Language Engineering for Online Reputation Management

26 May 2012

PROCEEDINGS

Editors:

Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, Irina Chugur

Language Engineering for Online Reputation Management

Workshop Programme

9:00 - 9:45 – Introduction to the NLP & ORM Challenge
Julio Gonzalo (UNED), *The RepLab Initiative: An Evaluation Campaign for Online Reputation Management*

Hugo Zaragoza (WebSays), *Online Reputation Management: Business Requirements and Scientific Challenges*

Miguel Lucas (Acteo), *Online Reputation Management: Analysis of Existing Commercial Tools*

9:45 - 10:30 – Position papers
Alexandra Balahur (JRC), *The Challenge of Processing Opinions in Online Contents in the Social Web Era*

Patrick Brennan (Juola & Associates), *Tagging Commentary with Demographic Data*

Fredrik Olsson, Jussi Karlgren, Magnus Sahlgren, Fredrik Espinoza, Ola Hamfors, (Gavagai), *Technical Requirements For Knowledge Representation For Reputation Mining On A Realistic Scale*

10:30 - 11:00 – Coffee Break

11:00 - 11:45 – Technical Papers
Chandra Mohan Dasari, Dipankar Das, Sivaji Bandyopadhyay (Jadavpur University), *Topic Identification from Blog Documents: Roles of Bigram, Named Entity and Sentiment*

Yue Dai, Ernest Aredarenko, Tuomo Kakkonen, Ding Liao (University of Eastern Finland), *Towards SoMEST – Combining Social Media Monitoring with Event Extraction and Timeline Analysis*

Damiano Spina (UNED), Edgar Meij, Andrei Oghina, Minh Thuong Bui, Mathias Breuss, Maarten de Rijke (University of Amsterdam), *A Corpus for Entity Profiling in Microblog Posts*

11:45 – 13:00 – Roadmap Discussion & Wrap-up
Jordi Atserias (Yahoo! Research Barcelona)
Adolfo Corujo (Llorente & Cuenca)
Julio Gonzalo (UNED)
Miguel Lucas (Acteo)
Edgar Meij (University of Amsterdam)
Maarten de Rijke (University of Amsterdam)
Hugo Zaragoza (WebSays) plus all workshop participants

Editors

Adolfo Corujo	Llorente & Cuenca, Spain
Julio Gonzalo	UNED, Spain
Edgar Meij	University of Amsterdam, The Netherlands
Maarten de Rijke	University of Amsterdam, The Netherlands
Irina Chugur	UNED, Spain

Workshop Organizers/Organizing Committee

Adolfo Corujo	Llorente & Cuenca, Spain
Julio Gonzalo	UNED, Spain
Edgar Meij	University of Amsterdam, The Netherlands
Maarten de Rijke	University of Amsterdam, The Netherlands

Workshop Programme Committee

Eugene Agichtein	Emory University, USA
Alexandra Balahur	JRC, Italy
Krisztian Balog	NTNU, Norway
Raymond Franz	Trendlight, The Netherlands
Donna Harman	NIST, USA
Eduard Hovy	ISI/USC, USA
Radu Jurca	Google, Switzerland
Jussi Karlgren	Gavagai/SICS, Sweden
Mounia Lalmas	Yahoo! Research, Spain
Jochen Leidner	Thomson Reuters, Switzerland
Bing Liu	U. Illinois at Chicago, USA
Alessandro Moschitti	U. Trento, Italy
Miles Osborne	U. Edinburgh, UK
Hans Uszkoreit	U. Saarbrucken, Germany
James Shanahan	Boston U., USA
Belle Tseng	Yahoo!, USA
Julio Villena	Daedalus/U. Carlos III, Spain

Table of contents

Author Index

Preface

This volume collects technical and position papers for the LREC Workshop on Language Engineering for Online Reputation Management held in Istanbul on May 26, 2012.

Online Reputation Management deals with the image that online media project about individuals and organizations. The growing relevance of social media and the speed at which facts and opinions travel in microblogging networks make online reputation an essential part of a company's public relations.

While traditional reputation analysis was based mostly on manual analysis (clipping from media, surveys, etc.), the key value from online media comes from the ability of processing, understanding and aggregating potentially huge streams of facts and opinions about a company or individual. Information to be mined includes answers to questions such as: What is the general state of opinion about a company/individual in online media? What are its perceived strengths and weaknesses, as compared to its peers/competitors? How is the company positioned with respect to its strategic market? Can incoming threats to its reputation be detected early enough to be neutralized before they effectively affect reputation?

In this context, Natural Language Processing plays a key, enabling role, and we are already witnessing an unprecedented demand for text mining software in this area. Note that, while the area of opinion mining has made significant advances in the last few years, most tangible progress has been focused on products. However, mining and understanding opinions about companies and individuals is, in general, a much harder and less understood problem.

The aim of the workshop was to bring together the Language Engineering community (including researchers and developers) with representatives from the Online Reputation Management industry, a fast-growing sector which poses challenging demands to text mining technologies. The goal was to establish a five-year roadmap on the topic, focusing on what language technologies are required to get there in terms of resources, algorithms and applications. The workshop is tightly connected to RepLab, an evaluation initiative for Online Reputation Management Systems which has its first edition as a CLEF 2012 lab, in September 2012. The outcome of the workshop is intended to serve as direct input to establish the research priorities of RepLab.

With this purpose in mind, the workshop included both research papers and position statements from industry and academia. Besides paper presentations, the agenda of the workshop includes a session introducing the problem from a dual business and academic perspective, and a discussion session aimed at establishing a roadmap for the topic. The workshop is partially supported by the EU project Limosine (under project number 288024, call FP7-ICT-2011-7).

Position Papers

The Challenge of Processing Opinions Expressed in Online Contents in the Social Web Era

Alexandra Balahur

European Commission Joint Research Centre
Institute for the Protection and Security of the Citizen
GlobeSec - OPTIMA (OPensource Text Information Mining and Analysis)
Via Fermi 2749, T.P. 267
I-21027 Ispra (VA), Italy

alexandra.balahur@jrc.ec.europa.eu

Abstract

In the new Social Web era, the globalization of markets combined with the fact that people can freely express their opinion on any product or company on forums, blogs or e-commerce sites led to a change in the companies' marketing strategies, in the rise of awareness for client needs and complaints, and a special attention for brand trust and reputation. Specialists in market analysis, but also IT fields such as Natural Language Processing (NLP), demonstrated that in the context of the newly created opinion phenomena, decisions for economic action are not only given by factual information, but are highly affected by rumors and negative opinions. In this context, analyzing online reputation and being able to understand the mechanisms through which opinions are spread and the extent and manner in which they influence the business, social and political spheres become necessary endeavors. The problem in this context is much more difficult to solve, as entities, as opposed to products, are related to different events and topics and there is no fixed set of "attributes" that are commented on by persons expressing opinions on these entities. Additionally, answering opinion questions is an issue that is far from being trivial. This paper describes the challenges related to mining opinions for reputation management in the Social Web context.

Keywords: online reputation management, sentiment analysis, opinion mining, Social Web.

1. Introduction

The era in which we live has been given many names. "Global village", "technotronic era", "post-industrial society", "information society", "information age", and "knowledge society" are just a few of the terms that have been used in an attempt to describe the deep changes that have occurred in the lives of societies and people worldwide as a result of the fast development of ICT technologies, the access to Internet and its transformation into a Social Web. In this context, more than ever before, people are more than willing and happy to share their lives, knowledge, experience and thoughts with the entire world, through blogs, forums, wikis, review sites or microblogs. They are actively participating to events, by expressing their opinions on them, by commenting on the news appearing and the events that take place in all spheres of the society. The large volume of subjective information present on the Internet, in reviews, forums, blogs, microblogs and social network communications has produced an important shift in the manner in which people communicate, share knowledge and emotions and influence the social, political and economic behavior worldwide. The radical shift in the method employed for communication and the content of this communication has brought with itself new challenges, but also many opportunities.

At the economic level, the globalization of markets combined with the fact that people can freely express their opinion on any product or company on forums, blogs or e-commerce sites led to a change in the companies' marketing strategies, in the rise of awareness for client

needs and complaints, and a special attention for brand trust and reputation. Specialists in market analysis, but also IT fields such as Natural Language Processing (NLP), demonstrated that in the context of the newly created opinion phenomena, decisions for economic action are not only given by factual information, but are highly affected by rumors and negative opinions. Wright (2009)[1] claims that "for many businesses, online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace".

In this context, analyzing online reputation and being able to understand the mechanisms through which opinions are spread and the extent and manner in which they influence the business, social and political spheres become necessary endeavors.

The problem in this context is much more difficult to solve, as entities, as opposed to products, are related to different events and topics and there is no fixed set of "attributes" that are commented on by persons expressing opinions on these entities. There is only one freely accessible system performing such as a task - Lydia (Skiena et al., 2007), which gathers news from portals and blogs and classifies opinions on different entities. However, both this system, as well as different approaches that have been presented for this problem in the research literature, show that the issue of entity-centered opinion mining and, additionally, the correlation of the results with facts over events where these entities are involved are not trivial (Balahur and Steinberger, 2009; Zhang and Skiena, 2010).

In the following sections, we first present an overview of the issues that sentiment analysis in general and online reputation management in particular are concerned and detail on the problems related to each of the presented issues. Finally, we draw some conclusions on the aspects presented.

2. Challenges of Online Reputation Management

The challenges for the future of this task relate to different problems. Further on, we detail on these issues.

First of all, there is a need to **define the task** and **the concepts** it involves, in order to prevent the same issues as in sentiment analysis and opinion mining. Here, there are different tasks that have been tackled under the same umbrella, with different aims in mind and considering very different definitions of the basic concepts (Balahur-Dobrescu, 2011). In the case of opinion, if one were to look at the term definition given in the Webster dictionary[2], they would find the following set of synonyms: "opinion", "view", "belief", "conviction", "persuasion", "sentiment", meaning "a judgment one holds as true". Out of this definition, it is important to stress upon the fact that these closely related, synonym terms, have slightly different meanings.

- ***Opinion*** implies a conclusion thought out yet open to dispute; it is:
  1. A): a view, judgment, or appraisal formed in the mind about a particular matter; B): approval, esteem;
  2. A): a belief stronger than impression and less strong than positive knowledge; B): a generally held view;
  3. A): a formal expression of judgment or advice by an expert; B): the formal expression (as by a judge, court, or referee) of the legal reasons and principles upon which a legal decision is based.
- ***View*** suggests a subjective opinion.
- ***Belief*** implies often deliberate acceptance and intellectual assent.
- ***Conviction*** applies to a firmly and seriously held belief.
- ***Persuasion*** suggests a belief grounded on assurance (as by evidence) of its truth.
- ***Sentiment*** suggests a settled opinion reflective of one's **feelings.**

The term **feeling** is defined as the conscious subjective experience of emotion. (Van den Bos, 2006). This is approximately the same definition as the one given by Scherer (2005), which states that *"the term feeling points to a single component of emotion, denoting the subjective experience process, and is therefore only a small part of an emotion"*.

This definition suggests that there are different types of opinions and that not all opinions are subjective (see the definition of "view"), as well as not all opinions have a sentiment associated to them. An "objective" opinion could

[1] www.nytimes.com/2009/08/24/technology/internet/24emotion.html?_r=1&ref=start-ups

[2] http://www.merriam-webster.com/

be considered to be the one of an expert (e.g. a doctor giving a diagnosis on the basis of observed symptoms). A "subjective" opinion is one that is based on personal criteria (depends on the individual taste, ideas, standards etc.). This same definition also pinpoints to the fact that sentiments are types of opinions, namely the ones that are "reflective of one's feelings", where "feeling" is the "conscious subjective experience of emotion". Thus, sentiment relates to emotion, in the sense that it is the expression of an evaluation based on the emotion the writer feels.

"Opinion mining", as a computational task, appeared for the first time in a paper by Dave et al. (2003), and it was defined as follows: *"Given a set of evaluative text documents D that contain opinions (or sentiments) about an "object" (person, organization, product etc.), opinion mining aims to extract attributes and components of the object that have been commented on in each document d in the set D and to determine whether the comments are positive, negative or neutral."* According to Pang and Lee (2008), the fact that this work appeared in the proceedings of the World Wide Web (WWW) 2003 conference explains the popularity of this terminology within the web search and retrieval research community. This also explains the fact that Esuli and Sebastiani (2006) define ***opinion mining*** *as "a recent discipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses".*

From the computational point of view, Kim and Hovy (2005) define **opinion** *"as a quadruple [Topic, Holder, Claim, Sentiment] in which the Holder believes a Claim about the Topic, and in many cases associates a Sentiment, such as good or bad, with the belief.* As far as sentiments are concerned, the authors define them as: *"Sentiments, which in this work we define as an explicit or implicit expression in text of the Holder's positive, negative, or neutral regard toward the Claim about the Topic. Sentiments always involve the Holder's emotions or desires, and may be present explicitly or only implicitly."*

This definition relates opinion with sentiment, in the sense that it states that some opinions carry a sentiment, while others do not. In order to illustrate the difference between opinions with sentiment and opinions without sentiment, Kim and Hovy (2005) provide the following examples:

(1) "I believe the world is flat."
(2) "The Gap is likely to go bankrupt."

These are sentences that express opinions, but they do not contain any sentiment. The following examples, taken from the same paper, explain the difference between explicitly versus implicitly expressed sentiment of opinions:

(3) "I think that attacking Iraq would put the US in a difficult position." (implicit)
(4) "The US attack on Iraq is wrong." (explicit)
(5) "I like Ike." (explicit)
(6) "We should decrease our dependence on oil." (implicit)

Another definition of the term **opinion** was given by Bing Liu (2010). The author is the one who defined the task of "feature-based opinion mining and summarization", which deals with the classification of opinions expressed on different features of products and their summarization (Hu and Liu, 2004).

According to Liu (2010):

- *"An **opinion** on a feature f is a positive or negative view, attitude, emotion or appraisal on f from an opinion holder."*
- *"The **holder of an opinion** is the person or organization that expresses the opinion."*
- *"An **explicit opinion** on feature f is an opinion explicitly expressed on f in a subjective sentence."*
- *"An **implicit opinion** on feature f is an opinion on f implied in an objective sentence."*
- *"An **opinionated sentence** is a sentence that expresses explicit or implicit positive or negative opinions. It can be a subjective or objective sentence."*
- *"**Emotions** are our subjective feelings and thoughts."*

All tasks defined within opinion mining aim at classifying the texts according to the "orientation of the opinion" (usually into three classes – of positive, negative and neutral). The classes of opinion considered have been denoted using different terms: ***opinion orientation***, ***sentiment polarity***, ***polarity***, ***sentiment orientation***, ***polarity of opinion***, ***semantic orientation***.

