, P., et al.: Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: Proceedings of the twelfth european conference on machine learning (ecml-2001) (2001)
[30] Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 417–424 (2002)
[31] Valitutti, R.: Wordnet-affect: an affective extension of wordnet. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation. pp. 1083–1086 (2004)
[32] Whitelaw, C., Garg, N., Argamon., S.: Using appraisal groups for sentiment analysis. In: 14th ACM International Conference on Information and Knowledge Management (CIKM 2005). pp. 625–631 (2005)

# Towards Passive Political Opinion Polling using Twitter

Nicholas A. Thapen[1], Moustafa M. Ghanem[2]

[1]Department of Computing, Imperial College London
180 Queens Gate London, SW7 2AZ, United Kingdom {nicholas.thapen12@imperial.ac.uk}

[2]School of Science and Technology, Middlesex University, London
The Burroughs, London NW4 4BT, United Kingdom {m.ghanem@mdx.ac.uk}

**Abstract.** Social media platforms, such as Twitter, provide a forum for political communication where politicians broadcast messages and where the general public engages in the discussion of pertinent political issues. The open nature of Twitter, together with its large volume of traffic, makes it a useful resource for new forms of 'passive' opinion polling, i.e. automatically monitoring and detecting which key issues the general public is concerned about and inferring their voting intentions. In this paper, we present a number of case studies for the automatic analysis of UK political tweets. We investigate the automated sentiment analysis of tweets from UK Members of Parliament (MPs) towards the main political parties. We then investigate using the volume and sentiment of the tweets from other users as a proxy for their voting intention and compare the results against existing poll data. Finally we conduct automatic identification of the key topics discussed by both the MPs and users on Twitter and compare them with the main political issues identified in traditional opinion polls. We describe our data collection methods, analysis tools and evaluation framework and discuss our results and the factors affecting their accuracy.

## 1 Introduction

Twitter is a social networking service set up in 2006 allowing users to publish a stream of short messages called 'tweets', consisting of 140 characters or less. The author of a tweet may add the "#" symbol as prefix to arbitrary words in its content which become known as 'hashtags'. These hashtags can be regarded as keywords for identifying related messages. Optionally, users may also auto-tag their geo-location to a tweet. The social network is structured so that users can 'follow' each other, thus adding the followed user's tweets to the follower's newsfeed. Unlike other social networks, following a user in Twitter is usually automatic and does not require authorization. It is more like subscribing to an RSS feed or news service than establishing a friendship. Other built-in features are the ability to rebroadcast, or 'retweet', another tweet and the ability to reply to specific users as well as mention them in tweets. These features give Twitter aspects of both a social network and a news medium. In 2012, the service had over 140 million active users, with over 340 million tweets sent daily [19].




Paper submitted to BCS SGAI Workshop on Social Media Analysis 2013, Cambridge, UK.

### 1.1 Towards Passive Political Opinion Polling from Twitter

Many politicians have embraced Twitter as a means to reach out directly to the public, bypassing traditional media sources. For example, Barack Obama launched his 2012 re-election campaign on Twitter and his victory tweet became the most retweeted of all time. Moreover, with a large number of users, the service has become an important arena for the dissimulation, discussion and creation of political news and opinion.

The fact that Twitter is seemingly open and accessible[1] makes it much more appealing to use for the purposes of research than other social networks such as Facebook, which have more emphasis on privacy. A tweet is generally intended to be read by a wide audience as part of the public record. While individual tweets contain very little information their brevity means that they are typically focused on a single issue. Moreover the aggregate of thousands or millions of tweets can potentially be fruitfully analyzed to discover different views around the same issue. The fact that political representatives use Twitter along with their constituents allows their language and interaction to be studied in order to discern what they say on Twitter about key issues and to discern whether they are out of step with the population at large. The analysis of what the public is discussing on Twitter could also be used for identifying their key concerns, and potentially also inferring their voting intentions.

Although a number of papers have been published on using Twitter data for predicting election results (See for example [1-3,5,6,8,10-12,14,15,17]), there is little work linking Twitter data with tracking opinion polls analyzing which key issues may be influencing the opinions of the people or the polls themselves. One paper [12] investigated Twitter economic sentiment and US presidential tracking polls. It found significant correlation between their data set and economic sentiment but little correlation on the political polls. Their study used simple keyword matching based on the 'Obama' and 'McCain' keywords. Whether more sophisticated approaches can find a meaningful correlation between the debate on Twitter and the opinion polls is clearly an open research question.

### 1.2 Motivation and Paper Overview

We start with two key questions:

1. Can we use Twitter to infer the proportion of the general public intending to vote for specific political candidates or parties?
2. Can we use Twitter data to infer the distribution of issues that the general public is concerned about?

We do not attempt to answer either question in this paper. However, a first reasonable step towards answering them is to compare the results of automated analysis of Twitter data with available poll data. This would allow us to gain a better understanding of what is needed to develop appropriate methodologies for conducting 'passive' opinion polling. Starting from two data sets that we collected on UK tweets in

[1] We discuss briefly in Section 3 some of the practical challenges for collecting historical Twitter data

2012/13, we experimented with automatic sentiment analysis tools and used both tweet volume and sentiment-weighted tweet volume as proxies for voting intention. We also developed and investigated the use of automatic topic identification techniques from the tweets and compared the outputs to key issues identified as important in opinion polls. Although our experiments are based on simple approaches, they provide many illuminating results that help in appreciating the questions better.

In Section 2, we review related work on election result prediction from Twitter data and discuss some of its key challenges. In Section 3, we describe the data sets used in our experiments. In Sections 4 and 5 we describe both the sentiment analysis tools and topic detection algorithms used and present and discuss the results for each case. Finally, in Section 6, we present our conclusions and discussion.

## 2 Related Work

Various researchers have investigated the use of Twitter for election result prediction. However, the successes of the approaches used have shown great variation. In an analysis of the 2009 German federal election [17] the authors were able to predict the vote shares in the election with a Mean Average Error of 1.65%, compared to an average error of 1.14% for six standard opinion polls. A study of the UK 2010 General Elections [18] reported a final average error of 1.75%. However, a study of the 2011 Singapore Elections in 2011 [15] found a greater error rate of 5.23%., whereas a study of the U.S Senate elections in 2010 [10] found far larger errors of around 17%.

Most work used the volumes of tweets mentioning particular candidates or parties as the measure of their popularity. However, some studies also investigated different methods for incorporating automated sentiment analysis of tweets' contents towards the contenders. The German study reported evidence that tweets about parties lying in similar places on the political spectrum contained similar emotional content. The US study reported that the final prediction error was reduced from 17% to 7.6% when sentiment analysis was applied. The German study simply used the Linguistic Inquiry and Word Count (LIWC) software tool to compute word and phrase statistics whereas an investigation of the 2011 Irish General Election [2] trained a classifier on a corpus of manually annotated positive and negative political tweets, then used tweet volume weighted by sentiment to report a final error of 5.85%. Given the prevalence of sarcasm and sophisticated humor in political discussions the reported results are encouraging.

One criticism [5] is that most studies are retrospective, performing backward-looking analysis rather than true prediction, and that their data selection methods arbitrarily influence their conclusions. One paper [8] showed that if the German study had included the German Pirate Party, much favored by Internet activists, they would have been predicted a landslide victory. We note that all studies also vary drastically in terms of data collection methods, sample sizes and how the analysis is conducted. There is usually no attempt at elucidating how the underlying assumptions of the studies may relate to standard opinion polling techniques, such as demographic weighting. It is rare that attempts are made at analyzing the context of the tweets or what is being

discussed. In many cases, there is also little attempt to remove the influence of spammers or 'Twitter bombs' [10] - deliberate campaigns by political activists sending out thousands of similar tweets in a form of campaign advertising.

Moreover, most studies in this sphere are typically single shot experiments focused on the technological aspects. There is little or no methodological framework describing how they should be repeated and no standard benchmark against which they could be measured or through which their effectiveness could be analyzed time after time.

## 3 UK Political Tweets and Poll Data

**UK Polling Report and Ipsos MORI Issues Index**

We retrieved the list of voting intention polls kept by the UK Polling Report website [22]. This list provides all voting intention polls in the UK since June 2012. The polls are from all polling companies, and are thus based on various methodologies, such as phone-polling, internet panels and face to face interviews.

To retrieve a list of the issues that the public is concerned about we used Ipsos MORI [7], which has published a monthly Issues Index for the UK since 1974. This is based on a face-face survey asking around 1,000 British people the following question: "What do you see as the main/other important issues facing Britain today?" Respondents normally give around three categories as being important issues and Ipsos MORI then condense the answers into categories such as 'Health' and 'Economy'.

Taking a list of topics from this source enables us to compare if political discussions on Twitter centre around the same topics or not. For our experiments we retrieved the Ipsos MORI Issues Index for the months of July 2012 - July 2013. To keep our analysis tractable, we consolidated the most frequent issues appearing in the poll data in the past year into 14 main categories, as well as an 'Other' category intended to catch all other issues. The categories chosen are:

*Crime, Economy, Education, Environment, EU, Foreign Affairs, Government Services, Health, Housing, Immigration, Pensions, Politics, Poverty, Unemployment*

In our classification, 'Economy' includes Inflation, Tax, Value of pound as well as the Ipsos-MORI Economy category, 'Foreign Affairs' includes all defense related matters, 'Environment' includes Rural Affairs, 'Pensions' includes Adult Social Care, and 'Politics' refers to Devolution and Constitutional Reform.

**UK MP Twitter Data and Political Discussion Data**

In order to identify UK political issues discussed on Twitter automatically we needed to collect a training data set that could be used in learning a lexicon of UK political terms. We focused on UK Members of Parliament (MPs) with the assumption that their tweets would mainly be focused on topical political issues. Moreover, the political orientation of these delegates is known and their tweets can be used to provide sanity checks on automated sentiment analysis methods as described later.

A list of the Twitter accounts of 423 UK MPs, classified by party affiliation, was retrieved from news website Tweetminster [18]. We retrieved 689,637 tweets from the publically available timelines of the MPs on 10th June 2013 using Twitter's REST API [20]. We note that timeline data returned by the API is capped at a maximum of 3,200 tweets for a single user's timeline. Although Twitter holds an archive of all Tweets posted since the service began, these Tweets are not held on the user's timeline and are indexed only by their unique id. Query access to such data is only possible through a number of commercial data providers [20].

In order to collect sample tweets relevant to UK political discussions, we considered collecting data using geo-location queries for the UK and then filtering by political keywords. This would have allowed us to look at geographic topic distributions and voting intentions. However, very few people enable geo-tagging due to privacy concerns. We thus decided to consider Twitter users who had mentioned recently the leaders of the three main political parties in their tweets. Our assumption is that most such users would be UK-based and more interested in UK political discussion than others. We thus retrieved the list of Twitter users who had recently mentioned the leaders of the three main political parties. We removed from those users known news sources to avoid news oriented tweets. We also ensured that none of them were in the existing UK MPs list. We then took a random sample of 600 of the remaining users. Similar to the MP data set, we retrieved the tweets from each user's timeline. This resulted in 1,431,348 tweets; retrieved in August 2013.

## 4 Sentiment Analysis and Voting Intentions

### 4.1 Sentiments of the MPs towards different parties

We experimented with different types of automated sentiment analysis techniques. In this paper we report on the results achieved using SentiStrength [16], a freely available sentiment analysis software tool which assigns sentiment scores based on look-ups to keywords in a sentiment polarity lexicon. We applied the tool to both the MP dataset and the political discussion datasets.

First, using the MP dataset, we extracted the tweets where political parties are mentioned. MPs discussing other parties can generally be assumed to be attempting to disparage them in some way, while when discussing their own parties they will usually use a positive spin. We used a keyword set containing the names, nicknames and shortenings of the names of the three main parties and then excluded from the dataset any tweets that mentioned more than one party. This resulted in a data set with 48,140 tweets (Labour: 23,070; Conservative: 18,034; Liberal Democrats: 7,036). The smaller number of Liberal Democrat tweets reflects the small size of the parliamentary party and activist base compared to the two main parties. The tweets were then split into groups depending on the party of the MP who tweeted them.

