Proceedings of the 2nd Workshop on

Making Sense of Microposts (#MSM2012)

Big things come in small packages

at the 21st International Conference on the World Wide Web (WWW'12)
Lyon, France
16th of April 2012

edited by

Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie,

Preface

The 2nd Workshop on *Making Sense of Microposts (#MSM2012)* was held in Lyon, France, on the 16th of April 2012, during the 21st International Conference on the World Wide Web (WWW'12). #MSM2012 follows on from a successful 1st workshop, #MSM2011, at the 8th Extended Semantic Web Conference (ESWC 2011), which, with approximately 50 participants, was the most popular workshop at ESWC 2011.

The #MSM series of workshops is unique in targeting both Semantic Web researchers and other fields, both within Computer Science, such as Human-Computer Interaction and Visualisation, and in other areas, particularly the Social Sciences. The aim is to harness the benefits different fields bring to research involving microposts. Moving the 2nd workshop to WWW allowed us to reach a wider and more varied audience.

Posting information about on-going events, exchanging information with one's social and working circles, or simply publishing one's train of thought on online social media platforms such as Twitter and Facebook, or contributing information about points of interest on Foursquare, is increasingly the norm in the online world. Support for the use of online social media on or via ubiquitous, small, mobile devices and near-permanent connectivity have further lowered the barrier to interaction with the online world. This has resulted in an explosion of small chunks of information being published with minimal effort (e.g. a 'tweet' or a 'check-in' on foursquare) – we refer to such user input as *microposts*. The reality of the trend of microposts' domination in online, end user-generated content can be seen in the appearance of new services that focus primarily on low-effort user input, such as Google+, whose aim is to bootstrap microposts in order to more effectively tailor search results to a user's social graph and profile.

The sheer scale of micropost data, generated using a variety of devices and on multiple platforms, by myriad users in as many different situations, requires new techniques to glean knowledge and provide useful services and applications sitting atop the amalgamation of this heterogeneous, distributed data. Further, the brevity of user expression in microposts imposes additional challenges for analysis. The #MSM workshop series was born to bring together researchers exploring novel methods for analysing microposts, and for reusing the resulting collective knowledge extracted from such posts, both on the Web and in the physical world. The #MSM2012 workshop discussed emerging to fairly advanced work on the research these challenges have engendered.

#MSM2012 continues to highlight the importance of maintaining a focus on the end user – ranging from the mainstream user of what is now ubiquitous technology, such as the mobile phone, tablet or desktop computer, with little to no technical expertise, to the Semantic Web expert – to ensure that appealing, useful and usable tools are designed and built, which harness the particular benefits of Semantic Web technology.

Many hearty thanks to all our contributors and participants, and also the Programme Committee whose valued feedback resulted in a rich collection of papers, posters and demos, each of which adds to the state of the art in leading edge research. We are confident that the #MSM series of workshops will continue to foster a vibrant community, and target the rich body of information generated by the many and varied authors whose social and working lives span the physical and online worlds.

Matthew Rowe KMi, The Open University, UK
Milan Stankovic Hypios / Université Paris-Sorbonne, France
Aba-Sah Dadzie The University of Sheffield, UK
#MSM2012 Organising Committee, April 2012

Introduction to the Proceedings

Out of a total of 19 paper submissions, 6 full and 3 short papers were accepted. This was in addition to a poster and demo session, to exhibit practical application in the field, and foster further discussion of the ways in which data extracted from Microposts is being reused. The accepted submissions cover an array of topics; we highlight these below.

The proceedings include also the abstract of the keynote, '*Information Theoretic Tools for Social Media*', presented by Greg Ver Steeg, of the Information Sciences Institute at the University of Southern California.

Sentiment and Semantics

Platforms, such as Twitter and Facebook, that support micropost publication allow users to vent their frustrations and express their opinions in a centralised and public space. Passive networks formed on such platforms are comprised of users *listening* to the signals produced by other users and consuming their published information. As a consequence, sentiment analysis of microposts has become a useful means for companies and organisations to gauge the collective sentiment and opinion regarding different entities and topics. In Saif et al.'s paper, '*Alleviating Data Sparsity for Twitter Sentiment Analysis*', the authors describe an approach to alleviate the data sparsity problem that affects sentiment analysis on Twitter through the use of semantic and sentiment-topic features. The authors demonstrate the efficacy of these additional features by outperforming a baseline model which neglects such additional information.

Semantics within microposts forms the basis for discussion in '*Small talk in the Digital Age: Making Sense of Phatic Posts*' by Radovanovic & Ragnedda. In this paper the authors present a theoretical discussion and analysis of the importance of microposts in providing diverse information across the Web. Following on from this theme of diversity and information utility is Zangerle et al.'s paper titled '*Exploiting Twitter's Collective Knowledge for Music Recommendations*'. In this work the authors demonstrate the utility of microposts in providing music recommendations through collective knowledge. In both works microposts are described as a useful source for diverse information that can in turn be used in differing applications and contexts.

Information Extraction

The masses of microposts published every day cover a wide range of subjects and topics. Extracting information from microposts related to the same entity or topic can provide a diverse perspective with regard to public perception and/or opinion. Prior to performing opinion analysis, entities must be recognised within the microposts and the information extracted accordingly. One issue of the diversity of microposts, however, is the prevalence of term ambiguity – where the same term can have multiple meanings. Context provides one mechanism for disambiguation; however, given the limited information size of a single micropost, obtaining such contextual information is challenging. This issue is addressed in the paper by Castro Reis et al. titled '*Extracting Unambiguous Keywords from Microposts Using Web and Query Logs Data*', by automatically detecting, and hence, enabling the extraction of terms in microposts which are not ambiguous. The authors present a handcrafted classifier that uses background knowledge of term features to yield high levels of precision, outperforming a Support Vector Machine in the same setting.

The second paper to address the topic of information extraction is '*Knowledge Discovery in distributed Social Web sharing activities*' by Scerri et al. In this work the authors address the challenges in managing and making optimal use of personal information, by arguing that social activity streams, both of a given user and members of his/her social network, provide useful means for the enrichment of personal information spaces. To this end the authors propose a framework for the extraction of information from disparate activity streams and the integration of the extracted information into existing personal information spaces, through the LivePost ontology presented.

Visualisation, Search and Networks

The scale and volume of microposts makes interpretation and analysis of such data limited to end users. One solution is to visualise microposts in a coherent and readable form, thereby facilitating sense-making and data exploration. The paper by Hubmann-Haidvogel et al. titled '*Visualizing Contextual and Dynamic Features of Microposts*' presents work that enables the visualisation of large volumes of microposts. The approach supports multi-faceted views to enable a range of information-seeking tasks. For instance, by presenting geographical information alongside topic-volume statistics, the end user is presented with an overview of microposts at a higher level of abstraction.

Search over microposts has recently become a topic of great interest, with the creation of the first 'Microblog' track[1] at the Text REtrieval Conference 2011. The diversity and ambiguity of terms found within microposts limits current retrieval paradigms and therefore requires the exploration of new methods for retrieval. In the paper by Tao et al. titled '*What makes a tweet relevant for a topic?*' the authors explore the effects of various features on retrieval performance over microposts. The features investigate the topic-sensitive and topic-independent effects on retrieval performance and find that by taking the former information into account performance is improved.

The final two papers investigate the dynamics and effects of networks associated with microposts. The first, by Wagner et al. titled '*When social bots attack: Modeling susceptibility of users in online social networks*', assesses which users are likely to fall foul of socialbot attacks and be influenced by the content the bots produce. The authors explore three different feature sets to describe users and find that users who engage a lot in conversational behaviour with their social network are more susceptible to attacks. The second paper in the area of networks is '*Understanding co-evolution of social and content networks on Twitter*' by Singer et al. In this work the authors explore how networks change over time through time-series analysis of social network measures. Their findings indicate that social networks have an influence on content networks.