As far as sentiment analysis as NLP task is concerned, most of the research in the field coincides with the

following definition: *"The binary classification task of labelling an opinionated document as expressing either an overall positive or an overall negative opinion is called sentiment polarity classification or polarity classification".* (Pang and Lee, 2008)

"The orientation of an opinion on a feature f indicates whether the opinion is positive, negative or neutral. Opinion orientation is also known as sentiment orientation, polarity of opinion, or semantic orientation."(Liu, 2010)

A related concept is ***valence***, defined as *"a negative or positively attitude"* (Polanyi and Zaenen, 2004). In relation to this concept, Polanyi and Zaenen (2004) define the so-called ***"contextual valence shifters"*** *(e.g. negatives and intensifiers, modals, presuppositional items, ironical formulations, connectors),* which are lexical items or formulations that change the orientation of the attitude.

The term ***"sentiment"*** in the context of a computational text analysis task is mentioned for the first time in the paper by Das and Chen (2001). According to the authors *"in this paper, 'sentiment' takes on a specific meaning, that is, the net of positive and negative opinion expressed about a stock on its message board."* . At the same time, Tong (2001) proposed a "new" task at the Workshop on Operational Text Classification (OTC2001), which concerned the detection and tracking of opinions in on-line discussions and the subsequent classification of the sentiment of opinion.

The aim of the paper by Turney (2002) is *"to classify reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations (e.g., "subtle nuances") and a negative semantic orientation when it has bad associations (e.g., "very cavalier")".*

Pang et al. (2002) propose different methods to determine the *"sentiment, or overall opinion towards the subject matter for example, whether a product review is positive or negative".*

Nasukawa and Yi (2003) entitled their paper, "Sentiment analysis: Capturing favorability using natural language processing". In this paper, they state that *"the essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject."*

Yi et al. (2003), in their paper "Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques", consider opinion an equivalent term to sentiment. Their approach approximates the task later known as "feature-based opinion mining and summarization" (Hu and Liu, 2004), as they extract sentiment in correlation to a specific topic.

Subjectivity analysis and sentiment analysis/opinion mining have been considered to be highly-related tasks. Pang and Lee (2003) state that subjectivity analysis performed prior to sentiment analysis leads to better results in the latter. Banea et al. (2010) states in this sense that *"while subjectivity classification labels text as either subjective or objective, sentiment or polarity classification adds an additional level of granularity, by further classifying subjective text as either positive, negative or neutral".*

However, according to Pang and Lee (2008): *"(...) nowadays many construe the term (sentiment analysis) more broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text."*

As we can observe, terminology employed in this field is highly variable. At times, the definitions used to denote one task or another and their related concepts are vague, inexact, overlap with definitions given for other terms, different terms are used to denote the same task or concept and the definitions are not consistent with the formal ones (that we can find, for example, in a dictionary). On top of their inconsistencies, there is also a large body of research performing emotion detection to improve sentiment analysis (Cambria et al., 2009), although no explicit relation between emotion, sentiment and opinion is presented.

Second of all, another challenge of this task is related to the retrieval of **relevant data sources**. In this context, it is important to have in mind the difference in quality between data sources, their reputation (e.g. tabloids versus trusted news agencies), the trust one may give them and their bias (e.g. if they belong to a certain public or private entity, who may have specific interests in the information presented). Bias or sentiment can be expressed by mentioning some facts while omitting others, or it can be presented through

subtle methods like sarcasm (e.g. "Google is good for Google, but terrible for content providers"). Even if some work has been done in this sense, the issues of sarcasm/irony detection and of bias detection are still far from having received a valid solution.

Thirdly, given the heterogeneity of the data sources from which the relevant information is extracted, appropriate methods have to be proposed for each type of text considered – be it newspaper articles, blogs, fora, microblogs. In the news domain, the source characteristics (reputation, bias, location) can be an important deterrent to the quality of the extracted data. The same applies to blogs, which, additionally, contain a mixture of newspaper-reporting style and free, informal comments. In fora or microblogs, the challenge is increased by the style of language involved and the characteristics of the opinion sources (i.e. of the people or entities whose opinion is expressed in that piece of text). sentiment analysis can be applied to different textual genres, at a coarser or finer-grained level and for different applications. The choice in the level of analysis normally depends on the on the type of text that one is processing and the final application – i.e. on the degree of detail that one wishes or requires in order to benefit from the process of automatic sentiment detection.

While detecting the general attitude expressed in a review on a movie suffices to take the decision to see it or not, when buying an electronics product, booking a room in a hotel or travelling to a certain destination, users weigh different arguments in favor or against, depending on the "features" they are most interested in (e.g. weight versus screen size, good location versus price).

Reviews are usually structured around comments on the product characteristics and therefore, the most straightforward task that can be defined in this context is the feature-level analysis of sentiment. The feature-level analysis is also motivated by the fact that on specific e-commerce sites, reviews contain special sections where the so-called "pros" and "cons" of the products are summarized, and where "stars" can be given – to value the quality of a characteristic of a product (e.g. on a scale from 1 to 5 "stars").

As far as the source of opinion is concerned, in this type of text, reviews are written on the same topic and by the same author. At the time of processing, thus, one is not usually interested in the author of the review, but rather on being able to extract as many opinions as possible from the reviews available.

In contrast to that, in newspaper articles, for example, sentiment can be expressed on many topics within the same piece of news, by different sources. Thus, in this kind of text, the source and the target of opinions are very important at the time of analyzing opinion. Moreover, in newspaper articles, the author might convey certain opinions, by omitting or stressing upon some aspect of the text and by thus inserting their own opinion towards the facts. Such phenomena, analyzed as part of work on perspective determination or news bias research, should also be taken into consideration at the time of performing opinion mining from this textual source. Moreover, in these texts, the news in itself is highly correlated with the opinion expressed; however, the positivity or negativity of the news content should not be mistaken for the polarity of the opinion expressed therein.

In blogs, we are facing the same difficulties – i.e. of having to determine the characteristics of the source, as well as ensure that the target of the opinions expressed is the required one. Moreover, blogs have a dialogue-like structure, and most of the times, the topic discussed is related to a news item that is taken from a newspaper article. The same phenomena are also present in forums, microblogs, social network comments and reviews, but the characteristics of these texts are different (e.g. shorter documents, different language used, single versus multiple targets of opinions, different means of referencing targets). In relation to that, there is an entire sub-area of sentiment analysis that deals with opinion holders (i.e. the source of an opinion) and opinion targets (i.e. the "object" – person, event, product, etc. – that the opinion is given on.

Fourth of all, **the retrieval of relevant, related information** is in itself a challenge. Given a specific entity, there is a need to perform additional processing (non-opinion related) in order to retrieve related entities, together with the relevant titles which can be employed to refer to them or to model the domain in which this entity may appear (e.g. politics, environment, economics, etc.). Different solutions have been given to this problem (Steinberger and Pouliquen, 2007), but the problem is far from being solved. Additionally, many Named Entities (NEs) are ambiguous (e.g. George Bush, a name that can refer to at least two different persons).

Further on, **answering the (opinion) type of questions** the task aims at is, again, far from trivial. The Text Analysis Conference[3] 2008 Opinion Pilot task and the subsequent attempts to improve the results obtained by systems performing this task have shown that the issue of "opinion" and "opinion polarity" is many times not related to the presence of explicit statements of sentiment, but to the need to infer such expressions from presentations of common-sense knowledge-related situations (e.g. "The coffee in Starbucks is yellow."). This is, again, a difficult problem in NLP. In (Balahur-Dobrescu, 2011), I described a series of experiments in opinion question answering within the TAC 2008 and the NTCIR 8 MOAT competitions and additional improvements. Based on the results presented in this work, have shown that performing traditional tasks in the context of opinionated text has many challenges and that systems that were designed to work exclusively with factual data are not able to cope with opinion questions. New methods and techniques must be designed to adapt question answering systems to deal with opinionated content.

In the case of opinion question answering systems, there is firstly a need to develop a benchmark for opinion questions' classification and the proposal of adequate methods to tackle each type of opinion queries, in a monolingual, multilingual and cross-lingual setting. Additionally, the framework for opinion question answering should be extended with appropriate resources to other languages. Further on, as we have seen from our experiments in the NTCIR 8 MOAT competition, there is an immediate need to include high-performing methods for temporal expression resolution and anaphora resolution. Unfortunately, due to the low performance of systems resolving these aspects, at this point the influence they have on the opinion question answering system's performance is negative. Another important issue in opinion question answering is the study of query expansion techniques that are appropriate for opinionated content. From what we have seen in our experiments, the use of a paraphrase collection that is not specifically designed for the sentiment-bearing textual content leads to a drop in performance of the final system.

Finally, another challenging issue to be tackled is related to the **mixture**, in texts describing events and entities, **of good and bad news**, **with** (explicitly or implicitly) **opinion** on the participating entities. Apart from the difficulty to separate the semantics of events from the polarity of sentiments expressed on entities, the "good" and "bad" are highly-dependent on the "side" from which the event that is present is "read", or, better yet, interpreted. If no user point of view is modeled, phrases such as "They sold weapons to the Israeli" are very difficult to classify as expressing a positive or negative appraisal. In this sense, in (Balahur and Steinberger, 2009), we proposed a 3 component model – author, text and reader – and a definition of sentiment analysis in dependence to the perspective (which of the components from the 3) from which the sentiment is judged.). From the *reader's point of view*, the interpretations of the text can be multiple and they depend on the personal background knowledge, culture, social class, religion etc. as far as what is normal (expected) and what is not are concerned. Lastly, the opinion stated *strictly in the text* is the one that one should concentrate on at this level, being expressed directly or indirectly, by the target, towards the source, with all the information needed to draw this conclusion on polarity present in the text. From the author's point of view, news bias or perspective determination should be concerned with discovering the

[3] http://www.nist.gov/tac/

ways in which expression of facts, word choice, omissions, debate limitations, story framing, selection and use of sources of quotes and the quote boundaries, for example, conveys a certain sentiment or not.

3. Conclusions

As can be seen, online reputation management, as a related task to sentiment analysis and opinion question answering, has to face many challenges.

Although much work has been done in opinion mining in the past years and although online reputation management can deeply benefit from the research done within the opinion mining community, much remains to be done to overcome the challenges posed by the treatment of opinionated data and its correlation to facts.

However, the increasing amount of research done in this field and its proven necessity show an optimistic outlook for applications that take advantage of the opportunities given by the constant production of opinionated data on the Social Web, in all spheres of the economic, political and social life.

4. References

Balahur, A. and Steinberger, R. (2009). Rethinking sentiment analysis in the news: from theory to practice and back. 'Workshop on Opinion Mining and Sentiment Analysis' (WOMSA), held at the 2009 CAEPIA-TTIA13th Conference of the Spanish Association for Artificial Intelligence, pp. 1-12. Sevilla, Spain.

Balahur-Dobrescu, A. (2011). Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types. Doctoral Thesis. University of Alicante, ISBN 978-84-694-8360-2.

Banea, C., Mihalcea, R. and Wiebe, J. (2010). Multilingual subjectivity: are more languages better? In Proceedings of the International Conference on Computational Linguistics (COLING 2010), p. 28-36, Beijing, China.

Cambria, E., Hussain, A., Havasi, C. and Eckl, C. (2009). Affective Space: Blending Common Sense and Affective Knowledge to Perform Emotive Reasoning. In Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA) 2009: 32-41.

Das, S. and Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Dave, K., Lawrence, S., and Pennock, D. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In Proceedings of WWW-2003, 519-528.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available resource for opinion mining. In Proceedings of the 6th International Conference on Language Resources and Evaluation, pp.417-422.

Godbole, N., Srinivasaiah, M. and Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. In Proceedings of ICSWM 2007.

Hu, M. and Liu, B. (2004). Mining Opinion Features in Customer Reviews. In Proceedings of Nineteenth National Conference on Artificial Intellgience AAAI-2004, San Jose, USA.

Kim, S.-M. and Hovy, E. (2005). Automatic detection of opinion bearing words and sentences. In Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Jeju Island, Korea.

Liu, B. (2010). Sentiment Analysis and Subjectivity. In Handbook of Natural Language Processing, eds. N.Indurkhya and F.J.Damenan, 2010.

Nasukawa, T. and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the Conference on Knowledge Capture (K-CAP), 2003: 70-77.

Pang, B. and Lee, L. (2003). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting of the ACL, 2003: 115–124.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, Vol 2, Nr. 1-2, 2008.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP-02, 2002: 79-86.

Polanyi, L. and Zaenen, A. (2004). Exploring attitude and affect in text: Theories and applications. Technical Report SS-04-07.

Scherer, K. (2005). What are emotions? and how can they

be measured? Social Science Information, 3(44), 695-729.

Steinberger, R. and Pouliquen, B. (2007). Cross-lingual Named Entity Recognition. In: Satoshi Sekine & Elisabete Ranchhod (eds.) Journal Linguisticae Investigationes, Special Issue on Named Entity Recognition and Categorisation, LI 30:1, pp. 135-162. John Benjamins Publishing Company. ISSN 0378-4169.

Tong, R.M. (2001). An operational system for detecting and tracking opinions in on-line discussion. In Proceedings of the Workshop on Operational Text Classification (OTC), 2001, New Orleans, USA.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002, 417-424, Philadelphia, USA.

Van den Bos, G. (2006). APA Dictionary of Psychology. Washington, DC: American Psychological Association.

Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the IEEE International Conference on Data Mining (ICDM). 2003: 427-434.

Zhang, W. and Skiena, S. (2010). Trading Strategies To Exploit Blog and News Sentiment. In Proceedings of ICSWM 2010. Scherer, K. (2001). Toward a dynamic theory of emotion. The component process of affective states. Cognition and Emotion, 1(1).

Position Statement: Technical Requirements For Knowledge Representation For Attitude Mining On A Realistic Scale

Fredrik Olsson, Jussi Karlgren, Magnus Sahlgren, Fredrik Espinoza, Ola Hamfors

Gavagai

Abstract

To be useful, a reputation mining system must cover a broad range of weakly, vaguely, and implicitly expressed human sentiments and cannot in the absence of prior knowledge rely on sampling the data stream of human-generated text. To achieve coverage, a reputation mining system must be sensitive to variation and change in the signal. These requirements pose a challenge which are an instance of more general semantic processing – this paper presents some design requirements used to design and implement a semantic layer for a processing stack for human-generated information.