To identify how accurate the sentiment detection was, 30 tweets were selected at random from each of the nine groups and manually annotated as 'Positive 'or 'Nega-

tive' based on the language used and the sense meant. The results are summarized in Table 1.

**Table 1.** Sentiment Accuracy on Test Data Set

| Class | Precision | Recall | F1 Measure |
|---|---|---|---|
| Negative | 0.583 | 0.483 | 0.528 |
| Positive | 0.651 | 0.737 | 0.691 |
| Overall | 0.617 | 0.610 | 0.614 |

Clearly, the low precision and recall values raise alarms about the accuracy of the results for individual tweets, but overall indicate that the sentiment score could still be usable for overall sentiment detection. To verify, we then applied SentiStrength to the nine data sets (Figure 1). Here the figure shows SentiStrength's average positive and negative classification over each group, on a scale ranging from 1 (least positive/negative) to 5 (most positive/negative). The results back the general hypothesis. The simple methods work over aggregate data and show that MPs from each party tweet more negatively about other parties. Yet, the high level of negative sentiment of MPs concerning their own parties would be a surprise to most of the MPs themselves and their parties, as is the fact that any Labour tweets about Tories and vice-versa were positive.

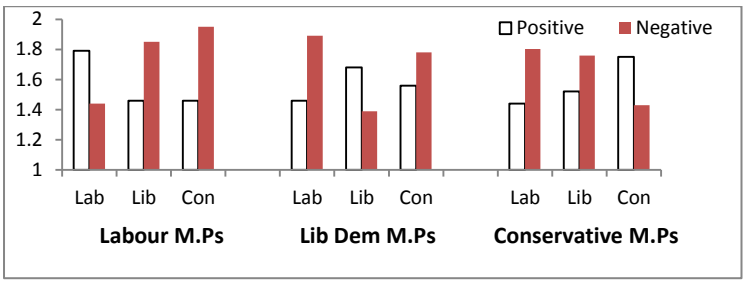


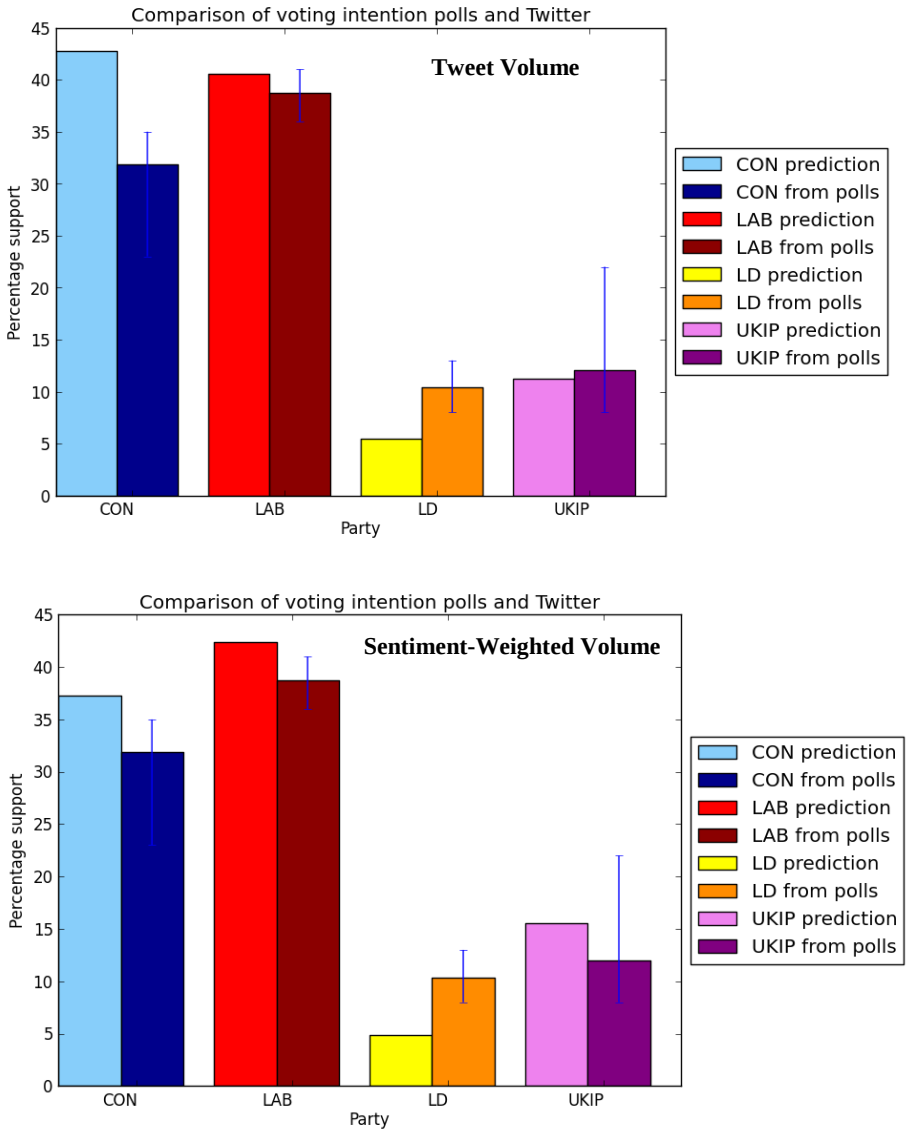

**Fig. 1.** Sentiment of MPs tweets towards each party based on SentiStrength

Despite both the simplicity of the method and its classification errors, the results show interesting insights. For example, they reveal that the parties tweet more negatively about their main opponents than the third party. Also, despite the coalition Lib Dems and Conservatives are not positive about each other, just less negative.

### 4.2 Investigating Voting Intentions

We proceeded to investigate whether the tweets in the political discussion dataset can be used as a proxy for inferring voting intentions. Our first experiment was simply to examine the proportion of tweets mentioning each party in the month of July 2013 (for which we had the most data in our data set) and to compare this to the average of the published opinion polls for the party. Firstly we obtained the numbers of tweets mentioning each party in that month. We excluded tweets which mentioned more than

one party, as these would not be useful for the later sentiment analysis steps. We then took the proportion of tweets mentioning each party and compared it to the average predicted share of the vote from the opinion polls.




**Figs. 2a and 2b** Voting Intentions vs. Tweet Volume and Sentiment-Weighted Volume

The results are shown in Figure 2.a and the error bars give the range of values from the different opinion polls. The comparison between the Twitter prediction and the polls has a Mean Absolute Error of 4.6%, which as a first attempt was a surprisingly

high correspondence. As shown in the figure, there is a close match for Labour and UKIP, but the Conservatives are given too much prominence and the Lib Dems too little. The ordering of Labour and the Conservatives is also incorrect.

Since many of the tweets mentioning the Conservatives are presumably negative, as they are the main party of government, we now moved on to weighting the results by sentiment to see if this could improve the fit of the data. In order to do so we adopted the sentiment-weighting methodology described in [12]. Adding in the sentiment weighting improved the error slightly, to 4.52%. More importantly all four parties are now in the correct rank order (Figure 2.b). The weighting was achieved by first running the sentiment analysis against all tweets to split them into positive and negative classes, then calculating sentiment weighted volume as follows:

$$weightedcount = count(party\ mentions) \times \frac{count(positive\ party\ mentions)}{count(negative\ party\ mentions)}$$

The fraction to compare against the polls is then: $\frac{weightedcount\ for\ party}{sum\ of\ weighted\ counts\ for\ all\ parties}$

**Investigating Temporal Effects**

We then looked at the same figures over the period July 2012 to July 2013. This revealed that the sentiment-weighted tweet data were extremely volatile, especially when looking at the earlier months in the chart. Before April 2013 they fail to match well with the voting intention figures at all. This analysis would seem to suggest that the close match between our data and the opinion polls for a single month could be a coincidence. However, the discrepancy could be accounted for by noting that we had much more data for recent months than for older ones due to the timeline retrieval limitations on Twitter. As mentioned earlier, the Twitter API allows retrieving only the most recent 3,200 tweets for each user. For example in our data set we have 9,979 tweets which mention party names in July 2013, but only 2,830 for April 2013.

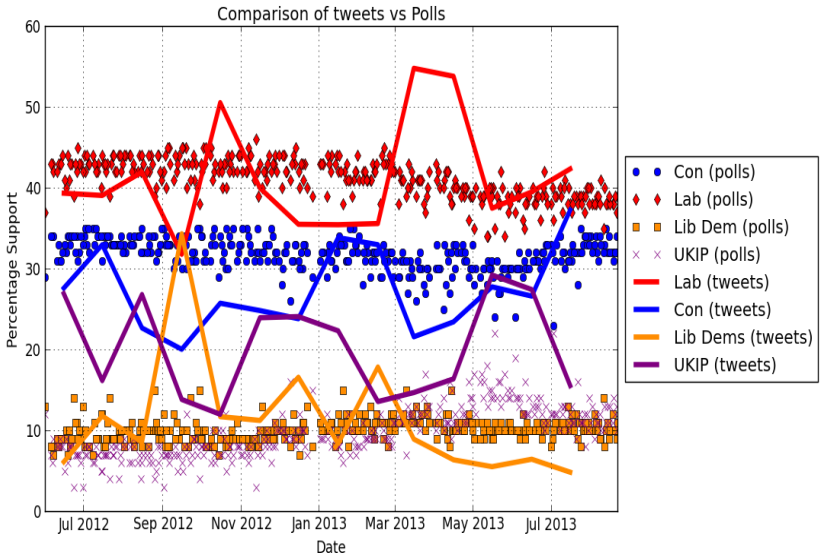



**Fig. 3.** Comparing Sentiment-Weighted Tweet Volume and Voting Intensions for 12 months

# 5 Topic Detection and Poll Issues

## 5.1 Topic Detection Algorithms

**Iterative Keyword-based Topic Detection**

We used an iterative snowballing method similar to [9], allowing the vocabulary to be built up gradually from the data, to develop the keyword-based classifier. We started with a manually constructed initial keyword list for each topic by consulting Wikipedia and our pre-existing political knowledge. These were used to classify the tweets in the MP dataset into the 14 categories. The keywords for each category were then augmented by additional keywords. This was done by first counting the frequency of 1, 2 and 3-grams in the tweets in each category and then applying a Fisher's Exact Test [21] to find n-grams which occurred most frequently within each category compared to outside the category. Human judgment was then applied to the candidate list to decide on which new keywords should be added. The process was iterated a number of times.

When using the initial keyword list on the MP dataset, the method was able to assign 109,644 tweets into the 14 categories but left 579,993 tweets, or 84% of the dataset, uncategorized. After the 5$^{th}$ iteration it could be seen that diminishing returns were setting in, since the final iteration categorized very few extra tweets. Snowballing allowed categorizing 94,647 extra tweets, leaving 70.4% of the dataset uncategorized. The large number of uncategorized tweets is expected since not all tweets will be discussing the 14 categories. We also experimented with using a Porter stemmer.

This reduced the number of uncategorized tweets to 63.5% of the total. However, it also slightly reduced classification accuracy, and was not taken forward.

To evaluate the keyword classifier, we used a data set of 300 tweets (20 tweets at random from those matched for each topic at the 5th iteration from the political discussion set) and manually annotated them. This resulted in Precision of 87.2%, Recall of 86.4% and F1-measure of 86,8%. These results ignore the large number of uncategorized tweets but indicate that the method is quite precise for our training purposes.

**Bayesian Topic Classification**

We then developed a Naïve Bayesian multi-label topic classifier that treats each tweet as a bag of words (similar to [4]). However, annotating a sufficient number of tweets for each topic to train the classifier would have been extremely time-consuming. We thus used output labels from the keyword-based classifier as training labels, giving thousands of tweets for each category. Moreover, since the training labels are noisy, the prior probabilities used by the classifier for the class labels were calculated from a probability distribution obtained by sampling 300 tweets and manually annotating them.

We trained a classifier for each topic separately, in order to allow for multi-label classification. If a topic classifier decides that the probability of the topic given the words in the tweet is greater than the probability of not belonging to the topic given the words, then the tweet is included in the topic label. If none of the classifiers assign a label to the tweet then the class with the greatest probability is selected as the single label.