[1] https://sites.google.com/site/microblogtrack/home

Workshop Awards

The Parisian Open Innovation startup, Hypios[2], sponsored an award for the submission that contributed best to making innovation happen on the Web. Best paper nominations were sought from the reviewers, and a final decision agreed by the Chairs, based on the nominations and review scores.

Additional Material

The call for participation and all paper, poster and demo abstracts are available on the #MSM2012 website[3]. The full proceedings are also available on the CEUR-WS server, as Vol-838[4]. The proceedings for the 1st workshop are available as CEUR Vol-718[5].

Programme Committee

Fabian Abel Leibniz University Hannover, Germany
Gholam R. Amin Sultan Qaboos University, Oman
Sofia Angeletou KMi, The Open University, UK
Pierpaolo Basile University of Bari, Italy
Uldis Bojars University of Latvia, Latvia
David Beer University of York
John Breslin NUIG, Ireland
A. Elizabeth Cano The University of Sheffield, UK
Óscar Corcho Universidad Politécnica de Madrid, Spain
Danica Damljanovic The University of Sheffield, UK
Ali Emrouznejad Aston Business School, UK
Guillaume Ereteo INRIA, France
Miriam Fernandez KMi, The Open University, UK
Fabien Gandon INRIA, Sophia-Antipolis, France
Andrés Garcia-Silva Universidad Politécnica de Madrid, Spain
Anna Lisa Gentile The University of Sheffield, UK
Jon Hickman Birmingham City University, UK
Seth van Hooland Free University of Brussels, Belgium
Jennifer Jones University of the West of Scotland, UK
Jelena Jovanovic University of Belgrade, Serbia
Vita Lanfranchi, The University of Sheffield, UK
Philipe Laublet Université Paris-Sorbonne, France
Pablo Mendes Kno.e.sis, Wright State University, USA
João Magalhães Universidade Nova de Lisboa, Portugal
Julie Letierce DERI, Galway, Ireland
Diana Maynard The University of Sheffield, UK
Pablo Mendes Freie Universität of Berlin, Germany
José M. Morales del Castillo Universidad de Granada, Spain
Alexandre Passant DERI, Galway, Ireland
Danica Radovanovic University of Belgrade, Serbia
Yves Raimond BBC, UK
Harald Sack University of Potsdam, Germany
Bernhard Schandl University of Vienna, Austria
Andreas Sonnenbichler KIT, Germany
Raphaël Troncy Eurecom, France
Victoria Uren Aston Business School, UK
Claudia Wagner Joanneum Research, Austria
Shenghui Wang Vrije University, The Netherlands
Katrin Weller University of Düsseldorf, Germany
Ziqi Zhang The University of Sheffield, UK

[2] http://hypios.com
[3] http://socsem.open.ac.uk/msm2012
[4] http://ceur-ws.org/Vol-838
[5] http://ceur-ws.org/Vol-718

Table of Contents

Information Theoretic Tools for Social Media

Greg Ver Steeg
Information Sciences Institute
The University of Southern California
California, USA
gregv@isi.edu

Abstract

Information theory provides a powerful set of tools for discovering relationships among variables with minimal assumptions. Social media platforms provide a rich source of information than can include temporal, spatial, textual, and network information. What are the interesting information theoretic measures for social media and how can we estimate these quantities? I will discuss how measures like information transfer can be used to quantify how predictive some variables are, e.g., how well one user's activity can predict another's. I will also discuss techniques for estimating entropies even when the data are sparse, as is the case for spatio-temporal events, or very high-dimensional, as is the case for textual information.

Alleviating Data Sparsity for Twitter Sentiment Analysis

Hassan Saif
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA, UK
h.saif@open.ac.uk

Yulan He
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA, UK
y.he@open.ac.uk

Harith Alani
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA, UK
h.alani@open.ac.uk

ABSTRACT

Twitter has brought much attention recently as a hot research topic in the domain of sentiment analysis. Training sentiment classifiers from tweets data often faces the data sparsity problem partly due to the large variety of short and irregular forms introduced to tweets because of the 140-character limit. In this work we propose using two different sets of features to alleviate the data sparseness problem. One is the semantic feature set where we extract semantically hidden concepts from tweets and then incorporate them into classifier training through interpolation. Another is the sentiment-topic feature set where we extract latent topics and the associated topic sentiment from tweets, then augment the original feature space with these sentiment-topics. Experimental results on the Stanford Twitter Sentiment Dataset show that both feature sets outperform the baseline model using unigrams only. Moreover, using semantic features rivals the previously reported best result. Using sentiment-topic features achieves 86.3% sentiment classification accuracy, which outperforms existing approaches.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

General Terms

Algorithms, Experimentation

Keywords

Microblogs, Sentiment Analysis, Opinion Mining, Twitter, Semantic Smoothing, Data Sparsity

1. INTRODUCTION

Few years after the explosion of Web 2.0, microblogs and social networks are now considered as one of the most popular forms of communication. Through platforms like Twitter and Facebook, tons of information, which reflect people's opinions and attitudes, are published and shared among users everyday. Monitoring and analysing opinions from social media provides enormous opportunities for both public and private sectors. for private sectors, it has been observed [21, 22] that the reputation of a certain product or company is highly affected by rumours and negative opinions published and shared among users on social networks. Understanding this observation, companies realize that monitoring and detecting public opinions from microblogs leads to building better relationships with their customers, better understanding of their customers' needs and better response to changes in the market. For public sectors, recent studies [3, 9] show that there is a strong correlation between activities on social networks and the outcomes of certain political issues. For example, Twitter and Facebook were used to organise demonstrations and build solidarity during Arab Spring of civil uprising in Egypt, Tunisia, and currently in Syria. One week before Egyptian president's resignation the total rate of tweets about political change in Egypt increased ten-fold. In Syria, the amount of online content produced by opposition groups in Facebook increased dramatically.

Twitter, which is considered now as one of the most popular microblogging services, has attracted much attention recently as a hot research topic in sentiment analysis. Previous work on twitter sentiment analysis [5, 13, 2] rely on noisy labels or distant supervision, for example, by taking emoticons as the indication of tweet sentiment, to train supervised classifiers. Other work explore feature engineering in combination of machine learning methods to improve sentiment classification accuracy on tweets [1, 10]. None of the work explicitly addressed the data sparsity problem which is one of the major challenges facing when dealing with tweets data.

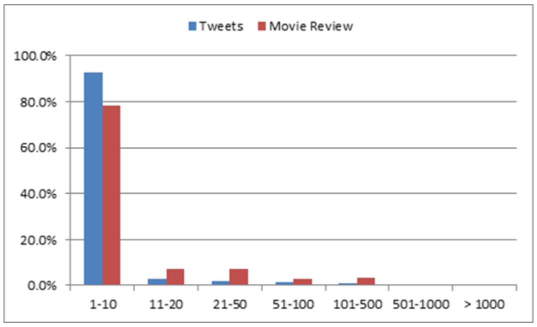


Figure 1: Word frequency statistics.

Figure 1 compares the word frequency statistics of the tweets data we used in our experiments and the movie review data[1]. X-axis shows the word frequency interval, e.g., words occur up to 10 times

Published as part of the #MSM2012 Workshop proceedings,
available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

[1] http://www.cs.cornell.edu/People/pabo/movie-review-data/

(1-10), more than 10 times but up to 20 times (10-20), etc. Y-axis shows the percentage of words falls within certain word frequency interval. It can be observed that the tweets data are sparser than the movie review data since the former contain more infrequent words, with 93% of the words in the tweets data occurring less than 10 times (cf. 78% in the movie review data).