1. It's a new kind of data and we have no choice but to cope with it

The language we see in user-generated content, such as social media, sms traffic, email conversations, etc.[1] have a number of characteristics that are normally not encountered in traditional collections of edited and published text: continuous vocabulary variation, multilinguality and code-switching, incompleteness, inconsistencies, noise and inconsistencies. Furthermore, language in streaming data has a temporal dimension not normally found in traditional corpora. These properties of streaming user-generated text data implies that we can no longer view information as something constant that can be extracted from a static knowledge repository. Instead, we need a knowledge representation that is dynamic by design and built from first principles to handle change and learn from it. We argue that a processing component deployed on a realistic scale of streaming user-generated content must be based on real-time processing of streaming data, and that the knowledge representation must be dynamic and able to seamlessly and continuously change and update its representation based on alterations in the incoming data. A traditional retrieval model of knowledge management may not be the most useful way to precede in this perspective; the interest in data streams are not necessarily based on sets of documents or mentions, but on a momentary or timely analysis of topical or attitudinal facets with respect to some topic or notion of interest, and a representation of how these change over some relevant time range.

These challenges are especially pertinent if what we are modelling are an aggregation of attitudes, opinions, and moods which tend to be less explicitly expressed in text — a system to handle implicitly formulated opinions must cover a very wide range of human expressions, many of which in themselves may seem to only have very weak signal.

2. Ethersource — our system solution

We have built a system to provide monitoring services for corporate needs for reputation management and related tasks. Ethersource is designed and implemented to constitute the Semantic Base Representation layer in the Big Data Stack, as illustrated in Figure 1.

[1] And, to extend the range of possible modalities, we may also include spoken language from telecom traffic, youtube videos, etc.

At its core, Ethersource computes and tracks relations between terms in symbols in streaming language data. These data are represented in a hyperdimensional vector space. Vector space models, the basis of many or even most information access systems today, use well established and well understood linear algebraic methods to access and manage the knowledge in them. Linguistic items such as terms or words are interpreted as points in a many-dimensional space, and similarity between terms as distances between those points. This is intuitively appealing and easy to talk about.

But vector space models are only as good as what is in them. In our case, the model is built on distributional data to build relations between terms based on their occurrence patterns. Distributional data are the basis for our semantic model - which is a solid theoretical standing point for a theory of meaning and a theory of meaning of meaning (0). Once you have a distributionally motivated model, you will be able to extract similarities between observed items in it and use those similarities to model conceptual abstraction in language. Distributional data can be aggregated in many different ways depending on what you want to find from those data. This is best done from an awareness of the basics of how language works.

Handling many-dimensional spaces poses computational challenges. Collecting data about millions of terms observed in use and the relations between them in a linear algeabric matrix may seem straightforward, but one rapidly finds that the matrix is huge and sparse. There are many many terms and many many documents (or other contexts they occur in). Even more unsettlingly we always will encounter new words: the matrix never stops growing!

There are several computational approaches to process huge matrices and to mine generalities from them such as matrix factorization techniques. Unfortunately such methods come at considerable computational cost.

The Gavagai word space model is based on a different approach in which distributional data are aggregated from observed language use incrementally, bypassing both the need for the huge matrix and the need for subsequent dimensionality reduction. Our approach is based on the practical evolution of recent techniques related to Random Indexing (0), which has several important advantages compared to other approaches: it does not require that we collect the data in

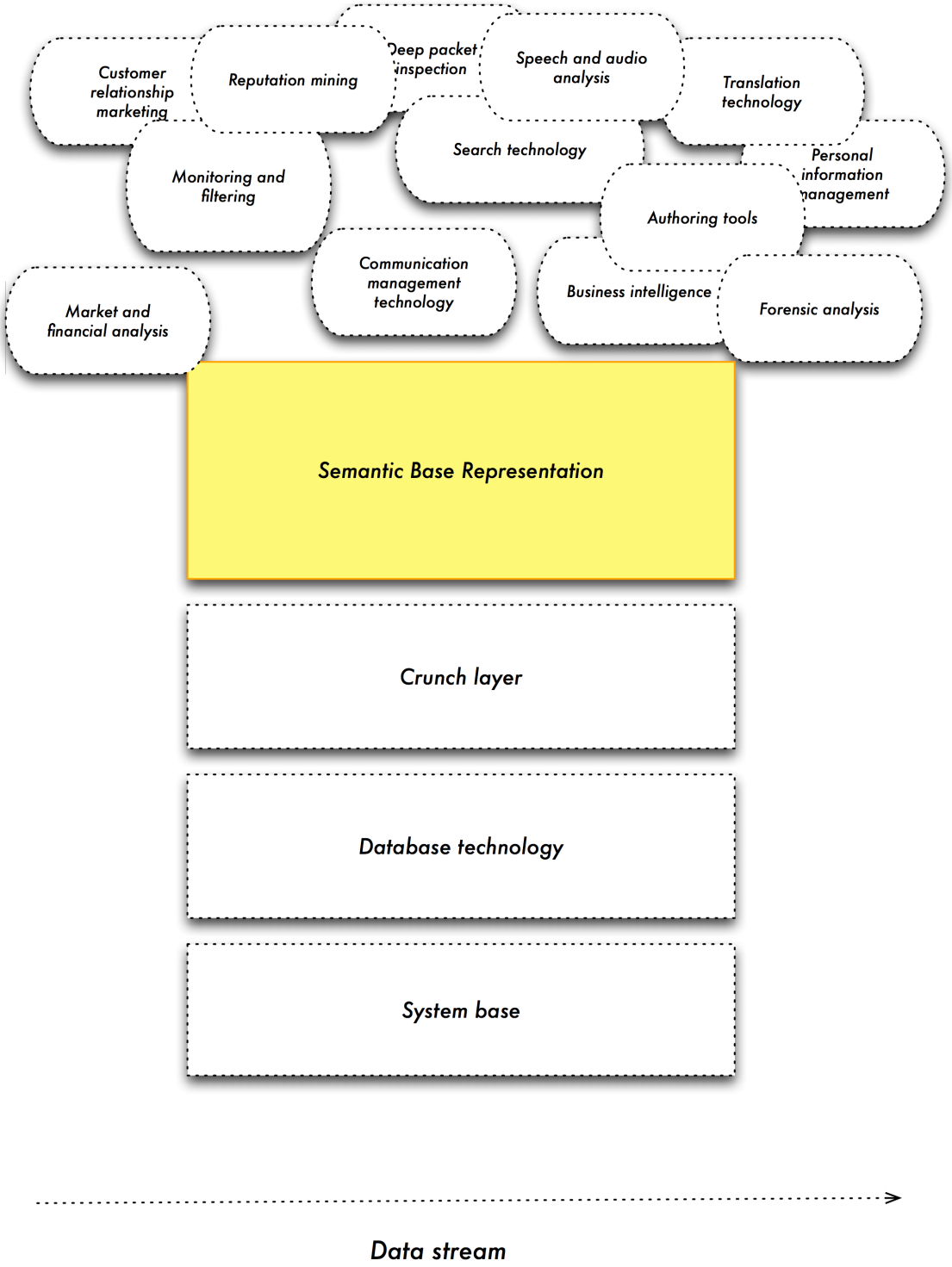

Figure 1: Ethersource is designed and implemented to constitute the Semantic Base Representation layer in the Big Data Stack.

a huge matrix and it does not require recompilation when new documents and words are encountered: the dimensionality is fixed and never increases.

The key characteristics of Ethersource include *completeness, scalability, timeliness, robustness, ability to learn*, and *multilinguality*.

2.1. Completeness, scalability, and timeliness

All sampling-based methods run the risk of missing out on crucial clues to the attitudes expressed toward a given entity. Ethersource is *complete* in the sense that it models the entire signal, that is, it is does not rely on sampling from incoming data streams.

Completeness only makes sense if the approach taken is also *scalable* to handle realistically sized data streams.

Based on neurophysiologically plausible models of information processing, Ethersource uses a fixed-size memory model whose size remains constant with growth of data. For Ethersource, the memory model and the processing model are identical.

Figure 2 serves to illustrate the orders of magnitude in difference between two commonly used memory models, (word-by-word, and word-by-document matrices), and the Ethersource model. All three models are designed to relate words to words based on their distribution. In this particular example, more than 11 000 000 Tweets concerning the state of the world, especially the Middle East and Northern Africa, were collected during the period of February 8 to 12, 2011. When building the three memory representations from the data, it turns out that the word-by-word memory model requires 190 times the number of matrix cells used by Ethersource. At the same number of Tweets, a word-by-document matrix would be 5 500 times larger than the representation used by Ethersource. The memory model employed by Ethersource grows sublinearly with the size of the input text stream.

To fully draw on the temporal qualities of the attitudes expressed toward a given target are maintained, the system has to ensure low latency in all its parts, that is, it should deliver actionable intelligence in a *timely* manner. Ethersource is, by virtue of its theoretical underpinnings, designed to allow for high throughput. As an example of timeliness, Figure 3 relates the official times at which the two artists Danny Saucedo and Loreen entered stage during the Swedish final of the Eurovision Song Contest 2012 (red annotations in the Figure), to the activity in Swedish social media, as measured with Ethersource. Note the short time from, e.g., Loreen entering the stage for the first time, and the corresponding outburst in (primarily) Tweets relating to that event. The time from the publication of a given Tweet on Twitter, until it is analyzed by Ethersource is typically less than a minute.

2.2. Robustness and learning

Robustness is a way of saying that a system does not choke if it encounters unexpected input. Ethersource is built on the presumption that "language is in order as it is", and is designed to cope - and thrive - with variability, noise and inconsistencies. Language is in a constant state of flux, and so is Ethersource.

Not only need a system be robust in the sense mentioned above, it should also be able to learn from the ever changing input. Ethersource is inherently and constantly learning, and is thus well equipped to pick up on language variations, misspellings, neologisms, etc, and turn such variations to a competitive advantage. In an unsupervised fashion, Ethersource continuously updates its knowledge representation as new data is encountered. The knowledge representation, in turn, is instantaneously accessible for queries about its current state, without having to resort to a update-retrain-redeploy cycle. With respect to learning, Ethersource does not rely on external language resources, or on human intervention.

As an example of coping with, and learning from linguistic variation, consider the following scenario. You are assigned with the task of monitoring the on-line mentions of the American football player Tim Tebow in English on-line social media. Now, the first question is what terms are suitable for looking for Tebow. When supplying Ethersource with the most obvious one, i.e., *Tebow*, it returns a number of not-so-obvious terms it has learned from the data that also refers to Tim Tebow and thus should be included in the target specification: *Twbow, Tibow, Tebox, Teboq, Tewbow, Teobow, Teabow, Teblow, Tebowm*

Furthermore, as a part of assessing the current state of his on-line presence, it may be useful to know the concepts associated with Tebow at any given point in time. Ethersource learns, and thus allows for the identification of concepts associated with the target tracked. In the case of Tim Tebow, the most prominent concepts associated with him, on the particular day we are looking at, include: *Broncos, Tim, Denver, quarterback, Tebowing, Tebowed*[2]

2.3. Multilinguality

In a realistic situation, all the above characteristics are required for multiple languages. Ethersource is inherently language agnostic in that it is designed to model what is common to all languages, rather than what makes the different from each other. The statistical regularities Ethersource exploits are consistent with current linguistic theory, and are sensitive to the generalities of natural language. Ethersource currently performs targeted processing on a range of typologically diverse languages, including English, Swedish, Chinese, Arabic, Russian, and Hindi. New languages can be added with minimal effort and without changing the system.

3. Conclusion

Ethersource is an implementation of a general purpose semantic model fulfilling the above technical requirements. We use Ethersource for monitoring attitudes in on-line media where — due to the nature of the task — the requirements of completeness, scalability, and timeliness on the one hand and the requirements of robustness and learning meet and come to the fore. We believe that these requirements are central to many tasks where situation awareness is crucial and we believe that we will find that the necessity for a general purpose semantic model to work with human language will be found to be necessary for numerous tasks to come.

4. References

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009.

Magnus Sahlgren. The distributional hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics)*, 20(1):33–53, 2008.

[2]Initially, the final two terms in the list puzzled us a bit. This is what we learned. *Tebowing* refers to the act of getting down on one knee and starting to pray, even if everyone around you is doing something completely different. *Tebowed*, on the other hand, has little to do with spirits as it denotes being run over while playing American football.

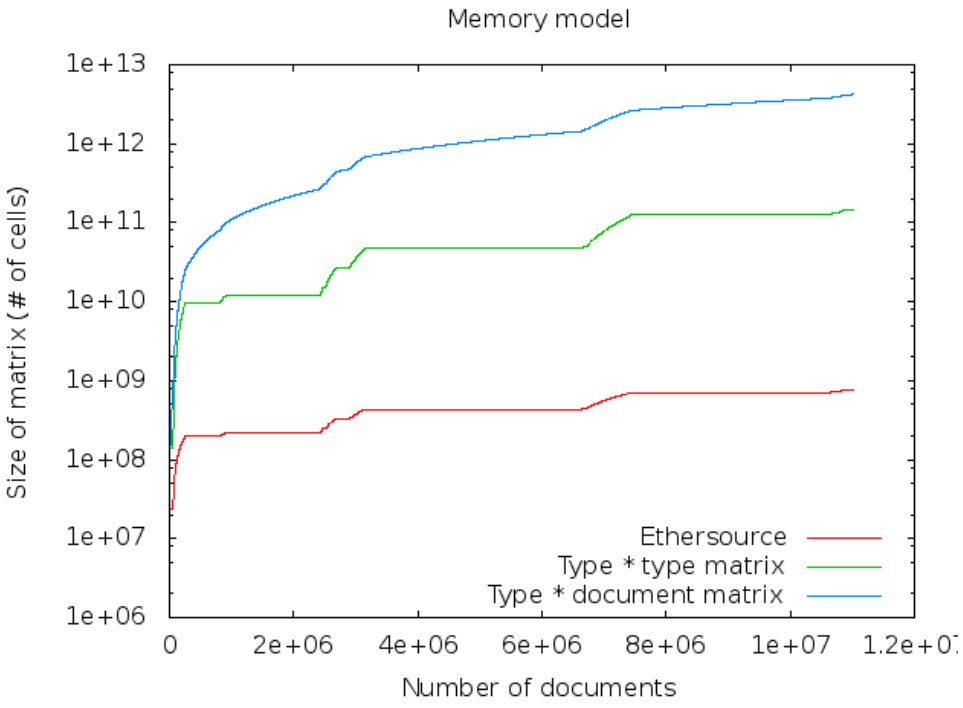

Figure 2: A theoretical comparison, in size, between three different ways of representing the same contents: a vanilla word-by-document matrix (blue), word-by-word matrix (green), and Ethersource (red).