An important caveat is that the distribution from the sample was fed into the Bayesian classifiers as prior knowledge. This means that classifiers are somewhat over-fitted. We thus prepared another randomly selected test data set of 300 tweets that was manually annotated. We then evaluated both classifiers on a randomly sampled manually annotated sample data set of 300 tweets. The results are summarized in Table 2. The results indicate that the Bayesian classifier is more accurate than the keyword-based one. Moreover, its accuracy is reasonable. Also, as can been seen, training the Bayesian classifier on stemmed data slightly improved both precision and recall. Nonetheless, the difference can be assumed not to be statistically significant.

**Table 2.** Classifier Evaluation on random data set

| Classifier | Precision | Recall | F1 Measure |
|---|---|---|---|
| Keyword matching ($5^{th}$ iteration) | 0.287 | 0.279 | 0.283 |
| Bayesian on non-stemmed data | 0.753 | 0.793 | 0.773 |
| Bayesian on stemmed data | 0.764 | 0.798 | 0.781 |

To gain further confidence in our Bayesian classifier, we used it to label all tweets in the MP dataset and compared the distribution of topics detected to that of yet a new annotated sample set. These results gave close agreement, with a Mean Absolute Error of 0.0138 for the standard and 0.0129 for the stemmed classifier, with most topics falling within the margin of error of the sample. To perform sanity checks, we also compared the distribution of topics based on MPs from different political parties. The

results (not shown because of space limitations) were consistent with political expectation. Labour MPs were concerned more with unemployment, poverty and housing than the Conservatives or Liberal Democrats. They also tweet more about the economy, reflecting their strong opposition to the governing coalition's austerity measures. The Conservatives' focus on the E.U compared to other parties is also evident, along with a slight extra interest in foreign affairs. Education receives a lot of emphasis from the Conservatives, perhaps due to their heavy focus on free schools and education reform in this parliament. It is somewhat surprising that Labour focus more on crime than the Conservatives and less on the environment.

**Comparing Topic Distribution**

We then compared the topic distribution between the MP data set and political one as shown in Figure 4. A greater proportion of the tweets here were identified as 'chatter', 70% rather than the 52% found amongst the MPs. Given that MPs are public figures, it was to be expected that a greater proportion of their tweeting would concern political topics. The higher proportion of tweets in the 'Other' category accounts for part of this, as does the fact that the keywords are explicitly political. The language used by MPs habitually uses many more of the political terms than the users. But, more, importantly, it was the MP data set that was used in training our methods.

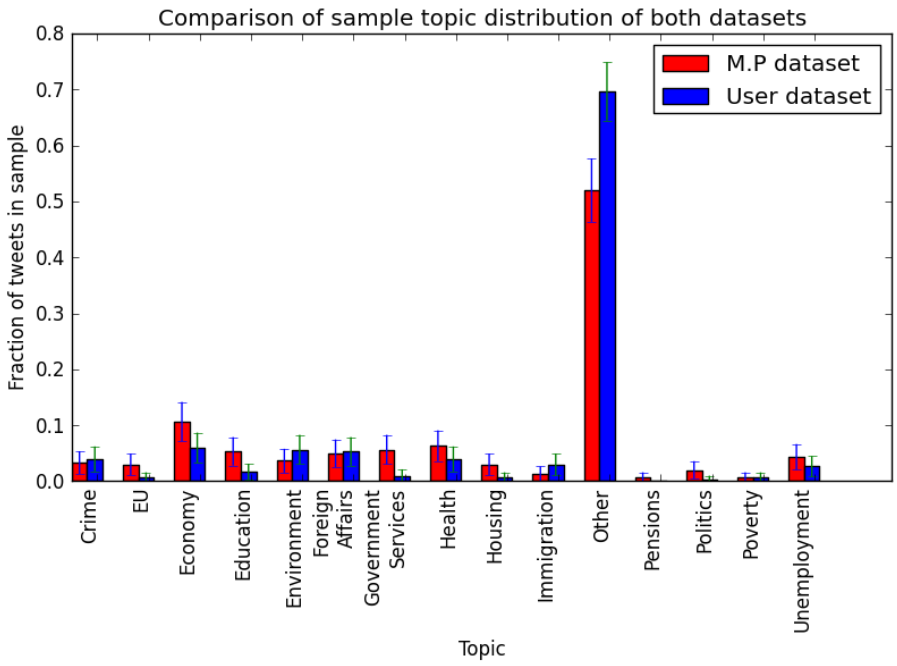



**Fig. 4.** Comparison between MP and Political Discussion Sets

### 5.2 Comparison with Polling Data

Finally, we proceeded to comparing the Ipsos MORI Issues Index to the topics extracted by the Bayesian classifier on both datasets. Again, we initially focused on the month of June, for which we had most data. The results are summarized in Figure 5. The Mean Absolute Error was 0.098 for the MPs and 0.11 for the users. One could interpret this slight difference in favour of the MPs being slightly more in touch with the concerns of the ordinary population than general Twitter users, since their topic distribution is found to be slightly closer to the polls. However, one must also notice that it was the MP data set used in the training of the classifier.

We do note the discrepancies between the polls and both the MPs and normal users in several categories; specifically, 'Economy', 'Immigration', 'Pensions' and 'Unemployment'. They all seem to be much less discussed on Twitter than in the poll. Analyzing the reasons of the mismatches is beyond the scope of this paper, but we cannot avoid making some comments. For example, one could also argue that normal users may not discuss the immigration issue too much over Twitter if they would be seen as racist by others. They could, however, be more likely to admit worries about it in private to a pollster than to broadcast them over Twitter.

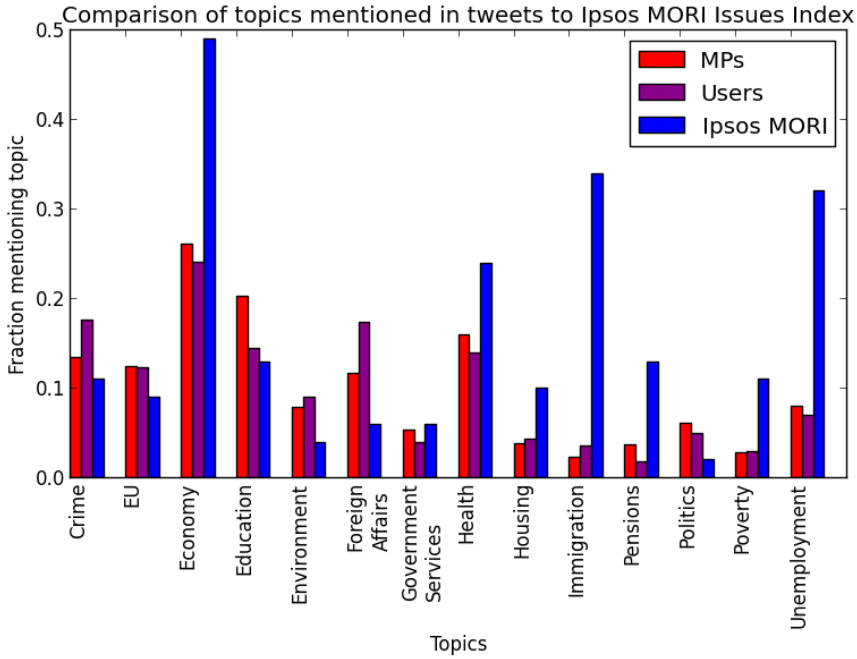



**Fig. 5.** Comparison of Twitter topics and Ipsos-MORI Issues for the month of June 2013

The demographics of Twitter users could potentially have had a big impact on the results. One could argue that Twitter users could be a much younger age group, and possibly one that is more affluent, than the broader spectrum taking part in the Ipsos MORI poll. However, there are no demographics in our Twitter data so we therefore examined the breakdown for the poll data itself for the 18-34 year old ABC1 group. This a social grade used by polling companies and market research organizations in the UK representing the segment of the population in high managerial down to administrative and professional positions, and is approximately 57% of the UK population [7]. We do not present the results here, but summarize that our experiments could not find any closer match in issues of this segment to the topics discussed on Twitter.

### Investigating Temporal Effects

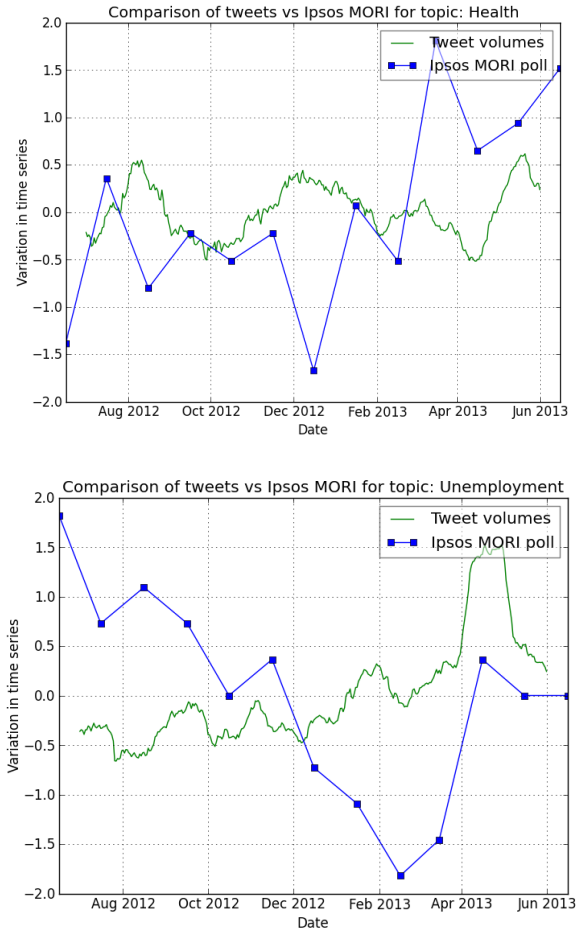



**Fig. 6.a and 6.b** 12-months comparison 'Health' and 'Unemployment' tweets and Ipsos-MORI category

Finally, we examined how the results varied for individual topics for the period June 2012 – June 2013. We provide only two examples here due to space limitations. Figure 6 shows time analysis for the topics 'Health' and 'Immigration' in the political discussion set vs. the poll data. The visual analysis of the graph does show some correlation between the trends of the respective time series, even if they do not match point-wise. The results are encouraging, but clearly indicate that more work is needed in developing the appropriate comparison methodology.

## 6 Summary and Discussion

In this paper, we presented our case studies conducted towards automated 'passive' political opinion polling from UK tweets. Namely, we looked at comparing volume and sentiment-weighted volume of tweets mentioning political parties with voting intentions reported from traditional polls. We also looked at detecting key political topics discussed over Twitter and comparing them to issues identified in polls. We described the techniques used and presented our results. Overall, the techniques yielded a close match and indeed showed that sentiment-weighted volume showed better matches for recent data. However, they showed volatility for the complete year. When comparing topics discussed in Twitter vs. those identified in polls, the task proved to more difficult, even if still promising.

Throughout the paper we identified all of our assumptions and described how our data collection methods could have influenced the results. The sample of tweets used in our work is not necessarily representative and our results are clearly not statistically rigorous. Our aim was not to conduct political analysis over the Twitter data but to investigate some of the key challenges that need to be addressed when doing so. Further development of the methodology for collecting the data and of the appropriate analysis methods is needed. Also more work is needed to understand how socio-political and demographics issues affect the results.

In the paper, we also showed how we used the known affiliation the MP to provide various sanity checks and also for training our lexicons. Clearly, the known affiliation could also be used in more interesting ways. For example we are currently investigating its use in developing Bayesian analysis techniques that take the context of a tweet into consideration when assigning a sentiment score. Moreover, we are investigating with various other Twitter data selection and sampling methods to avoid issues relating to political campaigning and also to increase the users under consideration.

If real-time, 'passive' opinion polling could be perfected it would be possible to cheaply canvas public opinion on a much broader range of issues than traditional polling. It could also potentially avoid 'framing' effects where respondents are influenced by the question asked. If such methods could be augmented by a theoretical underpinning, more sophisticated sentiment analysis and techniques such as demographic weighting then they could become a valuable tool in the political forecaster's arsenal and also for marketing analysts. However, more investigation is still required into developing and evaluating new appropriate methodologies for collecting the re-

quired data, developing more sophisticated software tools and algorithms for its analysis and developing standardized methods and benchmarks to evaluate the accuracy of results.