One possible way to alleviate data sparseness is through word clustering such that words contributing similarly to sentiment classification are grouped together. In this paper, we propose two approaches to realise word clustering, one is through semantic smoothing [17], the other is through automatic sentiment-topics extraction. Semantic smoothing extracts semantically hidden concepts from tweets and then incorporates them into supervised classifier training by interpolation. An inspiring example for using semantic smoothing is shown in Figure 2 where the left box lists entities appeared in the training set together with their occurrence probabilities in positive and negative tweets. For example, the entities "*iPad*", "*iPod*" and "*Mac Book Pro*" appeared more often in tweets of positive polarity and they are all mapped to the semantic concept "*Product/Apple*". As a result, the tweet from the test set "*Finally, I got my iPhone. What a product!*" is more likely to have a positive polarity because it contains the entity "*iPhone*" which is also mapped to the concept "*Product/Apple*".

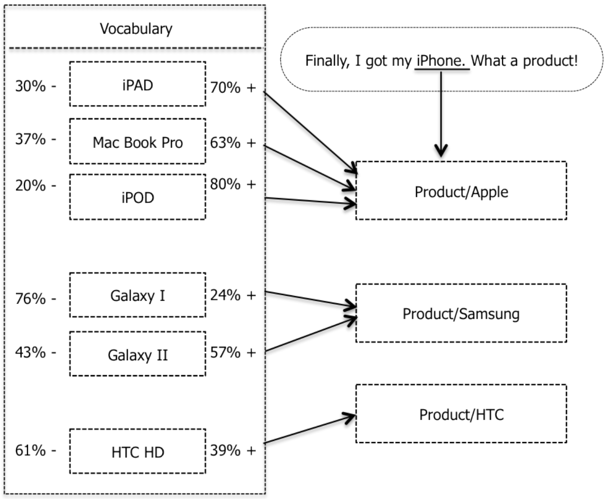


Figure 2: Incorporating semantic concepts for sentiment classification.

We propose a semantic interpolation method to incorporate semantic concepts into sentiment classifier training where we interpolate the original unigram language model in the Naïve Bayes (NB) classifier with the generative model of words given semantic concepts. We show on the Stanford Twitter Sentiment Data [5] that simply replaces words with their corresponding semantic concepts reduces the vocabulary size by nearly 20%. However, the sentiment classification accuracy drops by 4% compared to the baseline NB model trained on unigrams solely. With the interpolation method, the sentiment classification accuracy improves upon the baseline model by nearly 4%.

Our second approach for automatic word clustering is through sentiment-topics extraction using the previously proposed joint sentiment-topic (JST) model [11]. The JST model extracts latent topics and the associated topic sentiment from the tweets data which are subsequently added into the original feature space for supervised classifier training. Our experimental results show that NB learned from these features outperforms the baseline model trained on unigrams only and achieves the state-of-the-art result on the original test set of the Stanford Twitter Sentiment Data.

The rest of the paper is organised as follows. Section 2 outlines existing work on sentiment analysis with focus on twitter sentiment analysis. Section 3 describes the data used in our experiments. Section 4 presents our proposed semantic smoothing method. Section 5 describes how we incorporate sentiment-topics extracted from the JST model into sentiment classifier training. Experimental results are discussed in Section 6. Finally, we conclude our work and outline future directions in Section 7.

2. RELATED WORK

Much work has been done in the field of sentiment analysis. Most of the work follows two basic approaches. The first approach assumes that semantic orientation of a document is an averaged sum of the semantic orientations of its words and phrases. The pioneer work is the point-wise mutual information approach proposed in Turney [20]. Also work such as [6, 8, 19, 16] are good examples of this lexical-based approach. The second approach [15, 14, 4, 23, 12] addresses the problem as a text classification task where classifiers are built using one of the machine learning methods and trained on a dataset using features such as unigrams, bigrams, part-of-speech (POS) tags, etc. The vast majority of work in sentiment analysis mainly focuses on the domains of movie reviews, product reviews and blogs.

Twitter sentiment analysis is considered as a much harder problem than sentiment analysis on conventional text such as review documents, mainly due to the short length of tweet messages, the frequent use of informal and irregular words, and the rapid evolution of language in Twitter. Annotated tweets data are impractical to obtain. A large amount of work have been conducted on twitter sentiment analysis using noisy labels (also called distant supervision). For example, Go et al. [5] used emoticons such as ":-)" and ":(" to label tweets as positive or negative and train standard classifiers such as Naïve Bayes (NB), Maximum Entropy (MaxEnt), and Support Vector Machines (SVMs) to detect the sentiments of tweets. The best result of 83% was reported by MaxEnt using a combination of unigrams and bigrams. Barbosa and Feng [2] collected their training data from three different Twitter sentiment detection websites which mainly use some pre-built sentiment lexicons to label each tweet as positive or negative. Using SVMs trained from these noisy labeled data, they obtained 81.3% in sentiment classification accuracy.

While the aforementioned approaches did not detect neutral sentiment, Pak and Paroubek [13] additionally collected neutral tweets from Twitter accounts of various newspapers and magazines and trained a three-class NB classifier which is able to detect neutral tweets in addition to positive and negative tweets. Their NB was trained with a combination of n-grams and POS features.

Speriosu et al. [18] argued that using noisy sentiment labels may hinder the performance of sentiment classifiers. They proposed exploiting the Twitter follower graph to improve sentiment classification and constructed a graph that has users, tweets, word unigrams, word bigrams, hashtags, and emoticons as its nodes which are connected based on the link existence among them (e.g., users are connected to tweets they created; tweets are connected to word unigrams that they contain etc.). They then applied a label propagation method where sentiment labels were propagated from a small set of of nodes seeded with some initial label information throughout the graph. They claimed that their label propagation method outperforms MaxEnt trained from noisy labels and obtained an accuracy of 84.7% on the subset of the twitter sentiment test set from [5].

There have also been some work in exploring feature engineering to improve the performance of sentiment classification on tweets. Agarwal et al. [1] studied using the feature based model and the tree kernel based model for sentiment classification. They explored a total of 50 different feature types and showed that both the feature based and tree kernel based models perform similarly and they outperform the unigram baseline.

Kouloumpis et al. [10] compared various features including n-gram features, lexicon features based on the existence of polarity words from the MPQA subjectivity lexicon[2], POS features, and microblogging features capturing the presence of emoticons, abbreviations, and intensifiers (e.g., all-caps and character repetitions). They found that micoblogging features are most useful in sentiment classification.

3. TWITTER SENTIMENT CORPUS

In the work conducted in this paper, we used the Stanford Twitter Sentiment Data[3] which was collected between the 6th of April and the 25th of June 2009 [5]. The training set consists of 1.6 million tweets with the same number of positive and negative tweets labelled using emoticons. For example, a tweet is labelled as positive if it contains :), :-), :), :D, or =) and is labelled as negative if it has :(, :-(, or : (, etc. The original test set consists of 177 negative and 182 positive manually annotated tweets. In contrast to the training set which was collected based on specific emoticons, the test set was collected by searching Twitter API with specific queries including products' names, companies and people.

We built our training set by randomly selecting 60,000 balanced tweets from the original training set in the Stanford Twitter Sentiment Data. Since the original test set only contains a total of 359 tweets which is relatively small, we enlarge this set by manually annotating more tweets. To simplify and speed up the annotation efforts, we have built Tweenator[4], a web-based sentiment annotation tool that allows users to easily assign a sentiment label to tweet messages, i.e. assign a negative, positive or neutral label to a certain tweet with regards to its contextual polarity. Using Tweenator, 12 different users have annotated additional 641 tweets from the original remaining training data. Our final test set contains 1,000 tweet messages with 527 negative and 473 positive.

It is worth mentioning that users who participated in the annotation process have reported that using the annotation interface of Tweenator, as shown in Figure 3-a, they were able to annotate 10 tweet messages in 2 to 3 minutes approximately.

Recently, we have added two new modules to Tweenator by implementing our work that will be described in Section 4. The first module (see Figure 3-b) provides a free-form sentiment detection, which allows users to detect the polarity of their textual entries. The second module is the opinionated tweet message retrieval tool (see Figure 3-c) that allows to retrieve negative/positive tweets towards a specific search term. For example, a user can retrieve opinionated tweet messages about the search term "*Nike*".