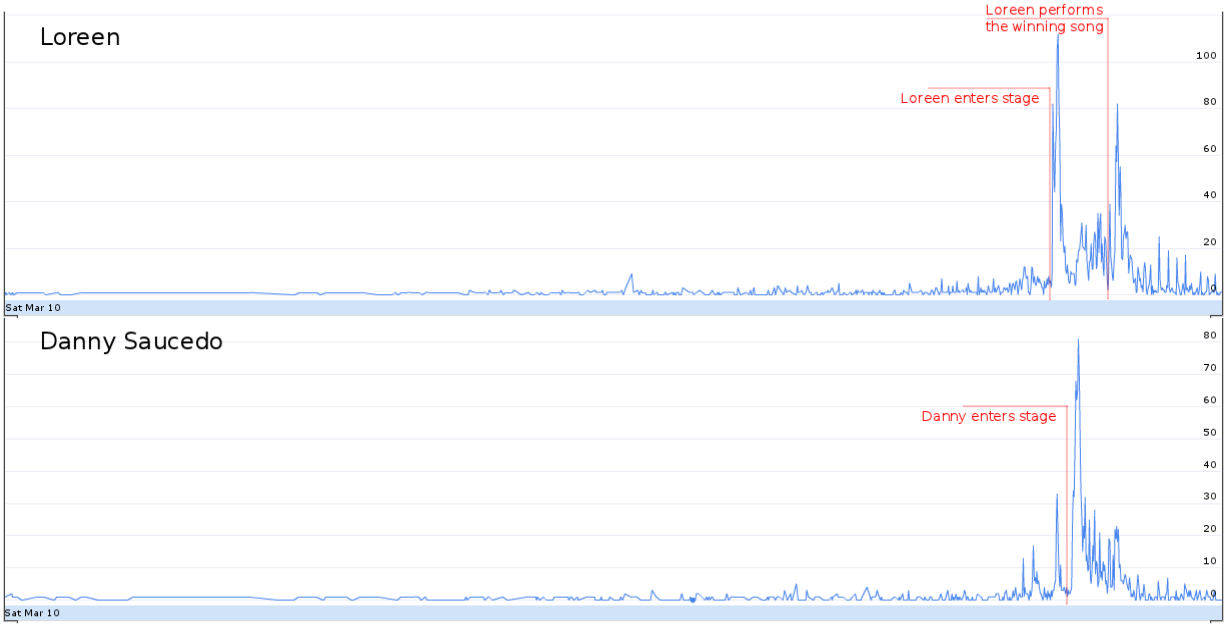

Figure 3: Illustrating the low latency of Ethersource. The popularity of the artists Loreen and Danny Saucedo, measured minute-by-minute during the day of the final of the Swedish part of the Eurovision Song Contest. The annotations in red denotes the appearance on stage by the two artists: Note the short delay between artist appearance and increase in data.

Uses of Computational Stylometry to Determine Demographics for Online Reputation Management

Patrick Brennan
Juola & Associates
E-mail: pbrennan@juolaassoc.com

Abstract

Computational stylometry can be used as a tool to gather better demographic data for the purposes of online reputation management. Computational stylometry is the study of linguistically style; in this case, applied to blog posts and comments on web sites. These sorts of messages are generally both anonymous and honest appraisals of products and services, so being able to gather more data about who these comments represent will provide businesses with a better idea of how they are doing with these demographic groups. Finally, we illustrate a product currently in the works that will tag comments with demographic tags through the use of computational stylometry.

Keywords: computational stylometry, authorship verification, online reputation management

1. The Problem

One of the core problems facing online reputation management is the inability of individuals and organizations to empirically identify the demographics that they are or are not servicing well. Put another way, companies have to take it on faith that the demographics that are giving them positive or negative feedback is, in fact, the demographic they say they are.

Anonymity has been a staple on the Internet since its inception. The fact of the matter is that it is actually easier to be anonymous on the Internet than it is to be public on the Internet. Furthermore, anonymity provides something of a security blanket to those who use the Internet; one that not many are willing to give up.

2. Additional Problems

A secondary problem caused by the abundance of anonymity is that of defamation and astroturfing. Given that there is no reliable way to ensure somebody on the Internet is really who they claim to be, businesses face the dual threats of having their own products negatively reviewed or a competitors products positively reviewed by a malicious third party.

Finally, we face the problem of volume. Social media data has skyrocketed over the past several years. It has moved past the point of being able to be reviewed by a human and must be reviewed by a computer.

3. Proof the Problems Exist

The case of the "Three Wolf Moon" shirt on Amazon is a classic example of when astroturfing occurs from a mischievous third party. A relatively unremarkable shirt displaying the graphic of three wolves and a moon was catapulted to being one of the top selling Amazon Clothing item because Internet jokesters gave the shirt five starts and wrote thousands of outlandish reviews for it on a lark.

At the end of the day, everything worked out well for The Mountain Company (the shirt-maker). It helped spur some Internet fame for the otherwise innocuous company that has led to an increase in sales. That being said, this sort of positive outcome is extremely unusual; and companies must remain diligent to ensure that their brands are not tarnished.

4. Generic Solution

As can be clearly seen, the twin problems of data volume and anonymity require there to be a large-scale automated solution. This solution needs to do two things:

1. Categorize and interpret these posts in a meaningful way so that analysts can spend their time following up posts that matter instead of ones that do not.
2. Assign demographic or personal data to comments and blog posts that are accurate.

5. Stylometry

Our technology answers these two problems in a lightweight, elegant, and novel manner. Our technology is based off of stylometry, the science of writing style. Just as every person has their own fingerprint and DNA, every person has their own writing style that can be used to identify them with confidence. Similarly, a person's writing style can speak volumes about their nationality, native language, age, education, social class, gender, and so forth. By analysing comments and blog posts, we can determine many of the characteristics of the person who wrote it.

One simplified example of stylometry in action is looking at the difference in word choice between "color" and "colour". If the word "color" is used, then the author is most likely from the United States. If the term "colour" is used, then the author is most likely from the United Kingdom, Canada, or Australia.

Stylometry has been used to identify gender for years, as seen by the Gender Genie. The Gender Genie is based off of work done as early as 2003 by Illinois Institute of Technology and Bar-Ilan University of Israel.

And while the Gender Genie is little more than an Internet toy, it is a stark demonstration of what stylometry can do.

Another example of stylometry being used to identify demographic traits comes out of the Evaluating Variations in Language Laboratory at Duquesne University. John Noecker Jr. and Michael Ryan were able to detect Myers-Briggs personality types through the use of stylometry. While personality type may not be all that useful for online reputation management it does show that stylometry can be used to identify certain "intangible" qualities that consumers have. After all, if something as ephemeral as personality type can be quantified and discovered in writing samples; what about preference for certain consumer products?

6. Our Solution

The system we are proposing would be a piece of middleware that will take unfiltered comments or blog posts and break them down along demographic lines through the use of stylometric classification technology. These posts will be tagged with essential metadata such as age, gender, nation of origin, educational background.

Our product would be integrated into the client's already existing infrastructure. The system will be connected to a website backend to scan comments as they come in for demographic data. If the user doesn't have direct access to data stream that they wish to scan, then external modules can be developed to facilitate the collection, cleaning and separating of posts for scanning.

Once a post is brought into the system, it goes into our language tagging system, which determines what language the comment is in. This step is important, as each of our other tagging modules will be language specific.

From here, the post is fed through a number of tagging modules that compares the post with the style associated with a particular group or person. These tagging modules could include but are not limited to: identifying male (or female) posters, identifying posts by people with college degrees, identifying adolescent posters, etc. Once the post proceeds through the tagging system its entry in the original data warehouse is updated with the accumulated tags.

From there, analysts and other programs can use the updated tagging information to very quickly categorize and process the comments. This provides analysts with a quick solution to both problems --- not only can they apply real data to user names and comments, they can do it quickly and reliably.

7. How It Will Work

The core technology that will drive our system is a technique of distractorless author verification developed by the Evaluating Variation in Language Lab at Duquesne University. While the technology was originally developed to determine authorship of a document; we will be using it to look for demographic traits in individuals.

Using this technique, we can quickly build a model of what an "average" individual who expresses a certain demographic trait looks like. The model is built by introducing carefully selected samples of writing from individuals why fit the demographic criteria we are looking for while making sure to hold the other demographic criteria constant.

After the models are build they are used by the system as a yardstick to determine how closely the post being checked resembles the post of an "average" person who represents the demographic in question. This comparison is represented by a percentage match and depending on how high this percentage is determines if the post is labelled or not. This cut-off percentage can be set to whatever level the user feels comfortable with.

Furthermore, the modules themselves evolve through use. The tagging model is developed using state-of-the-art machine learning techniques based on examples. A user has an option to confirm a tag on a post as long as he is sure that the tag is correct. A confirmed tag, in turn, can be used to "teach" the module a better categorization. Through the use of this human verification system, the modules will become more effective. Furthermore, using this system, users can even generate their own tagging modules. This works by submitting a number of posts that are known to have one point of commonality; for example, posts that were all written by a certain income bracket. From here, the system can dynamically build a preliminary tagging module that looks for additional posts that fit that criterion

8. Conclusion

In this paper, I hope we have shown just how useful stylometry is in the field of online reputation management as a means to determine better demographic data. Furthermore, we have illustrated one system that leverages stylometry effectively to provide that data in a lightweight, efficient, and scalable manner. That being said, the field is growing every day and more powerful and effective techniques will be developed as time goes on.

Technical Papers

Topic Identification from Blog Documents: Roles of Bigram, Named Entity and Sentiment

Chandra Mohan Dasari, Dipankar Das, Sivaji Bandyopadhyay
Computer Science and Engineering Department, Jadavpur University
Kolkata-700032, India
chanduthecm@gmail.com, dipankar.dipnil2005@gmail.com, sivaji_cse_ju@yahoo.com

Abstract

The rapid growth of blog documents in Web 2.0 and categorizing search applications based on topics motivates us to develop a system that identifies topic names of the blog documents using Bigrams, Named Entity (NE) and Sentiment features. We also associate the sentiment scores to the blog documents using the *SentiWordNet*. The individual module based on Bigrams, NE and Sentiment produces the topic bag for each blog document containing probable topic names of that blog. The probable topic names were evaluated manually based on top-n (n = 5, 10 and 20) ranking mechanism. Though the combined module of Bigram and Sentiment performs better than the combined module of Bigram and NE, the combination of all the three modules produces satisfactory results on evaluating 125 topic names with respect to 25 test documents on 5 different topics. The evaluation achieves the maximum accuracies of 60.0%, 72.0% and 84.0% for the combined module of Bigram and Sentiment and 76.0%, 86.0% and 92.0% for the combined module of Bigram and Named Entity with respect to top-5, top-10 and top-20 ranking mechanisms, respectively.

1. Introduction

Emails, weblogs, chat rooms, online forums and even twitter are being considered as the social media for discussing recent topics. Blog is the most important, communicative and informative repository of text based contents in the Web 2.0 (Yang et al., 2007). Many blogs act as online diaries of the bloggers for reporting the blogger's daily activities and surroundings. Sometimes, the blog posts are annotated by other bloggers. Therefore, blogs are being considered as one of the personal journals where people express their personal opinions on different aspects like products, travelled tourism places, politics, and current happenings in society. Especially, the blog posts contain instant views, updated views or influenced views regarding single or multiple topics.

Topic is the most frequently used, unexplained, term in the discourse analysis literature (Brown and Yule, 1983) or is the real world object, event or an abstract entity. Topic identification is one very important text processing technique that can help people scan through huge volume of texts, classify them into different categories, route them into relevant parties and summarize them (Lin, 1997). Topic identification is a central step for many automatic text processing tasks. Most of the related work uses topic identification as part of a specific task, such as automatic document indexing, text classification, text categorization, text summarization and information retrieval.

With exponentially increasing amounts of text being generated, it is important to find methods that can annotate and organize documents in meaningful ways. Thus, topic identification is also used for document ranking in Informational Retrieval systems. In addition to the content of the document itself, other relevant information about a document such as related topics can often enable a faster and more effective search or classification.

Topic identification is also essential in connection within categorizing search applications (Stein and Eissen, 2004). Categorizing search has attracted much interest recently; its potential has been realized by users and search engine developers in the same way. Categorizing search means to apply text categorization facilities to retrieval tasks where a large number of documents are returned. Consider for example the use of Internet search engines like Google or Lycos: Given a query they deliver a bulky result list *D* of documents. Categorizing search means to return *D* as a set of priori unknown categories such that thematically similar documents are grouped together.

Some of the applications of topic Identification are also used in Artificial Intelligence (AI) search. At the moment, being in the age before the Semantic Web, clustering technology has achieved considerable success in mastering this ad-hoc category formation task. AIsearch is a categorizing Meta search engine, which is developed in the institute (Stein and Meyer, 2002).

In addition to the above issues, topic of blogs is important as the recent trend of the people is to express their opinions on various situations, events and topics or discuss several important social issues on the blogs. The blog is becoming a rich source of various opinions in the form of product reviews, travel advice, social issue discussions, consumer complaints, movie review, stock market predictions, real estate market predictions, etc. The content in the blogs also contain names of important people, places, and organizations.