## References


1. Ampofo, L. et al. Trust, Confidence and Credibility: Citizen responses on Twitter to opinion polls during the 2010 UK General Election. Information, Communication & Society 14.6 (2011): 850-871.
2. Bermingham, A. & Smeaton, A., 2011. On using twitter to monitor political sentiment and predict election results. Workshop on Sentiment Analysis where AI meets Psychology.
3. Boutet, A. et al., 2012. What's in your tweets? I know who you supported in the UK 2010 general election. The International AAAI Conference on Weblogs and Social Media.
4. Frank, E. & Bouckaert, R.R., 2006. Naive Bayes for Text Classification with Unbalanced Classes. In PKDD'06 Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases.
5. Gayo-Avello, D., 2012. I wanted to predict elections with twitter and all I got was this lousy paper. A balanced survey on election prediction using twitter data. arXiv preprint arXiv:1204.6441 [Accessed June 3, 2013]. Available at: http://arxiv.org/abs/1204.6441
6. Hong, S and Nadler, D. Does the early bird move the polls?: the use of the social media tool Twitter by US politicians and its impact on public opinion." Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times. ACM, 2011.
7. Ipsos MORI [Accessed 4 September 2013] http://www.ipsos-mori.com/.
8. Jungherr, A., et al., 2012. Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment." Social Science Computer Review, 30, pp.229–234.
9. McCallum, A and Nigam, K. Text classification by bootstrapping with keywords, EM and shrinkage. ACL 1999 Workshop for the Unsupervised Learning in Natural Language Processing, pp. 52-58, 1999.
10. Metaxas, P.T. & Mustafaraj, E., 2012. Social Media and the Elections. *Science.* [Accessed May 31, 2013]. Available at: http://www.sciencemag.org/content/338/6106/472.full.pdf.
11. Mungiu-Pippidi, A and Munteanu, I Moldova's Twitter Revolution. Journal of Democracy 20.3 (2009): 136-142.
12. O'Connor, B. et al., 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. International AAAI Conference on Weblogs and Social Media.
13. Pew Research, 2012. Assessing the Representativeness of Public Opinion Surveys. [Accessed June 7, 2013]. Available at: http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.
14. Sang, E et al. Predicting the 2011 Dutch senate election results with twitter. Proceedings of the Workshop on Semantic Analysis in Social Media. Association for Computational Linguistics, 2012.
15. Skoric, M. et al., 2012. Tweets and Votes: A Study of the 2011 Singapore General Election. 2012 45th Hawaii International Conference on System Sciences, pp.2583–2591.
16. Thelwall, M. et al., 2010. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), pp.2544–2558.

17. Tumasjan, A. et al., 2010. Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment. Word Journal of The International Linguistic Association, 280(39), pp.178–185.
18. Tweetminster, 2010. Tweetminster predicts. [Accessed September 4, 2013]. Available at: http://www.scribd.com/doc/31208748/Tweetminster-Predicts-Findings.
19. Twitter, 2012. Twitter turns six. @Twitter. [Accessed September 4, 2013]. Available at: https://blog.twitter.com/2012/twitter-turns-six.
20. Twitter, 2013. Twitter API [Accessed December 1, 2013], Available at https://dev.twitter.com/
21. Weisstein, E.W., 2013. Fisher's Exact Test. *MathWorld--A Wolfram Web Resource*. Available at: http://mathworld.wolfram.com/FishersExactTest.html [Accessed August 28, 2013].
22. Wells, A., 2013. Voting Intention since 2010. [Accessed September 4, 2013]. Available at: http://ukpollingreport.co.uk/voting-intention-2.

# Mining Newsworthy Topics from Social Media

Carlos Martin[1], David Corney[1], Ayse Göker[1], and Andrew MacFarlane[2]

[1] IDEAS Research Institute, School of Computing & Digital Media,
Robert Gordon University, Aberdeen AB10 7QB
{c.j.martin-dancausa, d.p.a.corney, a.s.goker}@rgu.ac.uk
[2] School of Informatics, City University London, London EC1V 0HB
a.macfarlane-1@city.ac.uk

**Abstract.** Newsworthy stories are increasingly being shared through social networking platforms such as Twitter and Reddit, and journalists now use them to rapidly discover stories and eye-witness accounts. We present a technique that detects "bursts" of phrases on Twitter that is designed for a real-time topic-detection system. We describe a time-dependent variant of the classic *tf-idf* approach and group together bursty phrases that often appear in the same messages in order to identify emerging topics.
We demonstrate our methods by analysing tweets corresponding to events drawn from the worlds of politics and sport. We created a user-centred "ground truth" to evaluate our methods, based on mainstream media accounts of the events. This helps ensure our methods remain practical. We compare several clustering and topic ranking methods to discover the characteristics of news-related collections, and show that different strategies are needed to detect emerging topics within them. We show that our methods successfully detect a range of different topics for each event and can retrieve messages (for example, tweets) that represent each topic for the user.

**Keywords:** topic detection, Twitter, temporal analysis

## 1 Introduction

The growth of social networking sites, such as Twitter, Facebook and Reddit, is well documented. Every day, a huge variety of information on different topics is shared by many people. Given the real-time, global nature of these sites, they are used by many people as a primary source of news content [1]. Increasingly, such sites are also used by journalists, partly to find and track breaking news but also to find user-generated content such as photos and videos, to enhance their stories. These often come from eye-witnesses who would be otherwise difficult to find, especially given the volume of content being shared.

Our overall goal is to produce a practical tool to help journalists and news readers to find newsworthy topics from message streams without being overwhelmed. Note that it is not our intention to re-create Twitter's own "trending topics" functionality. That is usually dominated by very high-level topics and

memes, defined by just one or two words or a name and with no emphasis on 'news'.

Our system works by identifying phrases that show a sudden increase in frequency (a "burst") and then finding co-occurring groups to identify topics. Such bursts are typically responses to real-world events. In this way, the news consumer can avoid being overwhelmed by redundant messages, even if the initial stream is formed of diverse messages. The emphasis is on the temporal nature of message streams as we bring to the surface groups of messages that contain suddenly-popular phrases. An early version of this approach was recently described [2, 3], where it compared favourably to several alternatives and benchmarks. Here we expand and update that work, examining the effect of different clustering and topic ranking approaches used to form coherent topics from bursty phrases.

## 2 Related Work

Newman [4] discusses the central use of social media by news professionals, such as hosting live blogs of ongoing events. He also describes the growth of collaborative, networked journalism, where news professionals draw together a wide range of images, videos and text from social networks and provide a curation service. Broadcasters and newspapers can also use social media to increase brand loyalty across a fragmented media marketplace.

Petrovic et al. [5] focus on the task of first-story detection (FSD), which they also call "new event detection". They use a locality sensitive hashing technique on 160 million Twitter posts, hashing incoming tweet vectors into buckets in order to find the nearest neighbour and hence detect new events and track them. This work is extended in Petrovic et al. [6] using paraphrases for first story detection on 50 million tweets. Their FSD evaluation used newswire sources rather than Tweets, based on the existing TDT5 datasets. The Twitter-based evaluation was limited to calculating the average precision of their system, by getting two human annotators to label the output as being about an event or not. This contrasts with our goal here, which is to measure the topic-level recall, i.e. to count how many newsworthy stories the system retrieved.

Benhardus [7] uses standard collection statistics such as *tf-idf*, unigrams and bigrams to detect trending topics. Two data collections are used, one from the Twitter API and the second being the Edinburgh Twitter corpus containing 97 million tweets, which was used as a baseline with some natural language processing used (e.g. detecting prepositions or conjunctions). The research focused on general trending topics (typically finding personalities and for new hashtags) rather than focusing the needs of journalistic users and news readers.

Shamma et al. [8] focus on "peaky topics" (topics that show highly localized, momentary interest) by using unigrams only. The focus of the method is to obtain peak terms for a given time slot when compared to the whole corpus rather than over a given time-frame. The use of the whole corpus favours batch-mode processing and is less suitable for real-time and user-centred analysis.

Phuvipadawat and Murata [9] analysed 154,000 tweets that contained the hashtag '#breakingnews". They determine popularity of messages by counting retweets and detecting popular terms such as nouns and verbs. This work is taken further with a simple *tf-idf* scheme that is used to identify similarity [10]; named entities are then identified using the Stanford Named Entity Recogniser in order to identify communities and similar message groups. Sayyadi et al. [11] also model the community to discover and detect events on the live Labs SocialStream platform, extracting keywords, noun phrases and named entities. Ozdikis et al. [12] also detect events using hashtags by clustering them and finding semantic similarities between hashtags, the latter being more of a lexicographic method. Ratkiewitcz et al. [13] focus specifically on the detection of a single type of topic, namely political abuse. Evidence used include the use of hashtags and mentions. Alvanaki [14] propose a system based on popular seed tags (tag pairs) which are then tracked, with any shifts detected and monitored. These articles do use natural language processing methods and most consider temporal factors, but do not use $n$-grams.

Becker et al. [15] also consider temporal issues by focusing on the online detection of real world events, distinguishing them from non-events (e.g. conversations between posters). Clustering and classification algorithms are used to achieve this. Methods such as $n$-grams and NLP are not considered.

## 3 Methods

### 3.1 BNgrams

Term frequency-inverse document frequency, or *tf-idf*, has been used for indexing documents since it was first introduced [16]. We are not interested in indexing documents however, but in finding novel trends, so we want to find terms that appear in one *time period* more than others. We treat temporal windows as documents and use them to detect words and phrases that are both new and significant. We therefore define newsworthiness as the combination of novelty and significance. We can maximise *significance* by filtering tweets either by keywords (as in this work) or by following a carefully chosen list of users, and maximise *novelty* by finding bursts of suddenly high-frequency words and phrases.

We select terms with a high "temporal document frequency-inverse document frequency", or *df-idf*$_t$, by comparing the most recent $x$ messages with the previous $x$ messages and count how many contain the term. We regard the most recent $x$ messages as one "slot". After standard tokenization and stop-word removal, we index all the terms from these messages. For each term, we calculate the document frequency for a set of messages using $df_{ti}$, defined as the number of messages in a set $i$ that contain the term $t$.

$$df{-}idf_{ti} = (df_{ti} + 1) \cdot \frac{1}{\log\left(df_{t(i-1)} + 1\right) + 1}. \tag{1}$$

This produces a list of terms which can be ranked by their *df-idf*$_t$ scores. Note that we add one to term counts to avoid problems with dividing by zero or

taking the log of zero. To maintain some word order information, we define terms as $n$-grams, i.e. sequences of $n$ words. Based on experiments reported elsewhere [3], we use 1-, 2- and 3-grams in this work. High frequency $n$-grams are likely to represent semantically coherent phrases. Having found bursts of potentially newsworthy $n$-grams, we then group together $n$-grams that tend to appear in the same tweets. Each of these clusters defines a topic as a list of $n$-grams, which we also illustrate with a representative tweet. We call this process of finding bursty $n$-grams"BNgrams."

### 3.2 Topic Clustering

An isolated word or phrase is often not very informative, but a group of them can define the essence of a story. Therefore, we group the most representative phrases into clusters, each representing a single topic. A group of messages that discuss the same topic will tend to contain at least some of the same phrases. We can then find the message that contains the most phrases that define a topic, and use that message as a human-readable label for the topic. We now discuss three clustering algorithms that we compare here.

**Hierarchical clustering**. Here, we initially assign every $n$-gram to its own singleton cluster, then follow a standard "group average" hierarchical clustering algorithm [17] to iteratively find and merge the closest pair of clusters. We repeat this until no two clusters share more than half their terms, at which point we assume that each cluster represents a distinct topic. We define the similarity between two terms as the fraction of messages in the same time slot that contain both of them, so it is highly likely that the term clusters whose similarities are high represent the same topic. Further details about this algorithm and its parameters can be found in our previous published work [2].