4. SEMANTIC FEATURES

Twitter is an open social environment where there are no restrictions on what users can tweet about. Therefore, a huge number of infrequent named entities, such as people, organization, products, etc., can be found in tweet messages. These infrequent entities make the data very sparse and hence hinder the sentiment classification performance. Nevertheless, many of these named entities are semantically related. For example, the entities "*iPad*" and "*iPhone*" can be mapped to the same semantic concept "*Product/Apple*". Inspired by this observation, we propose using semantic features to alleviate the sparsity problem from tweets data. We first extract named entities from tweets and map them to their corresponding semantic concepts. We then incorporate these semantic concepts into NB classifier training.

4.1 Semantic Concept Extraction

We investigated three third-party services to extract entities from tweets data, Zemanta,[5] OpenCalais,[6] and AlchemyAPI.[7] A quick and manual comparison of a randomly selected 100 tweet messages with the extracted entities and their corresponding semantic concepts showed that AlchemyAPI performs better than the others in terms of the quality and the quantity of the extracted entities. Hence, we used AlchemyAPI for the extraction of semantic concepts in our paper.

Using AlchemyAPI, we extracted a total of 15,139 entities from the training set, which are mapped to 30 distinct concepts and extracted 329 entities from the test set, which are mapped to 18 distinct concepts. Table 1 shows the top five extracted concepts from the training data with the number of entities associated with them.

Concept	Number of Entities
Person	4954
Company	2815
City	1575
Country	961
Organisation	614

Table 1: Top 5 concepts with the number of their associated entities.

4.2 Incorporating Semantic Concepts into NB Training

The extracted semantic concepts can be incorporated into sentiment classifier training in a naive way where entities are simply replaced by their mapped semantic concepts in the tweets data. For example, all the entities such as "*iPhone*", "*iPad*", and "*iPod*" are replaced by the semantic concept "*Product/Apple*". A more principled way to incorporate semantic concepts is through interpolation. Here, we propose interpolating the unigram language model with the generative model of words given semantic concepts in NB training.

In NB, the assignment of a sentiment class c to a given tweet $\mathbf{w}$ can

[2] `http://www.cs.pitt.edu/mpqa/`

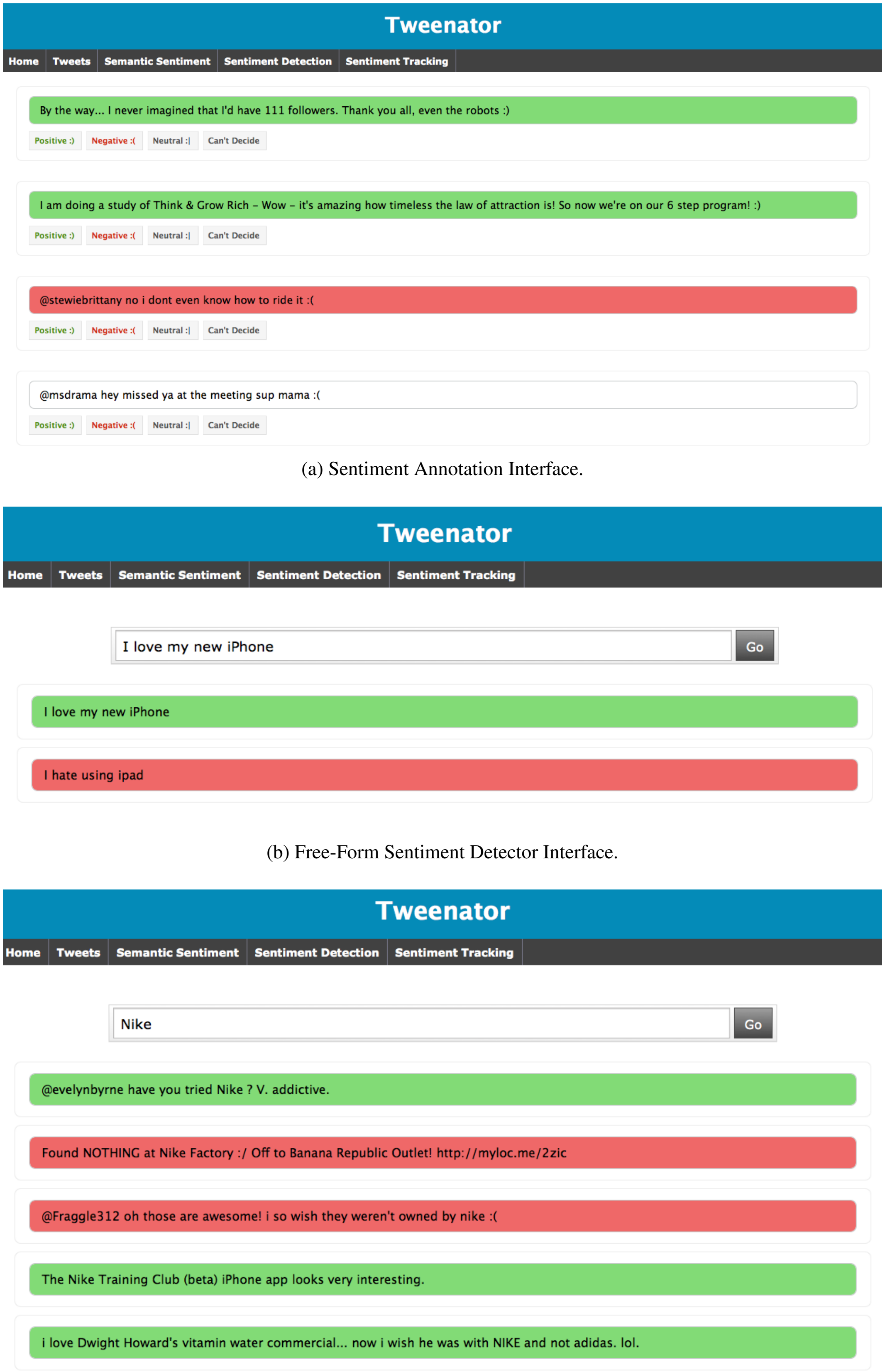
[3] `http://twittersentiment.appspot.com/`

[4] `http://atkmi.com/tweenator/`

[5] `http://www.zemanta.com/`

[6] `http://www.opencalais.com/`

[7] `http://www.alchemyapi.com/`



By the way... I never imagined that I'd have 111 followers. Thank you all, even the robots :)

I am doing a study of Think & Grow Rich – Wow – it's amazing how timeless the law of attraction is! So now we're on our 6 step program! :)

@stewiebrittany no i dont even know how to ride it :(

@msdrama hey missed ya at the meeting sup mama :(

(a) Sentiment Annotation Interface.

I love my new iPhone

I love my new iPhone

I hate using ipad

(b) Free-Form Sentiment Detector Interface.

Nike

@evelynbyrne have you tried Nike ? V. addictive.

Found NOTHING at Nike Factory :/ Off to Banana Republic Outlet! http://myloc.me/2zic

@Fraggle312 oh those are awesome! i so wish they weren't owned by nike :(

The Nike Training Club (beta) iPhone app looks very interesting.

i love Dwight Howard's vitamin water commercial... now i wish he was with NIKE and not adidas. lol.

(c) Opinionated Tweet Message Retrieval Interface.

Figure 3: Tweenator: Web based Sentiment Annotation Tool for Twitter

be computed as:

$$\begin{aligned}\hat{c} &= \arg\max_{c\in\mathcal{C}} P(c|\mathbf{w}) \\ &= \arg\max_{c\in\mathcal{C}} P(c) \prod_{1\leq i\leq N_{\mathbf{w}}} P(w_i|c), \qquad (1)\end{aligned}$$

where $N_{\mathbf{w}}$ is the total number of words in tweet $\mathbf{w}$, $P(c)$ is the prior probability of a tweet appearing in class c, $P(w_i|c)$ is the conditional probability of word w_i occurring in a tweet of class c.