Thus, in the present task, we have developed a system that identifies the topics from the blog documents using Bigrams, Named Entity and Sentiment features into account. Additionally, we have proposed a top-n ranking mechanism to evaluate the performance of our topic identification system. The ranking of the topics also helps to rank the blog documents. Blogs are very wide term and could be related with professional blogs of authors, conversational and discussion blogs or twitter-like reactions to various events, etc. But, in the present task, we have collected random samples of the blog documents without considering the specific syntactic and semantic

properties of the blog documents. The stop words, symbols were filtered from the blog documents to prepare a clean corpus. We have developed a topic identification system based on bigrams, NE and sentiment. We have used the Stanford Named Entity Recognizer [1] for identifying the NEs. We also find the sentiment scores for the blog documents based on the lexicon based approach. The SentiWordNet[2] and WordNet Affect[3] lists were used for identifying sentiment words and scores for the documents. We have evaluated the system on a test set of 25 blog documents on five latest hot topics (2G Scam in India, Bombay blasts, Common Wealth Games Scam, Separation of Telangana state in Andhra Pradesh and Cricket World Cup 2011). The system combining the modules of Bigram and Sentiment achieves the maximum accuracies of 60.0%, 72.0% and 84.0% whereas 76.0%, 86.0% and 92.0% accuracies have been obtained for the combined module of Bigram and Named Entity with respect to top-5, top-10 and top-20 topic names, respectively. The top-n evaluation technique was attempted to evaluate the system identified topic names against the manually defined topic names for each of the documents.

The rest of the paper is organized as follows. Section 2 describes related work. Preparation of clean blog corpus has been described in section 3. The description of topic identification from blog using different approaches is specified in section 4. Section 5 describes the procedure for finding sentiment scores for the blogs. The evaluation mechanisms and associated results are discussed in section 6. Finally section 7 concludes the paper.

2. Related work

Several research efforts have been conducted for topic identification. One of the important tasks that proposed various insights and solutions related to the topic identification was described in the dissertation (Lin, 1997). A corpus-based language model for topic Identification was also proposed in (Chen, 1995). They analyze the association of noun-noun and noun-verb pairs in LOB corpus.

A system was developed by (Kim and Hovy, 2004) that automatically finds the people who hold opinions about a given topic and the sentiment of each opinion. The system contains a module for determining word sentiment and another for combining sentiments within a sentence. (Chesley et al., 2006) presents experiments on subjectivity and polarity classifications of topic and genre independent blog posts, making novel use of a linguistic feature, verb class information, and of an online resource, the Wikipedia dictionary, for determining the polarity of adjectives. Each post from a blog is classified as *objective*, *positive*, or *negative*. On the other hand, a system that assigns scores indicating positive or negative opinion to each distinct entity in the text corpus was proposed in (Godbole et al., 2007). The emotion classification of web blog corpora using support vector machine (SVM) and conditional random field (CRF) machine learning techniques has been investigated in (Yang et al., 2007).

A method for automatic topic identification using an encyclopedic graph derived from Wikipedia was discussed in (Coursey et al., 2009). The system is found to exceed the performance of previously proposed machine learning algorithms for topic identification, with an annotation consistency comparable to human annotations. The problem of Named Entity Recognition in Query (NERQ) and classification of the named entity into predefined classes is potentially useful in many applications in web search and the whole discussion has been explained in (Guo et al., 2009).

There are several statistical methods for topic identification, such as topic datagram, TFIDF and others such as cache or weighted unigrams. All these are based on certain metrics. In the present task, we have employed three modules (Bigram Count, Named Entity Recognization and Sentiment Word Tagging) to identify the topic of the blog documents.

3. Preparation of Blog corpus

We have randomly collected a total of 46 blog documents from the web (during the session, 2010-2011) with a total of 20470 sentences. To prepare the test set, we selected 25 blog documents on five recent topics (2G Scam in India, Bombay blasts, Common Wealth Games Scam, Separation of Telangana state in Andhra Pradesh and Cricket World Cup 2011) from different blog sites. The statistics of the blog corpus is given in the Table 1.

Total number of documents in the corpus	46
Total number of Training documents	21
Total number of Test documents	25
Total number of sentences in the corpus	20470
Average number of sentences in a document	445
Total number of words in the corpus	290854
Average number of words in a document	6320

Table 1. Statistics of the whole Blog Corpus.

We have collected the source code (in HTML or XML format) of the blog documents and retrieved texts from the source code. The Example source code is shown in Figure 1. We removed stop words and special symbols from the corpus using a stop word list that contains 320 stop words. The stop words are small *"function words"* such as *the, is, at, which, on* etc. These words cause problems while searching for phrases or words including a few "*lexical words"*, such as *want.*

4. Topic Identification System

The present topic identification system consists of three different interconnected modules (Bigram Count, Named Entity Recognization and Sentiment Word Tagging). The system architecture is shown in Figure 2. The details of the individual module are as follows.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

[2] http://www.sentiwordnet.isti.cnr.it/

[3] http://www.cse.unt.edu/~rada/affectivetext/

```
<div class="commentBD description">
        <p>I guess, JPC is being pushed to ensure that
            scam booty is shared across all political
            parties. Will be great if there can be a measure
             to ensure money lost  come back to
             exchequer in the form of either
             penalties or cancelling the spectrum
             allocation and  punishment of guilties.
        </p>
    <div class="clear"></div>
</div>
```

Figure 1: The source code of the web blog data

4.1 Bigram Count Module

In the fields of computational linguistics and probability, an *n*-gram is a contiguous sequence of *n* items from a given sequence of text or speech (Jurafsky and Martin, 2009). An *n*-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram" and so on. We calculated unigram counts and then retrieved the unigrams for the blog documents with respect to five different topics. Primarily, it has been observed that the unigrams fail to produce complete topic names. For example, the top-5 relevant unigrams for the topic "2G Scam" are *'public',' money',' people', 'like',' scam'*.

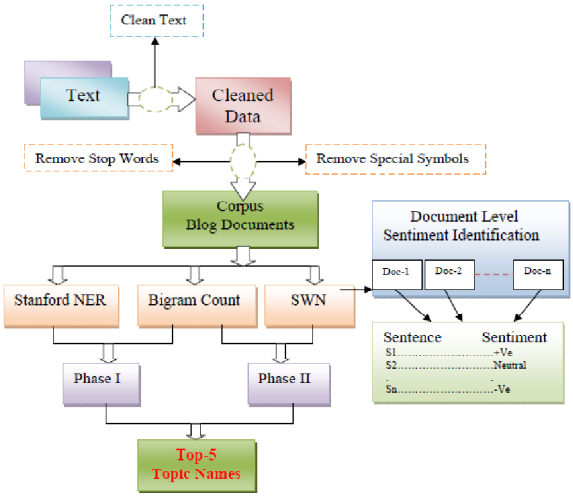

Figure 2: The System Diagram

Thus, to improve the performance of the topic identification system, we moved for Bigram count approach. Bigram counts maintain the same principle as monogram counts, but instead of counting occurrences of single words, bigram counts count the frequency of *pairs* of words. We calculated bigram word frequency and tagged these bigrams in the input file and retrieved top-5 bigrams based on the frequency count for the blog documents. The bigrams are also shown in Table 2. But, it is important to mention that the trigram count adds extra noise in the identified topic names.

4.2 Named Entity Based Approach

Named Entity Recognition (NER) is the task of processing text to identify and classify names. The NER enables the extraction of useful information from documents and is often performed using a statistical tagger which learns patterns for the recognition of names from manually-annotated textual corpora. (i.e. Person names (names of people), Organization names (companies, government organizations, committees etc.), Location names (cities, countries etc.), Miscellaneous names (Date, time, number, percentage, monetary expressions, number expressions, measurement expressions).

Topic Name	Bigram Approach	NER Approach	Sentiment Approach
2G scam	public money public participation black money 2g scam fight corruption	Priyanka Scam Nehru-Gandhi Ketan-Parekh Ram Jethmalani	jpc public money people scam
CWG	public money games village traffic jams hard earned spent cwg	India CWG Kalmadi Cayman Islands	games money india corruption cwg
Telangana	telangana movement news telangana raj news telangana latest Andhra Pradesh	Australia committee srikrishna Canberra Members	telangana news government movement channel

Table 2: top-5 Topic Names based on three approaches

We have used the Stanford Named Entity Recognizer for identifying named entities that are part of our topic identification task. The Stanford NER (also known as CRFClassifier) is a Java implementation of a Named Entity Recognizer that labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names.

For example, if the Stanford NER input is *"One man army Baba Ramdev is fighting against corruption"*, the corresponding output is *"One/O man/O army/O Baba/PERSON Ramdev/PERSON is/O fighting/O against/O corruption/O"*. We retrieve all the Named Entities from the blog documents using the Stanford NER and calculate the frequencies of all those words in the documents.

In addition to the individual system, we also tag the Named Entities in the blog documents where the Bigrams were also tagged. We retrieve the top-5 Bigram words that were also tagged by the Named Entity module. The combined module of Bigram and NE produces the topic names as shown in Table 4.

4.3 Sentiment Word Tagging Module

In the present work, we have used the *SentiWordNet-ver* 3.0.0, an enhanced lexical resource explicitly devised for supporting sentiment classification, emotional analysis and opinion mining applications. Each synset of the *SentiWordNet* (SWN) is associated to three numerical

scores *Pos*(s), *Neg*(s), and *Obj*(s) which indicate how *positive*, *negative*, and "*objective*" (i.e., *neutral*) the terms contained in the synset are.

We find the sentiment sentences if any word of the sentence appears in the *SentiWordNet* and retrieve only sentiment sentences from the blog documents. We have calculated the total number of *positive* and *negative* words in the test blog corpus. The statistics are shown in Table 3. We tag the sentiment words in the blog document in which Bigrams were also tagged. We retrieve the top-5 Bigram words commonly tagged by Bigrams and Sentiment modules both. The combined module produces the topic names as shown in Table 4.

Sentiment Statistics of the Blog Corpus	
Total number of Sentiment Sentences in the corpus	6918
Total number of Sentiment words in the corpus	35712
Total number of *Positive* sentiment words	20896
Total number of *Negative* sentiment words	14816

Table 3. Sentiment statistics of the blog corpus

Topic Names	**Bigram + NER**	**Bigram+ Sentiment**
2G scam	2g scam ketan parekh mr kalra baba ramdev	waste public dont think 2g scam fight corruption day day
CWG	public money spent cwg mr kalra delhi govt people india	hard earned common man closing ceremony opening ceremony completely agree
Telangana	news telangana telangana people according telangana state telangana telangana telangana	telangana movement raj news telangana latest latest news telangana news

Table 4: top-5 Topic Names based on two combined modules (Bigram + NER and Bigram + Sentiment).

5. Identification of Sentiment of Blogs

The topic-document model of information retrieval has been studied for a long time and systems are available publicly since last decade. On the contrary Opinion Mining/Sentiment Analysis is still an unsolved research problem. Although a few systems like Twitter Sentiment Analysis Tool, Tweet Feel are available in World Wide Web since last few years still more research efforts are necessary to match the user satisfaction level and social need. Blogs also express opinion of entities (people, places, things) while reporting on recent events.

Thus, in addition to identify the topics of the blog documents, we identify the sentiment of the documents based on phrases as well as sentences. Sentences can be considered as the basic information units of any document. For that reason, the overall document level sentiment identification process depends on the sentiment expressed by the individual sentences of that document which in turn is based on the sentiment expressed by the individual words or phrases (Das and Bandyopadhyay, 2009).

We tagged all sentiment words of the blog documents along with *positive* and *negative* scores extracted from the *SentiWordNet* (SWN). The sentences that contain sentiment words have been retrieved and the *positive* and *negative* scores of the sentences are calculated based on the sentiment words. One example sentence is as follows,

<Great>, 0.75,0 blog <pity>, 0, 0.75 happening let's <hope>, 0.25,0 <good>, 0.625,0 outcome →Positive Score=1.625 and Negative Score=0.75 POSITIVE STATEMENT

We have also calculated the number of *positive* and *negative* words in the document. Finally, we calculate the total *positive* and *negative* scores in the document and find the document level sentiment based on the maximum scores. The example of a document level sentiment is as follows.

Document Name: Telangana Blog
Total Positive Score=4547.75 Total Negative Score = 4242.375 POSITIVE DOCUMENT

6. Evaluation

The evaluation of the topic identification system has been conducted on 125 topic names for 25 blog documents containing a total of 10470 sentences. The blog documents have been collected with respect to five different topics such as CWG, 2G scam, Separate Telangana, CWC 2011 and Delhi Blasts as shown in Table 5. We have proposed two-way evaluation technique for measuring the performance of the system. In the first method, we have compared the system identified topic names against only one manually assigned topic name whereas the second method considers the evaluation with respect to five topic names suggested manually by the authors. Both of the methods use a top-n evaluation technique for measuring the performance of our topic identification system. The value n indicates the number of the system identified topic names for each of the documents. In the present task, we have classified the system identified topic names in only three categories, namely top-5, top-10 and top-20. We extracted the top-5, top-10, and top-20 topic names from each of the 25 blog documents.

	CWG (%)	**2G Scam (%)**	**Telangana (%)**	**Delhi Blasts (%)**	**CWC 2011 (%)**
Bigram Count	40	35	74	36	74
NER	48	54	42	45	16
Sentiment	42	54	73	42	50
Bigram+ NER	54	66	74	46	54
Bigram+ Sentiment	50	72	42	42	74

Table 5: Average Scores for each blog topics.

The accuracies for each of the blog topics are shown in graphical representation in Figure 3. In X-axis, the graph represents 5 different approaches for identifying the topic names and Y-axis represents the accuracy values for each of the approaches.

In the second method, we have manually assigned 5 topic names (t_{111}, t_{112}, t_{113}, t_{114}, and t_{115}) for each of the 5 documents (D_{11}, D_{12}, D_{13}, D_{14}, and D_{15}) with respect to each of the five latest blog topics (T_1, T_2, T_3, T_4, and T_5). Thus, a total of 125 topic names (t_{111}, t_{112} …….t_{535}……..t_{554}, t_{555}) have been considered for evaluation as shown in Figure 4.

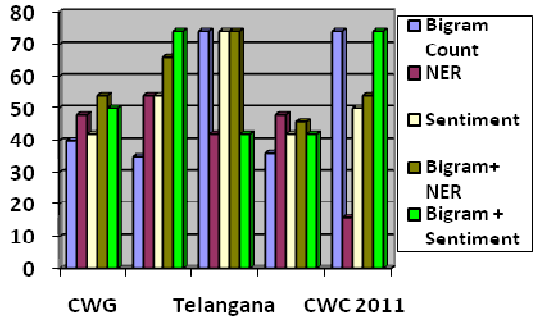

Figure 3. Percent of scores in different approaches

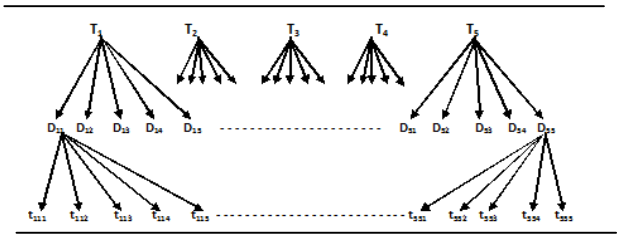

Figure 4. Manually Evaluated Topic names for 5 documents.