**Apriori algorithm**. The Apriori algorithm [18] finds all the associations between the most representative $n$-grams based on the number of tweets in which they co-occur. Each association is a candidate topic at the end of the process. One of the advantages of this approach is that one $n$-gram can belong to different associations (partial membership), avoiding one problem with hierarchical clustering. No number of associations has to be specified in advance. We also obtain maximal associations after clustering to avoid large overlaps in the final set of topic clusters.

**Gaussian mixture models**. GMMs assign probabilities (or strengths) of membership of each $n$-gram to each cluster, allowing partial membership of multiple clusters. This approach does require the number of clusters to be specified in advance, although this can be automated (e.g. by using Bayesian information criteria [19]). Here, we use the Expectation - Maximisation algorithm to optimise a Gaussian mixture model [20]. We fix the number of clusters at 20, although initial experiments showed that using more or fewer produced very similar results. Seeking more clusters in the data than there are newsworthy topics means that some clusters will contain irrelevant tweets and outliers, which can later be assigned a low rank and effectively ignored, leaving us with a few highly-ranked clusters that are typically newsworthy.

### 3.3 Topic Ranking

To maximise usability we need to avoid overwhelming the user with a very large number of topics. We therefore want to rank the results by relevance. Here, we compare two topic ranking techniques.

**Maximum *n*-gram $df - idf_t$.** One method is to rank topics according to the maximum $df - idf_t$ value of their constituent $n$-grams. The motivation of this approach is assume that the most popular $n$-gram from each topic represents the core of the topic.

**Weighted topic-length.** As an alternative we propose weighting the topic-length (i.e. the number of terms found in the topic) by the number of tweets in the topic to produce a score for each topic. Thus the most detailed and popular topics are assigned higher rankings. We define this score thus:

$$s_t = \alpha \cdot \frac{L_t}{L_{max}} + (1 - \alpha) \cdot \frac{N_t}{N_s} \tag{2}$$

where $s_t$ is the score of topic $t$, $L_t$ is the length of the topic, $L_{max}$ is the maximum number of terms in any current topic, $N_t$ is the number of tweets in topic $t$ and $N_s$ is the number of tweets in the slot. Finally, $\alpha$ is a weighting term. Setting $\alpha$ to 1 rewards topics with more terms; setting $\alpha$ to 0 rewards topics with more tweets. We used $\alpha = 0.7$ in our experiments, giving slightly more weight to those stories containing more details, although the exact value is not critical.

## 4 Experiments

Here, we show the results of our experiments with several variations of the BNgram approach. We focus on two questions. First, what is best slot size to balance topic recall and refresh rate? A very small slot size might lead to missed stories as too few tweets would be analysed; conversely, a very large slot size means that topics would only be discovered some time after they have happened. This low 'refresh rate' would reduce the timeliness of the results. Second, what the best combination of clustering and topic ranking techniques? Earlier, we introduced three clustering methods and two topic ranking methods; we need to determine which methods are most useful.

We have previously shown that our methods perform well [2]. The BNgram approach was compared to a popular baseline system in topic detection and tracking – Latent Dirichlet Allocation (LDA) [21] – and to several other competitive topic detection techniques, getting the best overall topic recall. In addition, we have shown the benefits of using $n$-grams compared with single words for this sort of analysis [3]. Below, we present and discuss the results from our current experiments, starting with our approach to evaluation.

### 4.1 Evaluation Methods

When evaluating any IR system, it is crucial to define a realistic test problem. We used three Twitter data sets focused on popular real-world events and compare the topics that our algorithm finds with an externally-defined ground truth.

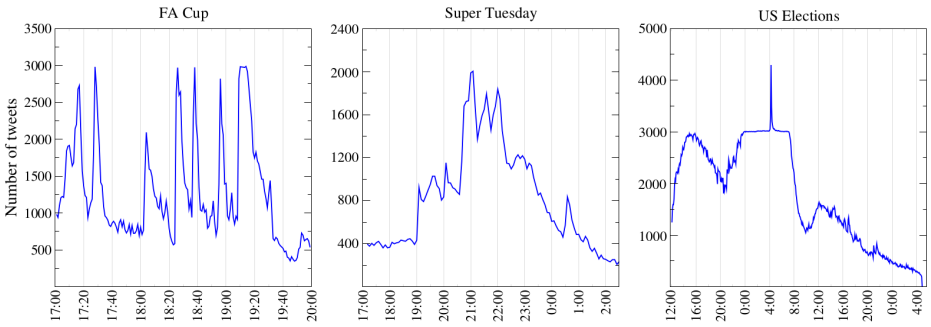



Fig. 1: Twitter activity during events (tweets per minute). For the FA Cup, the peaks correspond to start and end of the match and the goals. For the two political collections, the peaks correspond to the main result announcements.

To establish this ground truth, we relied on mainstream media (MSM) reports of the three events. This use of MSM sources helps to ensure that our ground truth topics are newsworthy (by definition) and that the evaluation is goal-focussed (i.e. will help journalists write such stories). We filtered Twitter using relevant keywords and hashtags to collect tweets around three events : the "Super Tuesday" primaries, part of the presidential nomination race of the US Republican Party; the 2012 FA Cup final, the climax to the English football season; and the 2012 US presidential election, an event of global significance. In each case, we reviewed the published MSM accounts of the events and chose a set of stories that were significant, time-specific, and represented on Twitter. For example, we ignored general reviews of the state of US politics (not time-specific), and quotes from members of the public (not significant events).

For each target topic, we identified around 5-7 keywords that defined the story to measure recall and precision, as discussed below. Some examples are shown in the first two columns of Table 4. We also defined several "forbidden" keywords. A topic was only considered as successfully recalled if all of the "mandatory" terms were retrieved and *none* of the "forbidden" terms. The aim was to avoid producing topics such as "victory Romney Paul Santorum Gingrich Alaska Georgia" that convey no information about who won or where; or "Gingrich wins", which is too limited to define the story because it doesn't name the state where the victory occurred.

Figure 1 shows the frequency of tweets collected over time, with further details in ref. [2]. We have made all the data freely available[3]. The three data sets differ in the rates of tweets, determined by the popularity of the topic and the choice of filter keywords. The mean tweets per minute (tpm) were: Super Tuesday, 832 tpm; FA Cup, 1293 tpm; and US elections, 2209 tpm. For a slot size of 1500 tweets these correspond to a "topic refresh rate" of 108s, 70s and

[3] `http://www.socialsensor.eu/results/datasets/72-twitter-tdt-dataset`

41s respectively. This means that a user interface displaying these topics could be updated every 1–2 minutes to show the current top-10 (or top-$m$) stories.

We ran the topic detection algorithm on each data set. This produced a ranked list of topics, each defined by a set of terms (i.e. $n$-grams). For our evaluation, we focus on the recall of the top $m$ topics ($1 \leq m \leq 10$) at the time each ground-truth story emerges. For example, if a particular story was being discussed in the mainstream media from 10:00-10:15, then we consider the topic to be recalled if the system ranked it in the top $m$ at any time during that period.

The automatically detected topics were compared to the ground truth (comprising 22 topics for Super Tuesday; 13 topics for FA Cup final; and 64 topics for US elections) using three metrics: **Topic recall**: Percentage of ground truth topics that were successfully detected by a method. A topic was considered successfully detected if the automatically produced set of words contained all mandatory keywords for it (and none of the forbidden terms, if defined). **Keyword precision**: Percentage of correctly detected keywords out of the total number of keywords for all topics detected by the algorithm in the slot. **Keyword recall**: Percentage of correctly detected keywords over the total number of ground truth keywords (excluding forbidden keywords) in the slot. One key difference between "topic recall" and "keyword recall" is that the former is a user-centred evaluation metric, as it considers the power of the system at retrieving and displaying to the user stories that are meaningful and coherent, as opposed to retrieving only some keywords that are potentially meaningless in isolation.

Note that we do not attempt to measure topic precision as this would need an estimate of the total number of newsworthy topics at any given time, in order to verify which (and how many) of the topics returned by our system were in fact newsworthy. This would require an exhaustive manual analysis of MSM sources to identify every possible topic (or some arbitrary subset), which is infeasible. One option is to compare detected events to some other source, such as Wikipedia, to verify the significance of the event [22], but Wikipedia does not necessarily correspond to particular journalists' requirements regarding newsworthiness and does not claim to be complete.

### 4.2 Results

Table 1 shows the effect on topic recall of varying the slot size, with the same total number of topics in the evaluation for each slot size. The mean is weighted by the number of topics in the ground truth for each set, giving greater importance to larger test sets. Overall, using very few tweets produces slightly worse results than with larger slot sizes (e.g. 1500 tweets), presumably as there is too little information in such a small collection. Slightly better results for the Super Tuesday set occur with fewer tweets; this could be due to the slower tweet rate in this set. Note that previous experiments [3] showed that including 3-grams improves recall compared to just using 1- and 2-grams, but adding 4-grams provides no extra benefit, so here we use 1-, 2- and 3-gram phrases throughout.

| *Slot size (tweets)* | 500 | 1000 | 1500 | 2000 | 2500 |
|---|---|---|---|---|---|
| Super Tuesday | **0.773** | 0.727 | 0.682 | 0.545 | 0.682 |
| FA Cup | 0.846 | 0.846 | **0.923** | **0.923** | **0.923** |
| US Elections | 0.750 | 0.781 | **0.844** | 0.734 | 0.766 |
| Weighted mean | 0.77 | 0.78 | **0.82** | 0.72 | 0.77 |

Table 1: Topic recall for different slot sizes (with hierarchical clustering).

Lastly, we compared the results of combining different clustering techniques with different topic ranking techniques (see Fig. 2). We conclude that the hierarchical clustering performs well despite the weakness discussed above (i.e. each $n$-gram is assigned to only one cluster), especially in FA Cup dataset. Also, the use of weighted topic-length ranking technique improves topic recall with hierarchical clustering in the political data sets.

The Apriori algorithm performs quite well in combination with the weighted topic length ranking technique (note that this ranking technique was specially created for the "partial" membership clustering techniques). We see that the Apriori algorithm in combination with the maximum $n$-gram $df - idf_t$ ranking technique produces slightly worse results, as this ranking technique does not produce diverse topics for the first results (from top 1 to top 10, in our case) as we mentioned earlier.

Turning to the EM Gaussian mixture model results, we see that this method works very well on the FA Cup final and US elections data sets. Despite being a "partial" membership clustering technique, the use of weighted topic length ranking technique does not make any representative difference, even its performance is worse in Super Tuesday dataset. Further work is needed to test this.

Table 2 summarises the results of the three clustering methods and the two ranking methods across all three data sets. The weighted-mean scores show that for the three clustering methods, ranking by the length of the topic is more effective than ranking by each topic's highest $df - idf_t$ score. We can see that for the FA Cup set, the Hierarchical and GMM clustering methods are the best ones in combination with the maximum $n$-gram $df - idf_t$ ranking technique. For Super Tuesday and US Elections data sets, "partial" membership clustering techniques (Apriori and GMM, respectively) perform the best in combination with weighted topic length ranking technique, as expected.

Finally, Table 3 shows more detailed results, including keyword precision and recall, for the best combinations of clustering and topic ranking methods of the three datasets when the top five results are considered per slot. In addition, Table 4 shows some examples of ground truth and BNgram detected topics and tweets within the corresponding detected topics for all datasets.