In multinomial NB, $P(c)$ can be estimated by $P(c) = N_c/N$ Where N_c is the number of tweets in class c and N is the total number of tweets. $P(w_i|c)$ can be estimated using maximum likelihood with Laplace smoothing:

$$P(w|c) = \frac{N(w,c)+1}{\sum_{w'\in V} N(w'|c) + |V|} \qquad (2)$$

Where $N(w,c)$ is the occurrence frequency of word w in all training tweets of class c and $|V|$ is the number of words in the vocabulary. Although using Laplace smoothing helps to prevent zero probabilities of the "unseen" words, it assigns equal prior probabilities to all of these words.

We propose a new smoothing method where we interpolate the unigram language model in NB with the generative model of words given semantic concepts. Thus, the new class model with semantic smoothing has the following formula:

$$\begin{aligned}P_s(w|c) =& (1-\alpha)P_u(w|c) \\ &+ \alpha \sum_j P(w|s_j)P(s_j|c) \qquad (3)\end{aligned}$$

Where $P_s(w|c)$ is the unigram class model with semantic smoothing, $P_u(w|c)$ is the unigram class model with maximum likelihood estimate, s_j is the j-th concept of the word w, $P(s_j|c)$ is the distribution of semantic concepts in training data of a given class and it can computed via the maximum likelihood estimation. $P(w|s_j)$ is the distribution of words in the training data given a concept and it can be also computed via the maximum likelihood estimation. Finally, the coefficient α is used to control the influence of the semantic mapping in the new class model. By setting α to 0 the class model becomes a unigram language model without any semantic interpolation. On the other hand, setting α to 1 reduces the class model to a semantic mapping model. In this work, α was empirically set to 0.5.

5. SENTIMENT-TOPIC FEATURES

The joint sentiment-topic (JST) model [11] is a four-layer generative model which allows the detection of both sentiment and topic simultaneously from text. The generative procedure under JST boils down to three stages. First, one chooses a sentiment label l from the per-document sentiment distribution π_d. Following that, one chooses a topic z from the topic distribution $\theta_{d,l}$, where $\theta_{d,l}$ is conditioned on the sampled sentiment label l. Finally, one draws a word w_i from the per-corpus word distribution $\phi_{l,z}$ conditioned on both topic z and sentiment label l. The JST model does not require labelled documents for training. The only supervision is word prior polarity information which can be obtained from publicly available sentiment lexicons such as the MPQA subjectivity lexicon.

We train JST on the training set with tweet sentiment labels being discarded. The resulting model assigns each word in tweets with a sentiment label and a topic label. Hence, JST essentially clusters different words sharing similar sentiment and topic. We list some of the topic words extracted by JST in Table 2. Words in each cell are grouped under one topic and the upper half of the table shows topic words bearing positive sentiment while the lower half shows topic words bearing negative polarity. It can be observed that words groups under different sentiment and topic are quite informative and coherent. For example, Topic 3 under positive sentiment is related to a good music album, while Topic 1 under negative sentiment is about a complaint of feeling sick possibly due to cold and headache.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Positive	dream	bought	song	eat	movi
	sweet	short	listen	food	show
	train	hair	love	coffe	award
	angel	love	music	dinner	live
	love	wear	play	drink	night
	goodnight	shirt	album	yummi	mtv
	free	dress	band	chicken	concert
	club	photo	guitar	tea	vote
Negative	feel	miss	rain	exam	job
	today	sad	bike	school	hard
	hate	cry	car	week	find
	sick	girl	stop	tomorrow	hate
	cold	gonna	ride	luck	interview
	suck	talk	hit	suck	lost
	weather	bore	drive	final	kick
	headache	feel	run	studi	problem

Table 2: Extracted polarity words by JST.

Inspired by the above observations, grouping words under the same topic and bearing similar sentiment could potentially reduce data sparseness in twitter sentiment classification. Hence, we extract sentiment-topics from tweets data and augment them as additional features into the original feature space for NB training. Algorithm 1 shows how to perform NB training with sentiment-topics extracted from JST. The training set consists of labeled tweets, $\mathcal{D}^{train} = \{(\mathbf{w}_n; c_n) \in \mathcal{W} \times \mathcal{C} : 1 \leq n \leq N^{train}\}$, where $\mathcal{W}$ is the input space and $\mathcal{C}$ is a finite set of class labels. The test set contains tweets without labels, $\mathcal{D}^{test} = \{\mathbf{w}_n^t \in \mathcal{W} : 1 \leq n \leq N^{test}\}$. A JST model is first learned from the training set and then infer sentiment-topic for each tweet in the test set. The original tweets are augmented with those sentiment-topics as shown in Step 4 of Algorithm 1, where $l_i\_z_i$ denotes a combination of sentiment label l_i and topic z_i for word w_i. Finally, an optional feature selection step can be performed according to the information gain criteria and a classifier is then trained from the training set with the new feature representation.

6. EXPERIMENTAL RESULTS

In this section, we present the results obtained on the twitter sentiment data using both semantic features and sentiment-topic features and compare with the existing approaches.

6.1 Pre-processing

The raw tweets data are very noisy. There are a large number of irregular words and non-English characters. Tweets data have some unique characteristics which can be used to reduce the feature space through the following pre-processing:

Algorithm 1 NB training with sentiment-topics extracted from JST.

```
Input: The training set $\mathcal{D}^{train}$ and test set $\mathcal{D}^{test}$
Output: NB sentiment classifier
Train a JST model on $\mathcal{D}^{train}$ with the document labels discarded
Infer sentiment-topic from $\mathcal{D}^{test}$
for each tweet $\mathbf{w}_n = (w_1, w_2, ..., w_m) \in \{\mathcal{D}^{train}, \mathcal{D}^{test}\}$ do
    Augment tweet with sentiment-topics generated from JST,
    $\mathbf{w}'_n = (w_1, w_2, ..., w_m, l_1\_z_1, l_2\_z_2, ..., l_m\_z_m)$
end for
Create a new training set $\mathcal{D}^{train'} = \{(\mathbf{w}'_n; c_n) : 1 \leq n \leq N^{train}\}$
Create a new test set $\mathcal{D}^{test'} = \{\mathbf{w}'_n : 1 \leq n \leq N^{test}\}$
Perform feature selection using IG on $\mathcal{D}^{train'}$
Return NB trained on $\mathcal{D}^{train'}$
```

Pre-processing	Vocabulary Size	% of Reduction
None	95,130	0%
Username	70,804	25.58%
Hashtag	94,200	0.8%
URLS	92,363	2.91%
Repeated Letters	91,824	3.48%
Digits	92,785	2.47%
Symbols	37,054	29.47%
All	37,054	61.05%

Table 3: The effect of pre-processing.

- All Twitter usernames, which start with @ symbol, are replaced with the term "USER".
- All URL links in the corpus are replaced with the term "URL".
- Reduce the number of letters that are repeated more than twice in all words. For example the word "looooooveeee" becomes "loovee" after reduction.
- Remove all Twitter hashtags which start with the # symbol, all single characters and digits, and non-alphanumeric characters.

Table 3 shows the effect of pre-processing on reducing features from the original feature space. After all the pre-processing, the vocabulary size is reduced by 62%.

6.2 Semantic Features

We have tested both the NB classifier from WEKA[8] and the maximum entropy (MaxEnt) model from MALLET[9]. Our results show that NB consistently outperforms MaxEnt. Hence, we use NB as our baseline model. Table 4 shows that with NB trained from unigrams only, the sentiment classification accuracy of 80.7% was obtained.