We have checked the manually assigned topic names with our system identified topic names by dividing in three different groups such as top5, top10 and top20 respectively. We count the number of system identified topic names matched with manually identified topic names. For example, if we consider ***m*** is the number of manually evaluated topic names for each of the documents and ***d*** is the number of documents for evaluation and $\mathbf{x_1, x_2 \ldots\ldots x_d}$ are the number of system identified topic names that match with top-***n*** manually identified topic names. Thus the accuracy is calculated as follows,

$$[((x_1+x_2+\ldots\ldots+x_d)/(m*n))*100]$$

For example, if we consider top-5 evaluation technique where 2, 4, 0, 3, and 5 are the number of system identified topic names matched with manually identified 5 topic names with respect to each of the 5 documents respectively, the accuracy is calculated as **(2 + 4+ 0+ 3+5 /5*5)*100 = 42 %.** We calculate the accuracies for top-10, and top-20 evaluation in the similar way. The accuracies for the training and test sets are shown in Table 6. From Table 6, the average accuracies for the top-5 Bigram counts for all of the blog topics are 39.20% for training set and 50.4% for test set respectively.

It has been observed that the Bigram count along with the Sentiment feature gives the highest accuracies in comparison with the system that identifies the topic names using the Named Entities only. The reason is that the Named Entity Recognizer identifies only person, location and organization names but fails to detect the temporal information and multi-word components that give the clues regarding the topic names. By using the sentiment as a feature along with the Bigram and Named Entities, the present system performs satisfactorily to produce better results in topic identification.

	Top- n Topic (s)	**CWG Train [Test] (%)**	**2G Scam Train [Test] (%)**	**Telangana Train [Test] (%)**	**Delhi Blast Train [Test] (%)**	**CWC 2011 Train [Test] (%)**
BC	n=5	32 [36]	40 [52]	44 [64]	36 [48]	44 [52]
	n=10	44 [40]	72 [64]	68 [76]	52 [60]	48 [56]
	n=20	68 [76]	80 [72]	92 [84]	52 [56]	48 [64]
NER	n=5	28 [40]	28 [36]	32 [32]	44 [52]	36 [52]
	n=10	32 [44]	48 [56]	48 [52]	56 [60]	44 [52]
	n=20	64 [52]	52 [56]	48 [60]	64 [68]	52 [60]
SNTI	n=5	52 [44]	32 [40]	36 [40]	28 [36]	28 [36]
	n=10	56 [60]	52 [48]	44 [48]	36 [44]	36 [40]
	n=20	64 [72]	64 [68]	52 [52]	52 [60]	44 [56]
BC + NER	n=5	44 [44]	32 [40]	76 [68]	36 [40]	36 [44]
	n=10	56 [60]	44 [52]	84 [80]	44 [52]	48 [48]
	n=20	76 [72]	48 [52]	92 [88]	64 [76]	52 [64]
BC+ SNTI	n=5	44 [40]	44 [60]	36 [40]	32 [40]	40 [52]
	n=10	64 [60]	56 [68]	72 [64]	36 [52]	52 [72]
	n=20	72 [84]	72 [78]	76 [68]	48 [52]	60 [72]

Table 6. Training and Test set accuracy values for Top-5, 10, 20 Topics.

It has been found that in some cases, the present system fails to identify some of the topic names. For example, the system based on Bigram and NER fails to detect the topic "*2G scam*" in top-5 evaluation technique while Bigram and Sentiment based system identifies the topic names *"2g scam", "fight corruption" and "waste public"* in top-5 evaluation technique.

On the other hand, the topic "*Common Wealth Games*" has been identified by using the Bigram count and NER but not identified using Bigram count and Sentiment feature. The reason may be the topic "*2G scam*" is mostly related to sentiment whereas *Common Wealth Games* is more related to Named Entity rather than sentiment. Sometimes, our system produces some irrelevant topic names using Bigram and Sentiment. For example, the system based on Bigram count and sentiment identifies the topic names like "*day day", "way way*" instead of the "2G scam".

7. Conclusion

In the present task, we have collected the blog corpus on recent topics and developed a prototype system for evaluating the performance of identifying topic names.

We have incorporated some simple features like Bigram, Named Entity and Sentiment Words to identify the topic names from the blog documents. The system identifies the topic names satisfactorily. Our future task is to improve the performance of the system by identifying topics both at sentence and document levels and adding machine learning frameworks with more number of features.

8. Acknowledgements

The work reported in this paper was supported by a grant from the India-Japan Cooperative Programme (DSTJST) 2009 Research project entitled "Sentiment Analysis where AI meets Psychology" funded by Department of Science and Technology (DST), Government of India.

9. References

Aery Manu, Naveen Ramamurthy, Y. Alp Aslandogan. (2005).Topic Identification of Textual Data. Project Report.

Attardi Giuseppe, Maria Simi. (2006). Blog Mining through Opinionated Words. In Proceedings of TREC 2006 the Fifteenth Text Retrieval Conference, pp.2-7.

Baccianella Stefano, Andrea Esuli, and Fabrizio Sebastiani.(2006).SENTIWORDNET 3.0: An Enhance Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-06, pp. 2200-2204.

Bal Krishna Bal, Patrick Saint Dizier. (2010). Towards Building Annotated Resources for Analyzing Opinions and Argumentation in News Editorials. Proceedings, LREC, pp. 1152-1158.

Brown Gillian and George Yule. (1983). Discourse Analysis. Cambridge Textbooks in Linguistics, Cambridge University, Press, Cambridge, UK.

Chen Kuang-hua. (1995).Topic Identification in Discourse. Proceeding EACL '95 Proceedings of the seventh conference on European chapter of the Association for Computational Linguistic.

Chesley Paula, Bruce Vincent, Li Xu, and Rohini K.Srihari.(2006). Using Verbs and Adjectives to Automatically Classify Blog Sentiment. pp.25-28.

Coursey Kino, Rada Mihalcea and William Moen. (2009). Using Encyclopedic Knowledge for Automatic Topic Identification. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL\'09),pp 210-218.

Das Dipankar and Sivaji Bandyopadhyay. (2009). Sentence Level Emotion Tagging. In the proceedings of the 2009 International Conference on Affective Computing & Intelligent Interaction (ACII-2009). pp. 375-380.

Das Dipankar and Sivaji Bandyopadhyay. (2009). Word to Sentence Level Emotion Tagging for Bengali Blogs. (ACL IJCNLP-2009), pp.149-152. Suntec, Singapore.

Ertöz L., M. Steinbach, and V. Kumar. (2001). Finding Topics in Collections of Documents:A Shared Nearest Neighbor Approach. In Proceedings of the Text Mine '01, Workshop on Data Mining, 1st SIAM International Conference on Data Mining.

Godbole Namrata, Manjunath Srinivasaiah, Steven Skiena.(2007). LargeScale Sentiment Analysis for News and Blogs. ICWSM'2007 Boulder, Colorado, USA,pp. 1-4.

Guo Jiafeng, Gu Xu, Xueqi Cheng, Hang Li.(2009). Named Entity Recognition in Query. *Proceedings of the 32nd* international ACM SIGIR conference, pp 267-274.

Kim Soo-Min, Eduard Hovy.(2004). Determining the Sentiment Of Opinions. Proceedings of the 20th international conferenc on Computational Linguistics COLING,pp. 1367-es.

Ku Lun-Wei, Yu-Ting Liang and Hsin-Hsi Chen. (2000). Opinion Extraction, Summarization and Tracking in News and Blog Corpora, Artificial Intelligence, pp 100-107.

Lin Chin-Yew. (1999). Robested Automated Topic identification. Faculty of the Graduate School, University of Southern California. ACL, pp.308-310.

Macherey Wolfgang and Hermann Ney. (2002).Towards Automatic Corpus preparation for a German Broadcast news Transcription system. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.

Mogadala Aditya and Vasudeva Varma.(2011).Finding Influence by Cross-Lingual Blog Mining through Multiple Language Lists. International Conference on Information systems for Indian Languages. pp. 54-59.

Jurafsky Daniel and James H. Martin. (2009). Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Stein Benno, Sven Meyer zu Eissen. (2004). Topic Identification: Framework and Application. Paderborn University, Germany. Proc of International Conference on Knowledge Management.

Yang Changhua Kevin Hsin-Yih Lin Hsin-Hsi Chen. (2007). Emotion Classification Using Web Blog Corpora. ACM International Conference on Web Intelligence, IEEE, pp.275-278.

Towards SoMEST – Combining Social Media Monitoring with Event Extraction and Timeline Analysis

Yue Dai, Ernest Arendarenko, Tuomo Kakkonen, Ding Liao
School of Computing
University of Eastern Finland
{yvedai, earendar, tkakkone, dliao}@cs.joensuu.fi

Abstract

We report on the development of a social media monitoring tool based on the novel *Social Media Event Sentiment Timeline* (SoMEST) model. The novelty of our model is that it combines opinion mining techniques with a timeline-based event analysis method and an information and event extraction tool. While *Event Timeline Analysis* (ETA) is an existing method utilized in analyzing the external environment of businesses, the SoMEST model and the BEECON (*Business Events Extractor Component based on Ontology*) tool as well as the OMS (*Opinion Miner for SoMEST*) component we report on are developed by the authors of the current paper.

1 Introduction

Successful business enterprises have a high level of awareness of the events that occur in their environment. Such events include various actions taken by the competitors, changes in legislation and technological advancements in the relevant branches of industry. In order to understand the needs and opinions of customers, companies also need to listen to the customer's voices that are presented, among other sources, in *social media* (SM). The volume of textual information available in online news outlets and SM, however, makes it extremely difficult to provide an integrated view of what the customers are saying online and the events that take place in the business environment.

To our knowledge, no practical method or software system exists that combines the two perspectives of monitoring the environment and listening to the voice of the customers. In the existing models and systems changes in customer opinions are not directly linked with the events that take place in the company's environment. The *Social Media Event Sentiment Timeline* (SoMEST) model that we introduced in (Dai, Kakkonen & Sutinen, 2011) is an analysis framework that aims at addressing this issue. The model is a combination of *event timeline analysis* (ETA), *opinion mining* (OM) techniques and information and *event extraction* (EE) methods that aims at deep exploration and understanding of business intelligence and competitive intelligence collected from online financial news and SM.

OM refers to the identification of opinions that a particular text through extracting and analyzing judgments on various aspects; it attempts to automatically classify human opinions (positive, negative, and neutral) from a text written in a natural language (Pang & Lee, 2008; Bhuiyan, Xu & Josang, 2009). The essential issues in OM research relevant to SoMEST include detection of topics (what is being talked about), opinion holders (who expressed the opinion), *opinion* polarity identification (positive or negative), and opinion intensity (ranking opinions based on their strength).

ETA refers to the systematic charting of events related to a specific topic or event; it provides a way of representing and explaining sequences of events (Qiu, Li, Qiao, Li & Zhu, 2008). Applied to the business domain, ETA has the potential to answer many crucial strategic questions and to predict the future development of industries and corporations. A variety of *natural language processing* (NLP) technologies can support ETA. For example, EE and tracking techniques have been used for environmental

scanning (Fleisher & Bensoussan, 2007; Liu, Shih, Liau & Lai, 2009).

The paper is organized as follows: Section 2 introduces the background of this work. In Section 3, we report on the process of implementing a software system based on the SoMEST model. Section 4 summarizes the main topics of the paper and outlines opportunities for future work.

2 Background

2.1 DAVID

The work on SoMEST is part of a larger effort to build the *Data Analysis and Visualization aId for Decision-making* (DAVID) system. The aim of DAVID is to derive from written texts information for business decision-making by using methods such as entity and event extraction, categorization, clustering, OM and visualization. In addition to the SoMEST model (Section 2.2), tools such as the CoProE ontology (Section 2.3), BEECON (Section 2.4) and OMS (Section 2.5) outlined in this paper are being developed as a part of the effort of building the DAVID system.

2.2 SoMEST Model

SoMEST is a unified model to combine EE and OM mining with a well-known competitive intelligence analysis method ETA (Dai, Kakkonen & Sutinen, 2011). The model is unique as it analyses simultaneously both the competitors and the customers by monitoring the market events abd by exploring and organizing SM content and thereby aggregating disparate pieces of information into meaningful *social media profiles*. Figure 1 outlines the SoMEST framework.

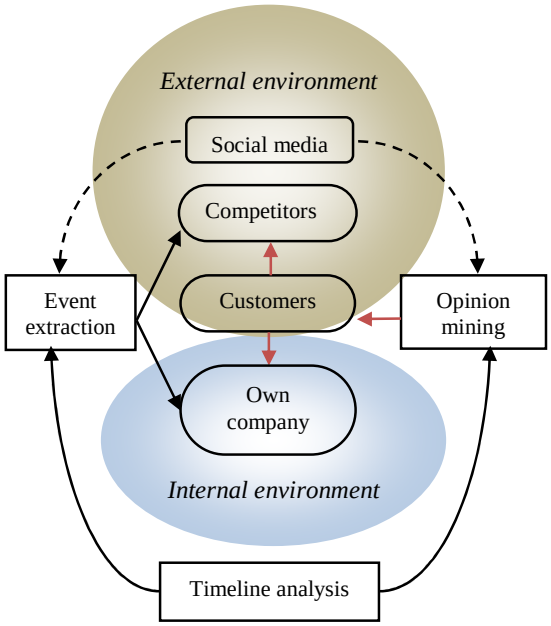

Figure 1. SoMEST framework

As illustrated in Figure 1, The SoMEST framework consists of three *objectives:* competitors, consumers and the company itself. SM is considered to be a part of the external environment. Competitors and consumers are the two major players in the external environment. Both competitors and customers generate new information to SM. In SoMEST, EE and OM are used to analyze pieces of information collected from SM and news articles. These two techniques have distinct foci of analysis. While EE is mostly concerned with analysing events from news and from texts published by companies, OM is used to understand customers' opinions towards one's own company and the competitors. While OM can analyse customers' opinions about brands, products, services and the whole company, it does not allow explaining why customers do or do not purchase certain products.