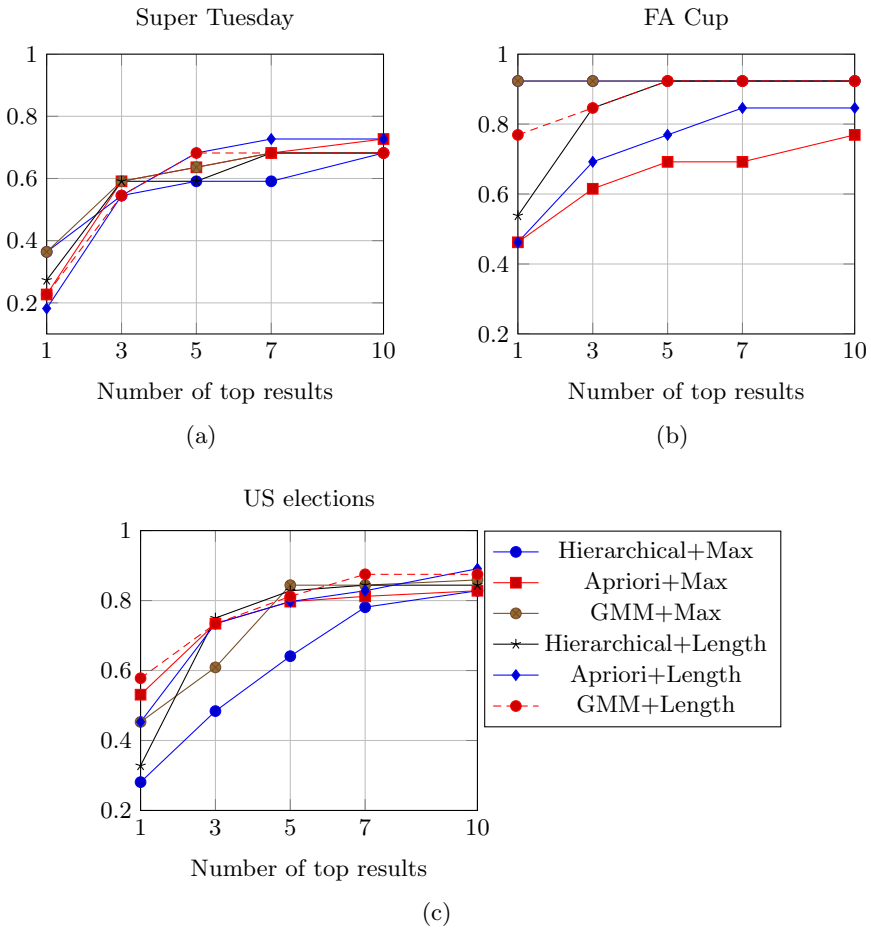



Fig. 2: Topic recall for different clustering techniques in the Super Tuesday, FA Cup and US elections sets (slot size = 1500 tweets).

## 5 Conclusions

If we compare the results between the three collections, one difference is particularly striking: the topic recall is far higher for football (over 90%) than for politics (around 60-80%; Table 2). This is likely to reflect the different nature of conversations about the events. Topics within a live sports event tend to be transient: fans care (or at least tweet) little about what happened five minutes ago; what matters is what is happening "now". This is especially true during key events, such as goals. In politics, conversations and comments tend to spread over hours (or even days) rather than minutes. This means that sports-related topics tend to occur over a much narrower window, with less overlapping chatter. In politics, several different topics are likely to be discussed at the same time,

| *Ranking* | Max. $n$-gram $df - idf_t$ | | | Weighted topic-length | | |
|---|---|---|---|---|---|---|
| *Clustering* | Hierar. | Apriori | GMM | Hierar. | Apriori | GMM |
| FA Cup | **0.923** | 0.677 | **0.923** | 0.861 | 0.754 | 0.892 |
| Super Tuesday | 0.573 | 0.605 | 0.6 | 0.591 | **0.614** | 0.586 |
| US Elections | 0.627 | 0.761 | 0.744 | 0.761 | 0.772 | **0.797** |
| Weighted Mean | 0.654 | 0.715 | 0.735 | 0.736 | 0.734 | **0.763** |

Table 2: Normalised area under the curve for the three datasets combining the different clustering and topic ranking techniques (1500 tweets per slot).

| **Method** | *T-REC@5* | *K-PREC@5* | *K-REC@5* |
|---|---|---|---|
| **Super Tuesday** | | | |
| *Apriori+Length* | 0.682 | 0.431 | 0.68 |
| *GMM+Length* | 0.682 | 0.327 | 0.753 |
| **FA Cup** | | | |
| *Hierar.+Max* | 0.923 | 0.337 | 0.582 |
| *Hierar.+Length* | 0.923 | 0.317 | 0.582 |
| *GMM+Max* | 0.923 | 0.267 | 0.582 |
| *GMM+Length* | 0.923 | 0.162 | 0.673 |
| **US elections** | | | |
| *GMM+Max* | 0.844 | 0.232 | 0.571 |

Table 3: Best results for the different datasets after evaluating top 5 topics per slot. T-REC, K-PREC, and K-REC refers to topic-recall and keyword-precision/recall respectively.

making this type of trend detection much harder. Looking back at the distribution of the tweets over time (Figure 1), we can see clear spikes in the FA Cup graph, each corresponding to a major event (kick-off, goals, half-time, full-time etc.). No such clarity is in the politics graphs, which instead is best viewed as many overlapping trends.

This difference is reflected in the way that major news stories often emerge: an initial single, focussed story emerges but is later replaced with several potentially overlapping sub-stories covering different aspects of the story. Our results suggest that a dynamic approach may be required for newsworthy topic detection, finding an initial clear burst and subsequently seeking more subtle and overlapping topics.

Recently, Twitter has been actively increasing its ties to television[4]. Broadcast television and sporting events share several common features: they occur a pre-specified times; they attract large audiences; and they are fast-paced. These features all allow and encourage audience participation in the form of sharing comments and holding discussions during the events themselves, such that the

[4] "Twitter & TV: Use the power of television to grow your impact" `https://business.twitter.com/twitter-tv`

| *Target topic* | *Ground truth keywords* | *Extracted keywords* | *Example tweet* |
|---|---|---|---|
| Newt Gingrich says "Thank you Georgia! It is gratifying to win my home state so decisively to launch our March Momentum" | Newt Gingrich, Thank you, Georgia, March, Momentum, gratifying | launch, March, Momentum, decisively, thank, Georgia, gratifying, win, home, state, #MarchMo, #250gas, @newtgingrich | **@Bailey_Shel**: RT @newtgingrich: Thank you Georgia! It is gratifying to win my home state so decisively to launch our March Momentum. #MarchMo #250gas |
| Salomon Kalou has an effort at goal from outside the area which goes wide right of the goal | Salomon Kalou, run, box, mazy | Liverpool, defence, before, gets, ambushed, Kalou, box, mazy, run, @chelseafc, great, #cfcwembley, #facup, shoot | **@SharkbaitHooHa_**: RT @chelseafc: Great mazy run by Kalou into the box but he gets ambushed by the Liverpool defence before he can shoot #CFCWembley #FACup |
| US President Barack Obama has pledged "the best is yet to come", following a decisive re-election victory over Republican challenger Mitt Romney | Obama, best, come | America, best, come, United, States, hearts, #Obama, speech, know, victory | **@northoaklandnow**: "We know in our hearts that for the United States of America, the best is yet to come," says #Obama in victory speech. |

Table 4: Examples of the mainstream media topics, the target keywords, the topics extracted by the $df\text{-}idf_t$ algorithm, and example tweets selected by our system from the collections.

focus of the discussion is constantly moving with the event itself. Potentially, this can allow targeted time-sensitive promotions and advertising based on topics currently receiving the most attention. Facebook and other social media are also competing for access to this potentially valuable "second screen" [23]. Television shows are increasingly promoting hashtags in advance, which may make collecting relevant tweets more straightforward. Even if topic detection with news requires slightly different methods compared to sport and television, both have substantial and growing demand.

**Acknowledgments** This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975. We wish to thank Nic Newman and Steve Schifferes of City University London for invaluable advice.

## References

1. Newman, N.: Mainstream media and the distribution of news in the age of social discovery. Reuters Institute for the Study of Journalism working paper (September 2011)
2. Aiello, L., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., Jaimes, A.: Sensing trending topics in twitter. Multimedia, IEEE Transactions on **15**(6) (2013) 1268–1282
3. Martin, C., Corney, D., Goker, A.: Finding newsworthy topics on Twitter. IEEE Computer Society Special Technical Community on Social Networking E-Letter **1**(3) (September 2013)

4. Newman, N.: #ukelection2010, mainstream media and the role of the internet. Reuters Institute for the Study of Journalism working paper (July 2010)
5. Petrovic, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to Twitter. In: Proceedings of NAACL. Volume 10. (2010)
6. Petrovic, S., Osborne, M., Lavrenko, V.: Using paraphrases for improving first story detection in news and Twitter. In: Proceedings of HTL12 Human Language Technologies. (2012) 338–346
7. Benhardus, J.: Streaming trend detection in Twitter. National Science Foundation REU for Artificial Intelligence, Natural Language Processing and Information Retrieval, University of Colarado (2010) 1–7
8. Shamma, D., Kennedy, L., Churchill, E.: Peaks and persistence: modeling the shape of microblog conversations. In: Proceedings of the ACM 2011 conference on Computer supported cooperative work, ACM (2011) 355–358
9. Phuvipadawat, S., Murata, T.: Breaking news detection and tracking in Twitter. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Volume 3. (2010) 120–123
10. Phuvipadawat, S., Murata, T.: Detecting a multi-level content similarity from microblogs based on community structures and named entities. Journal of Emerging Technologies in Web Intelligence **3**(1) (2011) 11–19
11. Sayyadi, H., Hurst, M., Maykov, A.: Event detection and tracking in social streams. In: Proceedings of International Conference on Weblogs and Social Media (ICWSM). (2009)
12. Ozdikis, O., Senkul, P., Oguztuzun, H.: Semantic expansion of hashtags for enhanced event detection in Twitter. In: Proceedings of VLDB 2012 Workshop on Online Social Systems. (2012)
13. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., Menczer, F.: Detecting and tracking political abuse in social media. Proc. of ICWSM (2011)
14. Alvanaki, F., Sebastian, M., Ramamritham, K., Weikum, G.: Enblogue: emergent topic detection in Web 2.0 streams. In: Proceedings of the 2011 international conference on Management of data, ACM (2011) 1271–1274
15. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on Twitter. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM11). (2011)
16. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28**(1) (1972) 11–21
17. Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. The Computer Journal **26**(4) (1983) 354–359
18. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB. Volume 1215. (1994) 487–499
19. Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal **41**(8) (1998) 578–588
20. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) (1977) 1–38
21. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3** (Mar 2003) 993–1022
22. Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., Ounis, I.: Bieber no more: First story detection using Twitter and Wikipedia. In: SIGIR 2012 Workshop on Time-aware Information Access. (2012)
23. Goel, V., Stelter, B.: Social networks in a battle for the second screen. The New York Times (October 2 2013)

# Multimodal Sentiment Analysis of Social Media

Diana Maynard[1], David Dupplaw[2], and Jonathon Hare[2]

[1] University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
diana@dcs.shef.ac.uk
[2] University of Southampton, Web and Internet Science
Southampton, Hampshire, SO17 1BJ, UK
dpd|jsh2@ecs.soton.ac.uk

**Abstract** This paper describes the approach we take to the analysis of social media, combining opinion mining from text and multimedia (images, videos, etc), and centred on entity and event recognition. We examine a particular use case, which is to help archivists select material for inclusion in an archive of social media for preserving community memories, moving towards structured preservation around semantic categories. The textual approach we take is rule-based and builds on a number of sub-components, taking into account issues inherent in social media such as noisy ungrammatical text, use of swear words, sarcasm etc. The analysis of multimedia content complements this work in order to help resolve ambiguity and to provide further contextual information. We provide two main innovations in this work: first, the novel combination of text and multimedia opinion mining tools; and second, the adaptation of NLP tools for opinion mining specific to the problems of social media.

## 1 Introduction

Social web analysis is all about the users who are actively engaged and generate content. This content is dynamic, reflecting the societal and sentimental fluctuations of the authors as well as the ever-changing use of language. Social networks are pools of a wide range of articulation methods, from simple "Like" buttons to complete articles, their content representing the diversity of opinions of the public. User activities on social networking sites are often triggered by specific events and related entities (e.g. sports events, celebrations, crises, news articles) and topics (e.g. global warming, financial crisis, swine flu).

With the rapidly growing volume of resources on the Web, archiving this material becomes an important challenge. The notion of community memories extends traditional Web archives with related data from a variety of sources. In order to include this information, a semantically-aware and socially-driven preservation model is a natural way to go: the exploitation of Web 2.0 and the wisdom of crowds can make web archiving a more selective and meaning-based process. The analysis of social media can help archivists select material for inclusion, while social media mining can enrich archives, moving towards structured preservation around semantic categories. In this paper, we focus on

the challenges in the development of opinion mining tools from both textual and multimedia content.