We extracted semantic concepts from tweets data using Alchemy API and then incorporated them into NB training by the following two simple ways. One is to replace all entities in the tweets corpus with their corresponding semantic concepts (*semantic replacement*). Another is to augment the original feature space with semantic concepts as additional features for NB training (*semantic augmentation*). With *semantic replacement*, the feature space shrunk substantially by nearly 20%. However, sentiment classification accuracy drops by 4% compared to the baseline as shown in Table 4. The performance degradation can be explained as the mere use of semantic concepts replacement which leads to information loss and subsequently hurts NB performance. Augmenting the original feature space with semantic concepts performs slightly better than *sentiment replacement*, though it still performs worse than the baseline.

With *Semantic interpolation*, semantic concepts were incorporated into NB training taking into account the generative probability of words given concepts. The method improves upon the baseline model and gives a sentiment classification accuracy of 84%.

Method	Accuracy
Unigrams	80.7%
Semantic replacement	76.3%
Semantic augmentation	77.6%
Semantic interpolation	**84.0%**
Sentiment-topic features	82.3%

Table 4: Sentiment classification results on the 1000-tweet test set.

6.3 Sentiment-Topic Features

To run JST on the tweets data, the only parameter we need to set is the number of topics T. It is worth noting that the total number of the sentiment-topics that will be extracted is $3 \times T$. For example, when T is set to 50, there are 50 topics under each of positive, negative and neutral sentiment labels. Hence the total number of sentiment-topic features is 150. We augment the original bag-of-words representation of the tweet messages by the extracted sentiment-topics. Figure 4 shows the classification accuracy of NB trained from the augmented features by varying the number of topics from 1 to 65. The initial sentiment classification accuracy is 81.1% with topic number 1. Increasing the number of topics leads to the increase of classification accuracy with the peak value of 82.3% being reached at topic number 50. Further increasing topic numbers degrades the classifier performance.

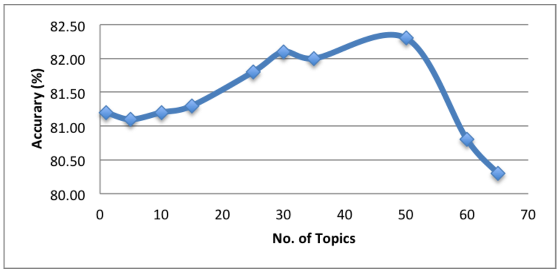


Figure 4: Classification accuracy vs. number of topics.

6.4 Comparison with Existing Approaches

In order to compare our proposed methods with the existing approaches, we also conducted experiments on the original Stanford Twitter Sentiment test set which consists of 177 negative and 182 positive tweets. The results are shown in Table 5. The sentiment classification accuracy of 83% reported in [5] was obtained using MaxEnt trained on a combination of unigrams and bigrams. It should be noted that while Go et al. used 1.6 million tweets for training, we only used a subset of 60,000 tweets as our training set.

[8] `http://www.cs.waikato.ac.nz/ml/weka/`
[9] `http://mallet.cs.umass.edu/`

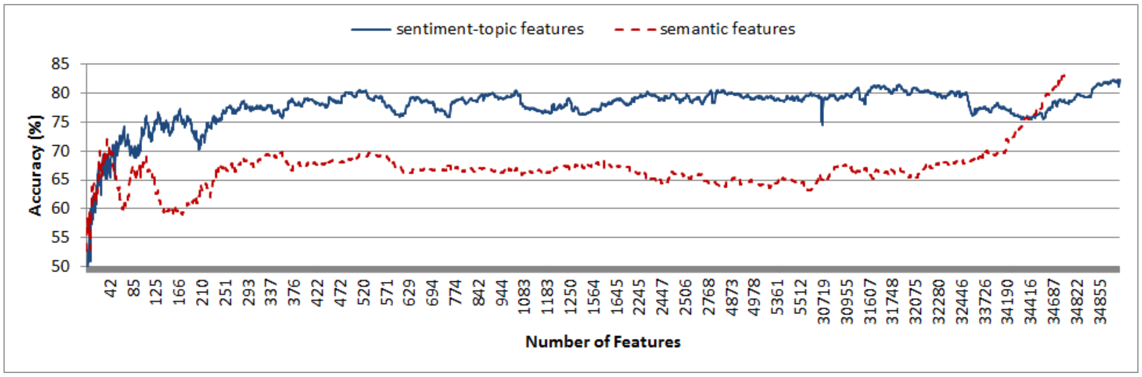


Figure 5: Classification accuracy vs. number of features selected by information gain.

Speriosu et al. [18] tested on a subset of the Stanford Twitter Sentiment test set with 75 negative and 108 positive tweets. They reported the best accuracy of 84.7% using label propagation on a rather complicated graph that has users, tweets, word unigrams, word bigrams, hashtags, and emoticons as its nodes.

It can be seen from Table 5 that *sentiment replacement* performs worse than the baseline. *Sentiment augmentation* does not result in the significant decrease of the classification accuracy, though it does not lead to the improved performance either. Our *semantic interpolation* method rivals the best result reported on the Stanford Twitter Sentiment test set. Using the sentiment-topic features, we achieved 86.3% sentiment classification accuracy, which outperforms the existing approaches.

Method	Accuracy
Unigrams	81.0%
Semantic replacement	77.3%
Semantic augmentation	80.45%
Semantic interpolation	84.1%
Sentiment-topic features	**86.3%**
(Go et al., 2009)	83%
(Speriosu et al., 2011)	84.7%

Table 5: Sentiment classification results on the original Stanford Twitter Sentiment test set.

6.5 Discussion

We have explored incorporating semantic features and sentiment-topic features for twitter sentiment classification. While simple *semantic replacement* or *augmentation* does not lead to the improvement of sentiment classification performance, *sentiment interpolation* improves upon the baseline NB model trained on unigrams only by 3%. Augmenting feature space with sentiment-topics generated from JST also results in the increase of sentiment classification accuracy compared to the baseline. On the original Stanford Twitter Sentiment test set, NB classifiers learned from sentiment-topic features outperform the existing approaches.

We have a somewhat contradictory observation here. Using sentiment-topic features performs worse than using semantic features on the test set comprising of 1000 tweets. But the reverse is observed on the original Stanford Twitter Sentiment test set with 359 tweets. We therefore conducted further experiments to compare these two approaches.

We performed feature selection using information gain (IG) on the training set. We calculated the IG value for each feature and sorted them in descending order based on IG. Using each distinct IG value as a threshold, we ended up with different sets of features to train a classifier. Figure 5 shows the sentiment classification accuracy on the 1000-tweet test set versus different number of features. It can be observed that there is an abrupt change in x-axis from around 5600 features jumping to over 30,000 features. Using sentiment-topic features consistently performs better than using semantic features. With as few as 500 features, augmenting the original feature space with sentiment-topics already achieves 80.2% accuracy. Although with all the features included, NB trained with semantic features performs better than that with sentiment-topic features, we can still draw a conclusion that sentiment-topic features should be preferred over semantic features for the sentiment classification task since it gives much better results with far less features.

7. CONCLUSIONS AND FUTURE WORK

Twitter is an open social environment where users can tweet about different topics within the 140-character limit. This poses a significant challenge to Twitter sentiment analysis since tweets data are often noisy and contain a large number of irregular words and non-English symbols and characters. Pre-processing by filtering some of the non-standard English words leads to a significant reduction of the original feature space by nearly 61.0% on the Twitter sentiment data. Nevertheless, the pre-processed tweets data still contain a large number of rare words.

In this paper, we have proposed two sets of features to alleviate the data sparsity problem in Twitter sentiment classification, semantic features and sentiment-topic features. Our experimental results on the Twitter sentiment data show that while both methods improve upon the baseline Naïve Bayes model trained from unigram features only, using sentiment-topic features gives much better results than using semantic features with less features.

Compared to the existing approaches to twitter sentiment analysis which either rely on sophisticated feature engineering or complicated learning procedure, our approaches are much more simple and straightforward and yet attain comparable performance.