ETA combines the results of OM and EE analyses into visual charts that help to identify trends and patterns and thus support business leaders in finding possible explanations and solutions. Hence, SoMEST can help recognize business threats and opportunities from the external environment based on texts collected from SM and online news articles.

2.3 CoProE

We developed the *company, product and event* (CoProE) ontology (Kakkonen & Mufti, 2011) for representing domain knowledge in the DAVID system. The ontology is based on reuse of existing freely available resources. As the name suggests, CoProE allows describing information relevant to business intelligence and competitive intelligence. The most important part of the ontology in the context of the work presented in this paper is the one that enables to represent 43 business event types related to companies (for instance, collaboration, bankruptcy, expansion, merger, investment) and products (for example, new product release, product recall and product-related issue).

2.4 BEECON

BEECON is a software tool that extracts from input texts business entities and relations between them. It makes use of the CoProE ontology as the source of knowledge to find information about companies and products of interest. The system is capable of recognizing 41

of the event types defined in CoProE. BEECON is built on top of the GATE (*General Architecture for Text Engineering*) platform (Bontcheva et al., 2004) by improving existing GATE processing resources and by adding new ones.

The two main components of BEECON are *event detector* and *company reference detector*. The first processing resource is based on a finite-state transducer. It consists currently of around 200 hand-crafted rules which define patterns for detecting business events and information related to them, such as timestamp, currency, percentage unit etc. The company reference detector looks for mentions of companies in a text and is capable of resolving cases such as "the largest U.S. oil company" or "the Swedish automaker" and matching them with the referenced company. Other development work done on BEECON thus far includes:

- updating GATE to better recognize times, dates and other *named entity* (NE) categories;
- enriching existing and adding new categories to the default GATE gazetteer to detect stock exchanges, analyst and rating agencies and financial metrics;
- improving NE recognition by adding more rules and supporting financial entities;
- updating Orthomatcher to better recognize company aliases.

All the processing resources are arranged into a single GATE pipeline which is executed on a corpus of input business documents. The extraction process is totally automatic and does not require any human involvement.

2.5 Opinion Mining

We are currently developing OMS - a *machine learning* (ML) based OM component to be used in the SoMEST-based system. OMS performs OM in three main steps: (1) identifying opinion-bearing words, (2) labelling the orientation and strength of sentiment for each word, and (3) calculating the overall polarity and sentiment strength for the input document based on the values for each of its components.

The system first goes through all the input documents, removes any redundant words and writes them into a database. OMS then finds those documents that contain expressions of opinion. The next processing step consists of extracting the opinion words and identifying their orientation with the help of a sentiment word list (Liu et al., 2005) that consists of around 2 000 positive and 4 700 negative terms. The orientation of a document is classified by using either *support vector machines* (SVMs) or *perceptron algorithm with uneven margins* (PAUM). The words extracted in the previous step as well as the roots of the word tokens are used as input features for ML. In the last processing step, the orientation of each opinion review is identified and a final document score is produced.

3 Implementing a Software System based on SoMEST

3.1 Introduction

Constructing a software system based on the SoMEST model involves designing and implementing an architecture that takes as input the outputs of BEECON and OMS, combines them into an ETA timeline and represents the results as graphical charts. This architecture and all the subcomponents of the system are written in Java.

3.2 Current status of system development

3.2.1 Entity and event extraction

In order to implement and improve BEECON, we have conducted two development and evaluation iterations using data sets consisting of 250 (test set A) and 550 sentences (test set B). We sourced the test data from well-known online news sources such as Wall Street Journal, Reuters and Financial Times websites, as well as from various corporate websites. We constructed the evaluation data by manually extracting from the collected news articles all the sentences that contained one or more relevant business events and annotated them with the event categories defined in CoProE. Accuracy of BEECON was evaluated by analyzing the whole documents with the system and comparing the event categories it assigned with the manually assigned event category tags.

The first test was conducted by using the initial untested rule set and test set A. On this first evaluation, the precision was 70% and recall 50%. Next, we improved NE and IE components of BEECON as well as wrote new event detection rules by using test set A as the development set. The aim was to achieve as high precision and recall as possible before moving on to the next development iteration. After the accuracy on data set A was deemed satisfactory, the system was tested by using data set B.

The precision and recall on the first test run on data set B were 95% and 67% respectively. We are currently using B as the development set in the ongoing development iteration. Preliminary evaluation results on a subset of the third test set consisting of 2 200 documents indicate that he precision is similar to what was achieved on test set B. However, the recall shows an improvement of around 6 percentage points.

3.2.2 Opinion mining

We have evaluated the OMS component by using the well-known movie review data set by Pang & Lee (2005). The data is labeled with polarity information. We randomly chose training data which consisted of 1500 positive and 1500 negative reviews. The performance on this data set on 5-fold cross validation test was precision 69% and recall 68% with SVM compared to 67% precision and 67% recall with PAUM.

3.2.3 Implementing the SoMEST framework

The SoMEST-based analysis process has three main processing phases: collection, extraction & classification and synthesis. Each timeline consists of consecutive *time points*. In each time point, one or more *social media records* (SMR) are automatically collected. SMRs consist of four features: time, content, publisher, and the number of views.

In the extraction and classification phase, the SMRs collected in the previous phase are analysed to form two types of extracts. The features of an *event extract* are time, actor, action, objectives and place. We can use BEECON to recognize time, actor, action and objectives from the content of the relevant SMR. The features of an *opinion extract* are time, topic, opinion holder, polarity and intensity. BEECON can help in detecting the topic and the opinion holder, while OMS is used for extracting the intensity and polarity of the opinion.

In the synthesis phase, extracts are combined into *social media profiles* that describe sequences of time points (i.e. time periods). A social media profile provides a unified view of all the events and opinions connected to brands, products, services and leaders of a company during a given time frame.

The current status of the system development in as follows: In addition to developing OMS and BEECON, we have established a database for storing SMRs, event extracts, opinion extracts, and social media profiles. We have also designed and implemented a visualizations component that will allow showing SoMEST reports to the users. A report consists of a timeline that visualizes the specified social media profile that shows both the relevant events (event extracts) as well as the changes in customer opinions (based on opinion extracts).

4 Conclusions

We have introduced the SoMEST model that combines SM and news monitoring with automatic event detection and timeline analysis. We described the steps we have taken towards implementing SoMEST in a software system. We are currently building the system on top of well-known Java tools for NLP, ML and information and event extraction. The current version of our EE tool achieves an acceptable level of accuracy (around 95% precision and 70% recall) on realistic test data. As our test data has been collected from various sources, these figures indicate that the system is reaching the point in which it can be used as a component of practical NLP systems after we conclude the third test and development iteration.

The OMS opinion mining component has the recall and precision of 69% and 68% respectively on a standard OM test set. These accuracy figures call for improvements, in particular in relation to the precision. For instance, the system reported by Hu and Liu (2004) is somewhat similar to ours. They achieved the precision 84% and recall 69% on a dataset consisting of customer reviews collected from Twitter. The training data they used was much larger than the one we have used so far; it consisted of 1 600 000 tweets.

Our ongoing and planned work on SoMEST involves implementing a fully functional system and evaluating it in real business environments. This work involves improving the coverage and accuracy of the information extraction and NE components as well as the event detection rules of BEECON. Our efforts on OMS will be in particular concentrated on improving the recall.

Acknowledgements

The research reported in this paper was funded by the project “Towards e-leadership: higher profitability through innovative management and leadership systems” which is funded by the European Regional Development Fund and TEKES – the Finnish funding agency for technology and innovation.

References

Bontcheva, K., Tablan, V., Maynard, D. & Cunningham, H. (2004). Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*. 10 (3/4), pp. 349-373.

Bhuiyan, T., Xu, Y. & Josang, A. (2009). State-of-the-art Review on Opinion Mining from Online Customers' Feedback. In *Proceedings of the 9th Asia-Pacific Complex Systems Conference*, Tokyo, Japan, pp. 385-390.

Dai, Y., Kakkonen, T. & Sutinen, E. (2011). SoMEST - a Model for Detecting Competitive Intelligence from Social Media. In *Proceedings of the 15th MindTrek Conference*, Tampere, Finland, pp. 241-248.

Fleisher, C. S. & Bensoussan, B. E. (2007). *Business and Competitive Analysis: Effective Application of New and Classic Methods*. Upper Saddle River, New Jersey, USA: Pearson Education, Inc.

Hu, M. & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM International Conference on Knowledge Discovery and Data Mining*. New York, USA, pp. 156-163.

Kakkonen, T. & Mufti, T. (2011). Developing and Applying a Company, Product and Business Event Ontology for Text Mining. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, Graz, Austria.

Liu, B., Hu, M. & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th International World Wide Web Conference*. Chiba, Japan, pp. 342-351.

Liu, D. R., Shih, M. J., Liau, C. J. & Lai, C. H. (2009). Mining the Change of Event Trends for Decision Support in Environmental Scanning. *Expert Systems with Applications*, 36, pp. 972-984.

Pang, B. & Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan, USA, pp, 115-124.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundation and Trends in Information Retrieval*, 2(1-2), pp. 1-135.

Qiu, J., Li, C., Qiao, S., Li, T. & Zhu, J. (2008). Timeline Analysis of Web New Events. In *Proceedings of The 4th International Conference on Advanced Data Mining and Applications*, Chengdu, China, pp. 123-134.

A Corpus for Entity Profiling in Microblog Posts

Damiano Spina*, Edgar Meij†, Andrei Oghina†, Minh Thuong Bui†, Mathias Breuss† , Maarten de Rijke†

* UNED NLP & IR Group
Juan del Rosal, 16
28040 Madrid, Spain
damiano@lsi.uned.es

†ISLA, University of Amsterdam
Science Park 904
1098 XH, Amsterdam, The Netherlands
edgar.meij@uva.nl, {oghina,mbui,mbreuss}@science.uva.nl, derijke@uva.nl

Abstract

Microblogs have become an invaluable source of information for the purpose of online reputation management. Streams of microblogs are of great value because of their direct and real-time nature. An emerging problem is to identify not only microblog posts (such as tweets) that are relevant for a given entity, but also the specific aspects that people discuss. Determining such aspects can be non-trivial because of creative language usage, the highly contextualized and informal nature of microblog posts, and the limited length of this form of communication. In this paper we present two manually annotated corpora to evaluate the task of identifying aspects on Twitter, both of them based upon the WePS-3 ORM task dataset and made available online. The first is created using a pooling methodology, for which we have implemented various methods for automatically extracting aspects from tweets that are relevant for an entity. Human assessors have labeled each of the candidates as being relevant. The second corpus is more fine-grained and contains opinion targets. Here, annotators consider individual tweets related to an entity and manually identify whether the tweet is opinionated and, if so, which part of the tweet is subjective and what the target of the sentiment is, if any.

1. Introduction

Online Reputation Management (ORM) deals with monitoring and handling the public image of entities, such as people, products, organizations, or brands, on the Web. In the field of ORM, much of the effort is focused towards analyzing mentions on social web streams that are relevant to the entity of interest. An emerging problem in this area is to identify not only microblog posts (such as tweets) that are relevant for a given entity, but also the specific *aspects* that people discuss.

Aspects refer to "hot" topics that people talk about in the context of an entity—the principal vectors that coagulate the public interest regarding the company. Aspects can cover a wide range of notions and they include, without being limited to, company products, services, key people, and events. They can change over time as public attention shifts from some aspects to others. For instance, when a company releases its quarterly earnings report, this can become, for a certain period of time, a topic of discussion and, hence, an aspect. Although aspects have been investigated in the context of, e.g., discussion fora (Thet et al., 2010), automatically determining aspects on streams of microblog posts is still an unsolved problem.

A well-known application in the context of ORM on social web streams is sentiment analysis (Jansen et al., 2009), with numerous online demos and tools. Since state-of-the-art methods for sentiment analysis still yield noisy results, it is common to measure aggregate sentiments, i.e., aggregating sentiment scores for a set of microblog posts. While measuring such "overall" sentiment has its merits, it also has obvious limitations. Especially in the context of entities such as large companies—which typically have many products or services to offer—a more fine-grained approach is needed.

Some current ORM tools such as UberVU[1] allow online reputation managers to monitor sentiment regarding a predefined set of keywords, such as product names (Amigó et al., 2010). However, the fluidity of microblogging streams renders this method too rigid, since aspects can have a dynamic nature, changing and emerging over time. Therefore, a better approach would be to extract the relevant, most discussed aspects of an entity in an automatic fashion.

To the best of our knowledge, there are no readily available datasets suitable to evaluate the task of identifying either aspects or opinion targets in the context of ORM on social web streams. In this paper we present two manually annotated corpora to fill this gap. Both of them are based upon the WePS-3 ORM task and will be made available online.[2]

The first dataset is created using a pooling methodology. Here, we have implemented various methods for automatically extracting aspects from tweets that are relevant for an entity. We subsequently generate a ranked list of aspects using each method, take the highest ranked aspects, and pool them. Then, human assessors consider each aspect and determine whether it is relevant in the context of the entity or not. The second dataset that we present is similar, but more fine-grained. Here, annotators consider individual tweets related to an entity and manually identify whether the tweet is opinionated and, if so, which part of the tweet is (i) sub-

[1] http://www.ubervu.com/walkthrough/

[2] http://nlp.uned.es/~damiano/datasets/entityProfiling_ORM_Twitter.html

jective and (ii) what the target of the sentiment is, if any.
In the next section, we briefly discuss related work and datasets. In Section 3. we describe the WePS-3 ORM task dataset, upon which our annotated corpora are based. In Sections 4. and 5. we introduce the corpus containing the entity aspects and the one containing the opinions, respectively. Section 6. briefly compares the two corpora, including an analysis of the overlap between them. We end with a concluding section.

2. Related Work

In other domains—such as product reviews or news—there exist various datasets to investigate aspects, typically in the form of opinion targets (Hu and Liu, 2004; Kim and Hovy, 2006; Wiebe et al., 2005). However, to the best of our knowledge, there are no manually annotated corpora to evaluate this task on microblog streams. Determining such aspects on streams of microblog posts can be non-trivial because of the creative language usage (including slang, emoticons, and acronyms), the highly contextualized and informal nature of microblog posts, and the limited length of this form of communication (Kaufmann and Kalita, 2010). This reduces the applicability of the techniques developed for other domains. Moreover, the amount of data produced on microblogging streams is substantially larger than that produced in customer reviews or news media, opening up opportunities for leveraging cross-post redundancy.
So far, most of the manually annotated corpora built upon Twitter are annotated at the level of individual tweets. For example, both the TwitterSentiment[3] and Sanders[4] corpora contain tweets labeled with subjectivity and polarity (i.e. positive, negative, and neutral).
In the TREC 2011 Microblog track,[5] the gold standard for the ad hoc real-time search task was built using a pooling methodology. The corpus used in this task was the Tweets2011 corpus[6]. Another recently released Twitter dataset contains semantic annotations, where each tweet is manually linked to a set of entities in the form of Wikipedia articles (Meij et al., 2012). Similarly, the WePS-3 ORM dataset links tweets to companies, as described in the next section.