We focus on two very different domains: socially aware federated political archiving (realised by the national parliaments of Greece and Austria), and socially contextualized broadcaster web archiving (realised by two large multimedia broadcasting organizations based in Germany: Sudwestrundfunk and Deutsche Welle). The aim is to help journalists and archivists answer questions such as what the opinions are on crucial social events, how they are distributed, how they have evolved, who the opinion leaders are, and what their impact and influence is.

Alongside natural language, a large number of the interactions which occur between social web participants include other media, in particular images. Determining whether a specific non-textual media item is performing as an opinion-forming device in some interaction becomes an important challenge, more so when the textual content of some interaction is small or has no strong sentiment. Attempting to determine a sentiment value for an image clearly presents great challenges, and this field of research is still in its infancy. We describe here some work we have been undertaking, firstly to attempt to provide a sentiment value from an image outside of any specific context, and secondly to utilise the multimodal nature of the social web to assist the sentiment analysis of either the multimedia or the text.

## 2 Related Work

While much work has recently focused on the analysis of social media in order to get a feel for what people think about current topics of interest, there are, however, still many challenges to be faced. State of the art opinion mining approaches that focus on product reviews and so on are not necessarily suitable for our task, partly because they typically operate within a single narrow domain, and partly because the target of the opinion is either known in advance or at least has a limited subset (e.g. film titles, product names, companies, political parties, etc.).

In general, sentiment detection techniques can be roughly divided into lexicon-based methods [22] and machine-learning methods, e.g. [1]. Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. Machine learning approaches make use of syntactic and/or linguistic features, and hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods. For example, [17] establish the polarity of reviews by identifying the polarity of the adjectives that appear in them, with a reported accuracy of about 10% higher than pure machine learning techniques. However, such relatively successful techniques often fail when moved to new domains or text types, because they are inflexible regarding the ambiguity of sentiment terms. The context in which a term is used can change its meaning, particularly for adjectives in sentiment lexicons [18]. Several evaluations have shown the usefulness of contextual information [26], and have identified context

words with a high impact on the polarity of ambiguous terms [8]. A further bottleneck is the time-consuming creation of these sentiment dictionaries, though solutions have been proposed in the form of crowdsourcing techniques[3].

Almost all the work on opinion mining from Twitter has used machine learning techniques. [19] aimed to classify arbitrary tweets on the basis of positive, negative and neutral sentiment, constructing a simple binary classifier which used n-gram and POS features, and trained on instances which had been annotated according to the existence of positive and negative emoticons. Their approach has much in common with an earlier sentiment classifier constructed by [9], which also used unigrams, bigrams and POS tags, though the former demonstrated through analysis that the distribution of certain POS tags varies between positive and negative posts. One of the reasons for the relative paucity of linguistic techniques for opinion mining on social media is most likely due to the difficulties in using NLP on low quality text [7]; for example. the Stanford NER drops from 90.8% F1 to 45.88% when applied to a corpus of tweets [14].

There have been a number of recent works attempting to detect sarcasm in tweets and other user-generated content [23, 13, 20, 5], with accuracy typically around 70-80%. These mostly train over a set of tweets with the #sarcasm and/or #irony hashtags, but all simply try to classify whether a sentence or tweet is sarcastic or not (and occasionally, into a set of pre-defined sarcasm types). However, none of these approaches go beyond the initial classification step and thus cannot predict how the sarcasm will affect the sentiment expressed. This is one of the issues that we tackle in our work.

Extracting sentiment from images is still a research area that is in its infancy and not yet prolifically published. However, those published often use small datasets for their ground truth on which to build SVM classifiers. Evaluations show systems often respond only a little better than chance for trained emotions from general images [27]. The implication is that the feature selection for such classification is difficult. [25] used a set of colour features for classifying their small ground-truth dataset, also using SVMs, and publish an accuracy of around 87%. In our work, we expand this colour-based approach to use other features and also use the wisdom of the crowd for selecting a large ground-truth dataset.

Other papers have begun to hint at the multimodal nature of web-based image sentiment. Earlier work, such as [11], is concerned with similar multimodal image annotation, but not specifically for sentiment. They use latent semantic spaces for correlating image features and text in a single feature space. In this paper, we describe the work we have been undertaking in using text and images together to form sentiment for social media.

## 3 Opinion Mining from Text

### 3.1 Challenges

There are many challenges inherent in applying typical opinion mining and sentiment analysis techniques to social media. Microposts such as tweets are, in

[3] `http://apps.facebook.com/sentiment-quiz`

some sense, the most challenging text type for text mining tools, and in particular for opinion mining, since the genre is noisy, documents have little context and assume much implicit knowledge, and utterances are often short. As such, conventional NLP tools typically do not perform well when faced with tweets [2], and their performance also negatively affects any following processing steps.

Ambiguity is a particular problem for tweets, since we cannot easily make use of coreference information: unlike in blog posts and comments, tweets do not typically follow a conversation thread, and appear much more in isolation from other tweets. They also exhibit much more language variation, and make frequent use of emoticons, abbreviations and hashtags, which can form an important part of the meaning. Typically, they also contain extensive use of irony and sarcasm, which are particularly difficult for a machine to detect. On the other hand, their terseness can also be beneficial in focusing the topics more explicitly: it is very rare for a single tweet to be related to more than one topic, which can thus aid disambiguation by emphasising situational relatedness.

In longer posts such as blogs, comments on news articles and so on, a further challenge is raised by the tracking of changing and conflicting interpretations in discussion threads. We investigate first steps towards a consistent model allowing for the pinpointing of opinion holders and targets within a thread (leveraging the information on relevant entities extracted).

We refer the reader to [2] for our work on twitter-specific IE, which we use as pre-processing for the opinion mining described below. It is not just tweets that are problematic, however; sarcasm and noisy language from other social media forms also have an impact. In the following section, we demonstrate some ways in which we deal with this.

### 3.2 Opinion Mining Application

Our approach is a rule-based one similar to that used by [22], focusing on building up a number of sub-components which all have an effect on the score and polarity of a sentiment. In contrast, however, our opinion mining component finds opinions relating to previously identified entities and events in the text. The core opinion mining component is described in [15], so we shall only give an overview here, and focus on some issues specific to social media which were not dealt with in that work, such as sarcasm detection and hashtag decomposition.

The detection of the actual opinion is performed via a number of different phases: detecting positive, negative and neutral words, identifying factual or opinionated versus questions or doubtful statements, identifying negatives, sarcasm and irony, analysing hashtags, and detecting extra-linguistic clues such as smileys. The application involves a set of grammars which create annotations on segments of text. The grammar rules use information from gazetteers combined with linguistic features (POS tags etc.) and contextual information to build up a set of annotations and features, which can be modified at any time by further rules. The set of gazetteer lists contains useful clues and context words: for example, we have developed a gazetteer of affect/emotion words from Word-

Net [16]. The lists have been modified and extended manually to improve their quality.

Once sentiment words have been matched, we find a linguistic relation between these and an entity or event in the sentence or phrase. A Sentiment annotation is created for that entity or event, with features denoting the polarity (positive or negative) and the polarity score. Scores are based on the initial sentiment word score, and intensified or decreased by any modifiers such as swear words, adverbs, negation, sarcasm etc, as explained next.

Swear words are particularly prolific on Twitter, especially on topics such as popular culture, politics and religion, where people tend to have very strong views. To deal with these, we match against a gazetteer list of swear words and phrases, which was created manually from various lists found on the web and from manual inspection of the data, including some words acquired by collecting tweets with swear words as hashtags (which also often contain more swear words in the main text of the tweet).

Much useful sentiment information is contained within hashtags, but this is problematic to identify because hashtags typically contain multiple words within a single token, e.g. #notreally. If a hashtag is camelcased, we use the capitalisation information to create separate tokens. Second, if the hashtag is all lowercase or all uppercase, we try to form a token match against the Linux dictionary. Working from left to right, we look for the longest match against a known word, and then continue from the next offset. If a combination of matches can be found without a break, the individual components are converted to tokens. In our example, #notreally would be correctly identified as "not" + "really". However, some hashtags are ambiguous: for example, "#greatstart" gets split wrongly into the two tokens "greats" + "tart". These problems are hard to deal with; in some cases, we could make use of contextual information to assist.

We conducted an experiment to measure the accuracy of hashtag decomposition, using a corpus of 1000 tweets randomly selected from the US elections crawl that we undertook in the project. 944 hashtags were detected in this corpus, of which 408 were identified as multiword hashtags (we included combinations of letters and numbers as multiword, but not abbreviations). 281 were camelcased and/or combinations of letters and nubers, 27 were foreign words, and the remaining 100 had no obvious token-distinguishing features. Evaluation on the hard-to-recognise cases (non-camel-cased multiword hashtags) produced scores of 86.91% Precision, 90% Recall, and an F-measure of 88.43%. Given that these hard-to-resolve combinations form roughly a quarter of the multiword hashtags in our corpus, and that we are entirely successful in decomposing the remaining hashtags, this means that the overall accuracy for hashtag decomposition is much higher.

In addition to using the sentiment information from these hashtags, we also collect new hashtags that typically indicate sarcasm, since often more than one sarcastic hashtag is used. For this, we used the GATE gazetteer list collector to collect pairs of hashtags where one was known to be sarcastic, and examined the second hashtag manually. From this we were able to identify a further set

of sarcasm-indicating hashtags, such as #thanksdude, #yay etc. Further investigation needs to be performed on these to check how frequently they actually indicate sarcasm when used on their own.

Finally, emoticons are processed like other sentiment-bearing words, according to another gazetteer list, if they occur in combination with an entity or event. For example, the tweet "They all voted Tory :-(" would be annotated as negative with respect to the target "Tory". Otherwise, as for swear words, if a sentence contains a smiley but no other entity or event, the sentence gets annotated as sentiment-bearing, with the value of that of the smiley from the gazetteer list.

Once all the subcomponents have been run over the text, a final output is produced for each sentiment-bearing segment, with a polarity (positive or negative) and a score, based on combining the individual scores from the various components (for example, the negation component typically reverses the polarity, the adverbial component increases the strength of the sentiment, and so on. Aggregation of sentiment then takes place for all mentions of the same entity/event in a document, so that summaries can be created.

## 4 Extracting Opinions from Images

### 4.1 Challenges

The main challenge with annotating non-textual media is that the underlying tokens within it are considerably less explicit than in textual media. In images and video, these underlying tokens are groups of pixels (compared with groups of characters [words] in text). As well as having multiple dimensions, the tokens have considerably more variation when representing exactly the same concept, and so using dictionaries and other traditional text-based tools becomes impossible. And so, we enter the world of image understanding and computer vision which, although over 30 years old, has made fewer revolutionary leaps than NLP. State of the art computer vision is still relatively basic for most general applications. This "semantic gap" between what computer vision can achieve and the level of understanding required for sentiment analysis is why extracting opinions from images is so difficult.

That said, certain specific applications have made advances recently - one of which is the application of computer vision for detecting and recognising faces of people. [24] developed a technique for face detection that is probably the most widespread computer-vision technique of all time, as most point-and-shoot cameras include face detection based on this algorithm. It uses some 1-dimensional peak features (Haar features) that are used to train a cascade of classifiers for general object detection. Trained on faces, these can detect faces in images robustly and efficiently.

Detecting the presence of a face is just the first part; fitting a model to a face can then provide some extra information about the shape and the expression of the face. Active Shape Models [3] (ASM) and Active Appearance Models [4] (AAM) are well-known algorithms for fitting a shape to an image

using the image's gradients to choose the best position for the vertices of the shape. As these models are parametric and generative (they are reconstructed using a small number of parameters), a large range of poses, expressions and appearances (skin textures) can be generated. Fitting a model to an image is a constrained optimisation problem in which the parameters of the model are iteratively updated in order to minimise the difference between the generated model and the image (hence Constrained Local Model [CLM]). Once a model is fitted to an image, the parameters can then be used as input to an expression classifier that can determine an expression label for the face.