There are a few possible directions we would like to explore as future work. First, in the semantic method all entities where simply replaced by the associated semantic concepts. It is worth to perform a selective statistical replacement, which is determined based on the contribution of each concept towards making a better classification decision. Second, sentiment-topics generated by JST model were simply augmented into the original feature space of tweets data. It could lead to better performance by attaching a weight to each extracted sentiment-topic feature in order to control the impact of the newly added features. Finally, the performance of the NB classifiers learned from semantic features depends on the quality of the entity extraction process and entity-concept mapping method. It is worth to investigate a filtering method which can automatically filter out low-confidence semantic concepts.

Acknowledgement This work is partially funded by the EU project ROBUST (grant number 257859).

8. REFERENCES

[1] AGARWAL, A., XIE, B., VOVSHA, I., RAMBOW, O., AND PASSONNEAU, R. Sentiment analysis of twitter data. In *Proceedings of the ACL 2011 Workshop on Languages in Social Media* (2011), pp. 30–38.

[2] BARBOSA, L., AND FENG, J. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of COLING* (2010), pp. 36–44.

[3] BHUIYAN, S. Social media and its effectiveness in the political reform movement in egypt. *Middle East Media Educator 1*, 1 (2011), 14–20.

[4] BOIY, E., HENS, P., DESCHACHT, K., AND MOENS, M. Automatic sentiment analysis in on-line text. In *Proceedings of the 11th International Conference on Electronic Publishing* (2007), pp. 349–360.

[5] GO, A., BHAYANI, R., AND HUANG, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* (2009).

[6] HATZIVASSILOGLOU, V., AND WIEBE, J. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1* (2000), Association for Computational Linguistics, pp. 299–305.

[7] HE, Y., AND SAIF, H. Quantising Opinons for Political Tweets Analysis. In *Proceeding of the The eighth international conference on Language Resources and Evaluation (LREC) - In Submission* (2012).

[8] HU, M., AND LIU, B. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), ACM, pp. 168–177.

[9] HUSSAIN, M., AND HOWARD, P. the role of digital media. *Journal of Democracy 22*, 3 (2011), 35–48.

[10] KOULOUMPIS, E., WILSON, T., AND MOORE, J. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the ICWSM* (2011).

[11] LIN, C., AND HE, Y. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 375–384.

[12] NARAYANAN, R., LIU, B., AND CHOUDHARY, A. Sentiment Analysis of Conditional Sentences. In *EMNLP* (2009), pp. 180–189.

[13] PAK, A., AND PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC 2010* (2010).

[14] PANG, B., AND LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004), Association for Computational Linguistics, p. 271.

[15] PANG, B., LEE, L., AND VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (2002), Association for Computational Linguistics, pp. 79–86.

[16] READ, J., AND CARROLL, J. Weakly supervised techniques for domain-independent sentiment classification. In *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (2009), pp. 45–52.

[17] SAIF, H., HE, Y., AND ALANI, H. Semantic Smoothing for Twitter Sentiment Analysis. In *Proceeding of the 10th International Semantic Web Conference (ISWC)* (2011).

[18] SPERIOSU, M., SUDAN, N., UPADHYAY, S., AND BALDRIDGE, J. Twitter polarity classification with label propagation over lexical links and the follower graph. *Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP* (2011), 53–63.

[19] TABOADA, M., AND GRIEVE, J. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)* (2004), pp. 158–161.

[20] TURNEY, P. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)* (2002).

[21] WARD, J., AND OSTROM, A. The internet as information minefield:: An analysis of the source and content of brand information yielded by net searches. *Journal of Business research 56*, 11 (2003), 907–914.

[22] YOON, E., GUFFEY, H., AND KIJEWSKI, V. The effects of information and company reputation on intentions to buy a business service. *Journal of Business Research 27*, 3 (1993), 215–228.

[23] ZHAO, J., LIU, K., AND WANG, G. Adding redundant features for CRFs-based sentence sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2008), pp. 117–126.

Small talk in the Digital Age: Making Sense of Phatic Posts

Danica Radovanovic
University of Belgrade
danica.radovanovic@gmail.com

Massimo Ragnedda
University of Sassari
ragnedda@gmail.com

ABSTRACT

This paper presents some practical implications of a theoretical web desktop analysis and addresses microposts in the Social Web contextual sense and their role contributing diverse information to the Web as part of informal and semi-formal communication and social activities on Social Networking Sites (SNS). We reflect upon and present the most pervasive and relevant socio-communication function of an online presence on microposts and social networks: the phatic communication function. Although some theorists such as Malinowski say these microposts have no practical information value, we argue that they have semantic and social value for the interlocutors, determined by socio-technological and cultural factors such as online presence and social awareness. We investigate and offer new implications for emerging social and communication dynamics formed around microposts, what we call here "phatic posts". We suggest that apparently trivial uses and features of SNS actually play an important role in setting the social and informational context of the rest of the conversation - a "phatic" function - and thus that these phatic posts are key to the success of SNS.

General Terms

Internet, Communication, Theory

Keywords

social network sites, microposts, phatic posts, phatic communication, online communication, social dynamics

1. INTRODUCTION

This paper is a theoretical and implication study of the communicative and social function of microposts on social network sites (SNS). We do not present statistical or applications-driven data or suggest some pattern, but we do offer qualitative implications, theories, and better understanding of the current social paradigm. This paper is implications-driven research and presents the relevance of microposts and phatic posts as derivatives of phatic communication, a term coined by Malinowski to describe the phenomenon of small talk. Phatic communication is "a type of speech in which ties of union are created by a mere exchange of words" and its purpose is to establish and maintain the social bonds of the interlocutors (Malinowski 1923: 151). We describe the socio-technological and communication dynamics that influence the formation of micro phatic posts. Living in an accelerating, interconnected world of information where the demand for instant updates and news is present here and now, different forms of communication dynamics are formed, referring to the socio-technological communication processes online.

Different SNS provide an expressive medium to share with others our feelings, needs, current status, or simple statements. Those simple and short statements can carry light information or low information such as: "I'm eating a dark chocolate", or "listening to new album by Air", or just "life is beautiful". It can also provoke a communication: "anyone there?", "does anyone know...?", etc. On the other side there are applications driven by small micro posts (built by social networks) that enable the creation of phatic expressions in the form of microblogs, Facebook updates, signal indications of "like", "poke", Instant messenger signals in the form of emoticons and wide variety of smileys, etc.

The aim of this paper is to argue the social consequences about the new way of communication on the SNS. In particular we are evolving from the concept of phatic communion coming from the anthropologist Malinowski and from phatic function coming from the linguist Jakobson. These two concepts can relate with networked sociality, the non dialogic and non-informational discussion on the social networks. Although some theorists such as Malinowski say that phatic messages do not have a practical information value, we are arguing in this paper that they do have semantic and social value for the interlocutors, determined by socio-technological and cultural factors. We are using in this paper a new coined term for such micro-posts that imply in their content or form a phatic communication function: a term phatic-posts. This phenomenon can be characterized as "new-word", which is employed here to describe the fact that it is both new and a word. New-words are clearly evident in all human culture. The paper consists of three main parts. First we will discuss the origin of phatic communication and phatic culture and the way they are presented on social networks. Second, we will discuss motivations for creating and consuming phatic posts and their importance for everyday communication and socializing. Finally we consider "small - talk language" on SNS, its dynamics, functions, and relevance to microposts.

Microposts are a dominant form in both virtual habitats (social networks, virtual communities) and their mobile extensions and they are of socio-technological value. Social signaling and online presence are both communication determinants for creating microposts. We will conclude with a few examples, based on web desktop sphere analysis (Hine, 2005), personal web observations and qualitative analysis of microposts and the semantics of phatic communication. In some social, linguistics,

Published as part of the #MSM2012 Workshop proceedings,
available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

and semantic theories, phatic may indicate communication being mundane, information-less, without any value. We show in this paper that do contain information messages, signals, values of staying up-to-date with micro and macro world of events and news, flirt, chat, public expressions of everyday life and emotions among the participants.