3. WePS-3 ORM

Determining aspects of an entity in the context of streams of microblog posts such as tweets involves two tasks. In the first task, tweets relevant to a given entity need to be identified, while in the second these tweets need to be analyzed in order to identify aspects. In this paper we focus mainly on the second task and base our annotations on the data used for the WePS-3 ORM Task (Amigó et al., 2010). Here, the task that participating systems needed to solve was to filter tweets containing a given company name depending on whether the post is actually related to the company or not. This is challenging for ambiguous names, such as *Apple* or *Fox*. In total, 99 companies were used, with around 450 tweets on average for each, summing up to a total of 45,201 tweets. Mechanical Turk was used to perform the relevance assessments; each tweet is annotated as being either *related* or *unrelated* to a given company.
For the annotations presented in this work, only the tweets that are related to each company are considered. For our first dataset pertaining to the identification of aspects, a total of 94 companies have been considered. This adds up to 17,775 tweets in total, with an average of 177 tweets per company. From this set, all the related tweets for 59 companies have been annotated in a second round, where we identify opinion targets and subjective phrases. The latter corpus constitutes our second dataset and includes 9,396 tweets in total, i.e., an average of 159 tweets per company.

4. Annotating Aspects

Let us consider the following profiling scenario: given a stream of tweets that are related to a company, we are interested in a ranked list of aspects representing the hot topics that are being discussed with respect to the company. Examples of aspects include products, services, key people, events, or entities that are associated with the company in a certain time frame.
This scenario can be formulated as an information retrieval task, where the goal of a system implementing a solution to this task is to provide a ranking of terms, extracted from tweets that are relevant with respect to the company.[7] We have implemented various methods addressing this task. For each company, each method returns a ranked list of terms associated with each company. The underlying principle for all methods is a comparison of the contents of the relevant tweets—henceforward, the *foreground* corpus—with a common *background* corpus, e.g., the whole WePS-3 collection. Using this comparison we identify and score terms based on their relative occurrence. Our methods include TF.IDF (Salton and Buckley, 1988), the log-likelihood ratio (Dunning, 1993) and parsimonious language models (Hiemstra et al., 2004). Since aspects can be opinion targets, we also applied an opinion-oriented method (Jijkoun et al., 2010) that extracts potential targets of opinions to generate a topic-specific sentiment lexicon. We use the targets selected during the second step of this method.
This dataset is then created using a pooling methodology (Harman, 1995): the 10 highest ranking terms from each method are merged and randomized. Then, human assessors consider each term and determine whether it is relevant in the context of the company or not.

4.1. Annotations

The annotators were presented with an annotation interface, where they could select one of the companies from a list. Once a company is selected, the interface shows a randomized list of aspects. The interface also facilitated looking up

[3] http://twittersentiment.appspot.com
[4] http://www.sananalytics.com/lab/twitter-sentiment/
[5] http://sites.google.com/site/microblogtrack/2011-guidelines/
[6] http://trec.nist.gov/data/tweets/
[7] In our current setup, we only consider unigrams as aspects. When a unigram is an obvious constituent of a larger, relevant aspect, it is considered relevant.

a term; when clicked, the system would present all tweets that are relevant to the company and contain that particular term. The annotators could indicate one of the following labels for each aspect:

- **Relevant:** A relevant aspect can include, e.g., product names, key people, events, etc. Relevant aspects are in general nouns, but can also be verbs, and (rarely) adjectives. Relevant aspects can include terms from compound words, mentions or hashtags. Aspects should provide some insight into the hot topics discussed regarding a company, topics that would also differentiate it from other more general discussions, or its competitors.
- **Not relevant:** Common words and words not representing aspects or sub-topics are not relevant.
- **Competitor:** A term is (part of) a competitor name, including an opponent team name, a competing company or a product from a competing company.
- **Unknown:** If, even after inspecting the tweets were the term occurs, the judge still cannot use the other labels.

In this work we treat the label *Competitor* as being *Relevant*, although the data set contains this explicit label for possible follow-up work. Table 1 shows some examples of the aspects annotated in the corpus.

Entity	Aspects
A.C. Milan	milanello, ac, football, milan, galliani, berlusconi, brocchi, leonardo
Apple Inc.	ipad, iphone, prototype, apple, store, gizmodo, employee, gb
Sony	advertising, set, headphones, digital, pro, music, sony, xperia, dsc, x10, bravia, camera, vegas, battery, ericsson, playstation
Starbucks	coffee, latte, tea, frappuccino, starbucks, shift, pilot, barista, drink, mocha

Table 1: Examples of aspects annotated for some of the entities in the corpus.

4.2. Analysis

In order to determine the level of agreement between the three annotators J_i, we calculate *Cohen's kappa* and *Fleiss' kappa* (Landis and Koch, 1977) and compare the annotators both pairwise and overall. The results are given in table 2. All of the obtained kappa values are above 0.6, which indicates a substantial agreement.

Method	J_1-J_2	J_1-J_3	J_2-J_3	All
Cohen's κ	0.691	0.62	0.676	-
Fleiss' κ	0.69	0.62	0.676	0.662

Table 2: Inter-annotator agreement for the aspects dataset.

In the WePS-3 ORM dataset, the number of tweets relevant to each company is highly variable (Amigó et al., 2010). Thus, one could expect correlations between the ratio of relevant tweets and the ratio of relevant aspects annotated for each company.

Tweets	|C|	AvgTw	AvgTer	AvgRel	Rel%
0-10	19	4.05	12.47	2.79	22.36%
11-50	15	22.20	22.00	8.53	38.79%
51-150	12	97.67	26.75	13.58	50.78%
151-300	25	219.40	28.80	16.40	56.94%
301+	28	381.43	30.64	19.46	63.52%

Table 3: Distribution of relevant aspects, binned by the number of relevant tweets per company.

Table 3 shows the number of tweets, the number of extracted terms (*AvgTer*), and the number of identified relevant aspects (*AvgRel*) based on the annotations. For this, we consider all terms included in the pooling, and divide the entities in five groups, based on the number of tweets available for each company (0-10, 11-50, 51-150, 151-300, 301+). For each group C, we count how many companies are part of the group ($|C|$) and the average number of tweets for these entities (*AvgTw*). We also compute the percentage of the aspects that are relevant (*Rel%*).

We observe that the percentage of relevant aspects across increases with the amount of data available. For companies that have no more than 10 tweets each, only 22.36% of extracted aspects are annotated as being relevant. On the other hand, for entities with more than 300 tweets, 63.52% of all extracted aspects were annotated as being relevant. This suggests that the amount of data available plays an important role in the performance of the methods used for the pooling.

5. Annotating opinion targets

The second dataset we present consists of the tweets of 59 entities from the WePS-3 dataset, manually annotated at the phrase-level. Here, we aim to identify opinion targets in tweets, related to an aspect of a company. We define an opinion target as a phrase p that satisfies the following properties: (i) p is an aspect of the entity, (ii) p is included in a sentence that contains a direct subjective phrase (i.e. an expression that explicitly manifests subjectivity or an opinion) and (iii) p is the target of the expressed opinion.

5.1. Annotations guidelines

The annotators were asked to indicate the following.

- **Subjectivity:** Tweet-level annotation that indicates whether the tweet contains an explicit opinionated expression.
- **Subjective phrase:** If the tweet is opinionated, identify the phrase that express subjectivity. In our annotation schema, we only considered direct private states (Wiebe et al., 2005).
- **Opinion target:** If the tweet contains opinionated phrases, identify the target of the opinion expressed in that phrases.

Table 4 show some examples of opinionated tweets.

Phrase-level annotation require much more effort than tweet-level annotations or aspect assessments. In order to maximize the number of annotated entities, 59 entities were randomly distributed over seven different annotators, making a disjoint assignment of annotators to data.

Entity	Tweet
Linux	Lxer: A Slimline Debian Install: Its Easier Than You Might Think: There are some *superb* **desktop Linux distributions**... http://bit.ly/8ZSaF
MTV	@MTV has the *best* **shows** ever. i watch it all day every day (:
Oracle	IMHO, the *best part of* Oracle now **owning Java** is that whenever **Java** is *criticized* for something, Oracles name is attached.
Sony	@user Welll Im not getting one then. **Sony** is *expensive*
Starbucks	**The Dark Cherry Mocha** from @Starbucks is just *the best Mocha ever!*

Table 4: Examples of phrase-level annotated tweets, having subjective phrases (italic) and opinion targets (boldface).

5.2. Analysis

In total, 9,396 tweets were annotated. Only 1,427 (15.16%) tweets contain subjective phrases and 1,308 (13.82%) contain opinion targets. There are 119 tweets where the annotators identified subjective phrases but not opinion targets. Most of them are tweets containing either emoticons or phrases expressing subjectivity at tweet-level (e.g. `LOL`, `Yay!`, `#fail`).

Analogous to the first dataset, we divided the annotated entities in groups based on the number of annotated tweets and computed the average of tweets with subjective phrases (*AvgSubj*) and opinion targets (*AvgOT*). Table 5 reports these averages as well as the averaged percentage of subjective tweets (*Subj%*).

Tweets	|C|	AvgTw	AvgSubj	AvgOT	Subj%
0-10	7	3.57	0.85	0.85	35.11%
11-50	11	23.36	3.64	3.09	14.24%
51-150	9	96.22	11.77	10.33	11.88%
151-300	19	218.68	25.21	23.10	14.22%
301+	13	392.54	61.23	56.61	15.8%

Table 5: Distribution of subjective phrases and opinion targets, binned by the number of relevant tweets per company.

6. Aspects vs. Opinion targets

In this section we analyze the vocabulary overlap between the terms identified in the two corpora presented in this paper, i.e., between aspects and opinion target terms.

For the first dataset we consider a majority vote, labeling terms as relevant when they are annotated as such by two or more judges. We further restrict ourselves to the same 59 entities annotated with opinion targets in the second dataset. We tokenize the phrases identified as opinion targets, keeping the constituent terms that occur in them after removing stopwords and symbols. As an example, Table 6 shows opinionated aspects for some of the entities in the datasets.

From a total of 783 aspects, 209 (26.69%) occur in opinion target phrases. Vice versa, the total number of terms extracted from the opinion target phrases is 1650; only 12.66% of those are also identified as relevant aspects. The

Entity	Aspects in opinion targets
Jaguar Cars Ltd.	jaguar (0.26), xj (0.06), cars (0.02), rover (0.01), car (0.01), auto (0.01), xf (0.01)
Linux	linux (0.12), multitouch (0.02)
Sony	sony (0.05), music (0.04), vegas (0.03), headphones (0.02), battery (0.02), xperia (0.01), pro (0.01), ericsson (0.01), x10 (0.01), playstation (0.01), bravia (0.01), camera (0.01)
Starbucks	starbucks (0.33), coffee (0.11), tea (0.06), frappuccino (0.03), drink (0.03), latte (0.02)

Table 6: Examples of aspects that are included in opinion target phrases, with the frequency in opinion targets in parentheses.

overlap between aspects and opinion targets is lower than expected. The low overlap is probably given by the different methodologies used to annotate aspects and opinion targets. While aspects were annotated using a pooling methodology that considers the 10 highest ranking terms retrieved from each method, opinion targets were manually annotated inspecting the tweets related to each company.

We observe that, instead of an aspect, the actual name of the entity has a tendency to occur as a target. However, the remaining aspects occur only a few times, suggesting a power-law distribution. In fact, terms in opinion targets are very sparse. The average occurrence of a term in an opinion target equals 1.78 and more than 75% of all terms occur only once. This suggests that the WePS-based sample of around 150 tweets per entity might not be enough for opinion-based entity profiling. We leave verifying this hypothesis (and possibly creating a larger dataset) for future work.

7. Conclusions

An emerging problem in the field of online reputation management consists of identifying the key aspects of an entity commented in microblog posts. Streams of microblogs are of great value because of their direct and real-time nature and synthesizing them in form of entity profiles facilitates reputation managers to keep a track of the public image of the entity.

In this paper we have presented two manually annotated corpora to evaluate the task of identifying aspects on Twitter, both of them based upon the WePS-3 ORM task dataset and made available online. The first dataset we release contains aspects that are strongly related to a given company in a stream of tweets, while the second contains phrases in tweets that represent the targets and opinions expressed towards entities in those tweets. The low overlap between relevant aspects and terms occurring in opinion target phrases shows the different nature of the two corpora built. We believe that these resources will allow to evaluate different entity profiling systems in microblog posts and to make progress in the use of human language technologies for online reputation management.

8. Acknowledgements

This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.005, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, under COMMIT project Infiniti, the Spanish Ministry of Education (FPU grant nr AP2009-0507), the Education Council of the Regional Government of Madrid, MA2VICMR (S-2009/TIC-1542), the Innovation project Holopedia (TIN2010-21128-C02-01) and by the ESF Research Network Program ELIAS.

We would like to thank Dr. Irina Chugur and Dr. Wouter Weerkamp for helping us with the definition and annotation of the opinion target corpus.

9. References

E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. 2010. WePS-3 evaluation campaign: Overview of the online reputation management task. In *CLEF 2010 Labs and Workshops Notebook Papers*.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19.

D. Harman. 1995. Overview of the Fourth Text REtrieval Conference (TREC-4). In *TREC-4*.

D. Hiemstra, S. Robertson, and H. Zaragoza. 2004. Parsimonious language models for information retrieval. In *Proceedings of the 27th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04)*.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (SIGKDD '04)*.

B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188.

V. Jijkoun, M. de Rijke, and W. Weerkamp. 2010. Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*.

M. Kaufmann and J. Kalita. 2010. Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing (ICNLP '10)*.

S.M. Kim and E. Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *ACL Workshop on Sentiment and Subjectivity in Text*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33.

E. Meij, W. Weerkamp, and M. de Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12)*.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513 – 523.

T.T. Thet, J.C. Na, and C.S.G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.