This model fits well with the Facial Action Coding System (FACS) which aims to provide a standardised way of describing the expressions of faces. Codes represent muscular actions in the face (such as "inner eyebrow raising", or "lip corner puller") and when combined they represent emotions (for example, activation of the lip corner puller AU6 and the cheek raiser AU12 actions imply happiness). These muscular movements map to combinations of parameters in the face model, so a classifier can be trained to recognise these actions. Of course, this relies on accurate face model fitting, but it is difficult to build a shape model (ASM, AAM or CLM) that will accurately fit all faces and poses, which is essential for the accurate measurement of the shape parameters needed for expression classification. Another problem is that accurate detection of a face is required to initialise the fitting of a face model; whilst face detection techniques are quite mature, they can still have major problems working in real-world images where the faces are not exactly frontal to the camera, or there are shadows or contrast issues.

### 4.2 Detecting Sentiment in Images

Figure 1 shows an example of a programme that recognises the expressions in a laboratory setting. In the wild, we found that inaccuracies in the face model alignment would regularly cause misclassification of the action units, and therefore the expressions.

In less constrained multimedia, we cannot rely on there being faces in the images, and sentiment may be carried by other visual traits. Indeed, images may intrinsically have sentiment associated with them through design (say a poster for a horror film) or through association with a specific subject matter which may be context sensitive (say a photo of wind generators in the context of climate change). For these situations there are no specific algorithms we can use for extracting the sentiment. However, we can perform standard feature-label correlations using classifiers over ground-truth datasets. Unfortunately, large, well labelled datasets for image sentiment are very thin on the ground. For that reason, we turned to the "wisdom of the crowd" for generating a large dataset to use for classification. Using SentiWordNet, we were able to query Flickr for the words that had the strongest positive and negative sentiments, and retrieve sets of images for each of them. Combined, these formed our ground-truth for positive and negative sentiment images. The details of this work are described in [21], but we will summarise the conclusions here.

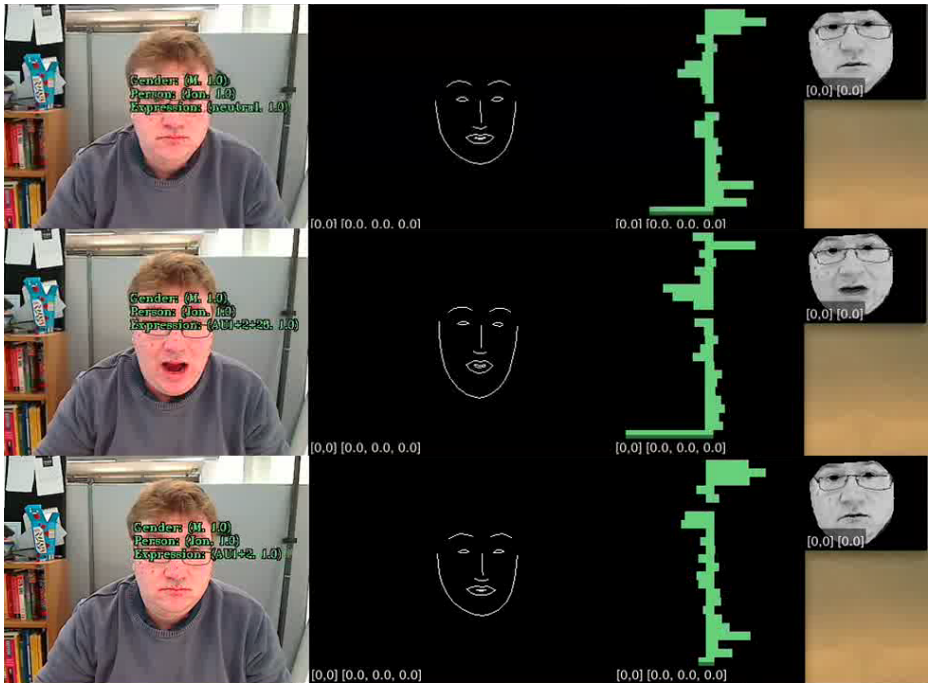



**Figure 1.** Recognition of expressions in a laboratory setting

We gathered images for the 1000 strongest sentiment words from SentiWordNet. This resulted in 586,000 images, most of which had a resolution of more than 1 megapixel. We extracted global and local colour features and SIFT local features from the images, and trained an SVM classifier to recognise positive/negative sentiment. We can observe that for small recall values, precision values of up to 70% can be reached. Due to the challenging character of this task, for high recall values, the precision degrades down to the random baseline. Interestingly, using mutual information, we were able to reverse engineer the correlations in the classifier to determine which features were correlated to which labels. We found that positive images had overall warm colours (reds, oranges, yellows, skin tones) and negative images had colder colours (blues, dark greens). The location of the colour had no real significance. The negative SIFT features seem dominated by a very light central blob surrounded by a much darker background, while the positive SIFT features are dominated by a dark blob on the side of the patch.

Clearly, from a basis where there is no context, it is only possible to achieve a limited amount of understanding. However, using the contextual information (e.g. co-located text) it is possible to aggregate various forms of analysis and make further estimates of an object's sentiment. To do that, it is necessary to find the text and images which are co-located. In web pages, we can extract the

'important' part of the page using boilerplate removal tools, such as our tool Readability4J [12]. In tweets, images are usually presented as links, usually to a URL shortener. It is necessary to follow the links to their destination, then to parse the final destination for a the "co-located" image. Once we have images related to the text, we look for entities within the visual content. As described in Section 3, we extract entities from the text and associate a sentiment value with them based on the textual context. These entities will be people, locations, or organisations and can be used to guide our analysis of the associated images. It is impractical to consider an entity recognition system that would recognise any entity (e.g. any person or any place), so we can use the entities in the text to reduce the search space. For example, we can use the detected person entities to train a face recognition system (for previously unseen people, on-the-fly using the image search results from major search engines), the location entites to fix a prior on a world-based search algorithm (as our work in [6]), or the organisation entities to train a logo detector.

One of the interesting insights into the social web is to see how media is spread – how it is reused and talked about and whether the sentiment associated with the media changes. We developed a system called Twitter's Visual Pulse [10] which finds near-duplicate images from a live or static Twitter stream. We used a technique called Locality Sensitive Hashing (LSH) of SIFT features extracted from the images, and determine near-duplicates by finding connected components in a graph where nodes are hashed features and edges are weighted based on the number of matching hashes. By extracting the sentiment from the tweets associated with these duplicate images, we can find out how the image is used in different contexts. In many cases, the image may be reused in contexts which are, overall, sentimentally ambivalent; however, there may be cases where an image is used in a consistent way - for example, a particular image may be used in consistently positive tweets. We form a discrete probability distribution for images falling in specific sentiment categories, which we can use to assign sentiment probabilities to the image when it is further reused, particularly in cases where the textual sentiment analysis is inconclusive.

## 5 Conclusions

In this paper, we have described the general approach we undertake to the analysis of social media, using a combination of textual and multimedia opinion mining tools. It is clear that both opinion mining in general, and the wider analysis of social media, are difficult tasks from both perspectives, and there are many unresolved issues. The modular nature of our approach also lends itself to new advances in a range of subtasks: from the difficulties of analysing the noisy forms of language inherent in tweets, to the problems of dealing with sarcasm in social media, to the ambiguities inherent in such forms of web content that inhibit both textual and multimedia analysis tools. Furthermore, to our knowledge this is the first system that attempts to combine such kinds of textual and multimedia analysis tools in an integrated system, and preliminary results

are very promising, though this is nevertheless very much ongoing research. Future work includes further development of the opinion mining tools: we have already begun investigations into issues such as sarcasm detection, more intricate use of discourse analysis and so on.

## Acknowledgments

This work was supported by the European Union under grant agreements No. 270239 Arcomem[4] and No. 610829 DecarboNet[5].

[4] http://www.arcomem.eu
[5] http://www.decarbonet.eu

# Bibliography


[1] Boiy, E., Moens, M.F.: A machine learning approach to sentiment analysis in multilingual web texts. Information Retrieval 12(5), 526–558 (2009)

[2] Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M.A., Maynard, D., Aswani, N.: TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics (2013)

[3] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models – their training and application. Comput. Vis. Image Underst. 61(1), 38–59 (Jan 1995), http://dx.doi.org/10.1006/cviu.1995.1004

[4] Cootes, T., Edwards, G., Taylor, C.: Active appearance models. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23(6), 681–685 (2001)

[5] Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. pp. 107–116. Association for Computational Linguistics (2010)

[6] Davies, J., Hare, J., Samangooei, S., Preston, J., Jain, N., Dupplaw, D., Lewis, P.H.: Identifying the geographic location of an image with a multimodal probability density function. In: MediaEval 2013 / Placing: Geo-coordinate Prediction for Social Multimedia (October 2013), http://eprints.soton.ac.uk/358836/

[7] Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K.: Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. ACM (2013)

[8] Gindl, S., Weichselbraun, A., Scharl, A.: Cross-domain contextualisation of sentiment lexicons. In: Proceedings of 19th European Conference on Artificial Intelligence (ECAI-2010). pp. 771–776 (2010)

[9] Go, A., Bhayani, R., , Huang, L.: Twitter sentiment classification using distant supervision. Tech. Rep. CS224N Project Report, Stanford University (2009)

[10] Hare, J., Samangooei, S., Dupplaw, D., Lewis, P.H.: Twitter's visual pulse. In: 3rd ACM International conference on multimedia retrieval. pp. 297–298 (April 2013)

[11] Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: A linear-algebraic technique with an application in semantic image retrieval. In: Sundaram, H., Naphade, M.R., Smith, J.R., Rui, Y. (eds.) CIVR. Lecture Notes in Computer Science, vol. 4071, pp. 31–40. Springer (2006), http://dblp.uni-trier.de/db/conf/civr/civr2006.html#HareLES06

[12] Hare, J.S., Samangooei, S., Dupplaw, D.P.: OpenIMAJ and ImageTerrier: Java libraries and tools for scalable multimedia analysis and indexing of

images. In: Proceedings of the 19th ACM International Conference on Multimedia. pp. 691–694. ACM, New York, NY, USA (2011)
[13] Liebrecht, C., Kunneman, F., van den Bosch, A.: The perfect solution for detecting sarcasm in tweets# not. WASSA 2013 p. 29 (2013)
[14] Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 359–367 (2011)
[15] Maynard, D., Bontcheva, K., Rout, D.: Challenges in developing opinion mining tools for social media. In: Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012. Turkey (2012)
[16] Miller, G.A., Beckwith, R., Felbaum, C., Gross, D., Miller, C.Miller, G.A., Beckwith, R., Felbaum, C., Gross, D., Miller, C.Minsky, M.: Five papers on WordNet (1990)
[17] Moghaddam, S., Popowich, F.: Opinion polarity identification through adjectives. CoRR abs/1011.4623 (2010)
[18] Mullaly, A., Gagné, C., Spalding, T., Marchak, K.: Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning. The Mental Lexicon 5(1), 87–114 (2010)
[19] Pak, A., Paroubek, P.: Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 436–439 (2010), http://www.aclweb.org/anthology/S10-1097
[20] Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Language Resources and Evaluation pp. 1–30 (2013)
[21] Siersdorfer, S., Hare, J., Minack, E., Deng, F.: Analyzing and predicting sentiment of images on the social web. In: ACM Multimedia 2010. pp. 715–718. ACM (October 2010), http://eprints.ecs.soton.ac.uk/21670/
[22] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics 1(September 2010), 1–41 (2011)
[23] Tsur, O., Davidov, D., Rappoport, A.: Icwsm–a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 162–169 (2010)
[24] Viola, P., Jones, M.: Robust real-time object detection. In: International Journal of Computer Vision (2001)
[25] Wei-ning, W., Ying-lin, Y., Sheng-ming, J.: Image retrieval by emotional semantics: A study of emotional space and feature extraction. In: Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on. vol. 4, pp. 3534–3539 (2006)
[26] Weichselbraun, A., Gindl, S., Scharl, A.: A context-dependent supervised learning approach to sentiment detection in large textual databases. Journal of Information and Data Management 1(3), 329–342 (2010)
[27] Yanulevskaya, V., Van Gemert, J., Roth, K., Herbold, A.K., Sebe, N., Geusebroek, J.M.: Emotional valence categorization using holistic image features. In: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. pp. 101–104 (2008)