2. MALINOWSKI AND JAKOBSON: THE ORIGIN OF "PHATIC POSTS"

Bronislaw Malinowski, an anthropologist who carried out a lot of research in ethnographical fields, introduces in his book "Coral Gardens and Their Magic'' two fundamental concepts for the study of language: context of situation and context of culture (1935: 73). He introduced three major ideas into his semantic theory: the first is related to the context of linguistic data; the second idea concerns the range of meaning and finally the third is that the context of situation may allow one to disambiguate sentences that are semantically unclear. All these three new ideas are important here. In particular it is interesting to underline the first and the last one. In the first one Malinowski clearly said that the real linguistic fact is the full countenance within its context of situation and in the last one that it is the context of situation that permits one to understand ambiguous sentences.

In this paper we are arguing that the origin of modern, social web micro posts (tweets, Facebook status updates, likes, pokes, geo-check-ins on Foursquare, Flickr comments, etc.) – which we call here "*phatic posts*" - have their origins in the human need for phatic communication, i.e. communication for social upkeep. The quality of the information being communicated has no practical value and is rather mundane and comes from Malinowski's concept of phatic communion. In particular phatic communion has three phatic functions: a social function to establish and maintain social connections; a communicative function to demonstrate that the channel of communication is open and present oneself as a potential communication partner; a validation and recognition function to indicate recognition of one's interlocutor as a potential communicative partner. To these three main functions, Philip Riley has added another three functions: to provide indexical information for social categorization (that is to signal different aspects of social identity); to negotiate the relationship, in particular relative status, roles and affectivity (which clearly could be seen operating if we look at the various forms of greetings and address that some individuals use according to his or her social or affective relationship with the interlocutor); to reinforce social structure (Riley 2007: 131-32).

Another important concept useful to better understand phatic culture (Miller 2008) and its social implication in everyday life is the term "phatic function" coined by Roman Jakobson. As is well known, Jakobson included the metalinguistic (verifying the code), as one of five general functions of language, along with: Emotive (expressing the sender's state); Conative (inciting the receiver's response); Phatic (tries to maintain contact with the receiver); Referential (relating to a context); and Poetic (existing as a construct for its own sake). Clearly depending upon the meaning of a particular speech act, one of these functions will come to prevail while the others remain subordinate. In particular in our discussion we are arguing about the phatic function of online communication in the context of this theoretical framework and we are going to discuss why the phatic function that tries to maintain contact with the receiver is important on SNS for maintaining and strengthening existing relationships. This is more evident in the case of Facebook where its primary purpose is to re-establish relationships lost in time, such as those between former classmates or older friends.

3. MOTIVATIONS FOR CREATING AND CONSUMING PHATIC POSTS

At this point we are explaining why phatic communication practices are useful. Beside the demands of constant online connected presence in an increasingly networked world, we are exploring motivations why phatic communication is being supported, encouraged, and practiced by social media services. The importance of phatic communication has already been recognized by software engineers defining protocols for use in messaging. Notably, the SIP (Session Initiation Protocol) and SIMPLE (Session Initiation Protocol for Instant Messaging and Presence Leveraging Extensions) protocols draw extensively on the idea of "presence" as a signal to networks of users that communication is possible and of the disposition of other users to communicate. In some senses, phatic posts up-level the same principle.

However, we are focusing on the phatic function. It is crucial because what really counts in human interaction is to stay in touch and let others know that "I'm here too". To do this participants just write "nonsense", expressing their thoughts freely and making witty comments. This apparently "nonsense writing", has an intimate purpose, not so much in what has been written, but keeping in contact and reinforcing relationship. For example, Twitter, beside micro-blogging, implies social networking, interacting, text messaging, learning, enabling communication both through the internet and mobile devices. These communications are designed to be read as soon as they are sent; essentially they are updates creating the notion and feeling of intimacy by being constantly connected online, in real time with others, globally. These practices have resulted in forming 'phatic media' (Miller 2008) in which communication without content has taken precedence. Indeed, these phatic messages tend to reinforce existing relationships and facilitate further relation without giving information or adding to the messages.

For many users, the point of Twitter is the maintenance of connected presence, very similar to saying "what's up?" in an analogue space as you pass someone on the street when you have no intention of finding out what is actually going on. The phatic function is communication practice that simply indicates the possibility that communication may occur.

Furthermore, one of the contemporary digital media scholars, Mizuko Ito, described the appearance of phatic communication processes among Japanese teens in "low-content text message" groups, whose purpose is simply to stay in contact with others. These mundane communication exchanges represent the kind of communication that arises among people who are overwhelmed with other forms of communication. For example, in Japanese culture, phatic function is called *aizuchi*. Aizuchi tweets are real time, continuous, two-sided communication where, if one drops out of the communication thread, the dynamics of the aizuchi is lost. Aizuchi also involves very short expressions of approval or disapproval and expressions and connotations of someone's online presence. Aizuchi has a social function: to keep connectedness with others. The stage of connectedness is always characterized by a very high degree of alertness. We have conducted a set of interviews[1], including an interview with

[1] Radovanovic, D, Qualitative research, set of semi-structured interviews (N -31) from 2010 to 2012.

Takashi Ota, Japanese software developer and Wikimedian, in order to clarify aizuchi. One can assume aizuchi as a sign of the confirmation of presence: "I'm listening", "it's your turn", "I won't interrupt you" or "you're expected to keep talking". This is a typical effect when interlocutors use aizuchi during direct conversation or phone calls, but it may be applied to online conversations as well. When being used online, aizuchi “makes you think as if your counterpart is talking in front of you. It makes you feel we are connected”.

All those examples, again, show how the phatic function is fundamental in SNS because the aim is to maintain and reinforce relationship: This is why Twitter and Facebook are the virtual realms of constant connections, sharing and relationships between people, interactive playgrounds where the phatic function is really important, if not fundamental. Online awareness streams that indicate online presence *are* incredibly good at providing phatic communication. Phatic function being the language we use for the purpose of being social, not so much for sharing information or ideas, though these two are not excluded: it is in the virtual communication ‘what’s up?’ or ‘how’re you doing?”

Our ancestors used to check in at different places, using chalk, pieces of wood and stones to signal their presence or potential danger to their community, in order to establish social contact in everyday life. Computer-mediated and mobile-mediated environments today provide the channel of communication to be open and to present oneself as a potential communication partner. Pokes, likes, signals, phatic posts and other small, micro-symbols indicate the recognition of one’s interlocutor (presence and validation) as a potential communicative partner. Once the connection is established, there are a variety of communicative processes happening on the walls of SNS profiles with the important consequence of keeping social and communication dynamics alive.

4. THE DYNAMICS OF PHATIC POSTS

Facebook exists to make the world “more and more connected”, and by that it encourages, among other dynamics, the phatic function of interaction and communication through sociable applications, games, and add-ons. For example: the basic two phatic expression functions are the “Like” button and “Poke”. Here a couple of examples coming from qualitative research on a social network. David, (engineer, 50) talking about the Poke function said: *“I have a few people I have been exchanging "pokes" with for ever - in most cases I have no recollection who started it! They simply mean "I was online and thought of you".”*Another example comes from Corky (programmer, 39): *“I respond to pokes, but I very rarely initiate them. I saw a post once that said "'Like' buttons mean "I like your post, but I am far too lazy or not interested enough to make an actual comment, or in a hurry" - I think poking is similar. I am thinking of you, or I noticed your profile photo in my feed or whatever, and I poked you to let you know you crossed my mind, but I'm far too lazy, or uninterested or busy to take the time to write a message. "*

Communicative dynamics established with the web 2.0 paradigm shift and the development of microblogging culture and the usage of social media and SNS using mobile communication, encouraged users to practice in everyday life what we call here: *a phatic display of connected presence.* This phatic display of a connected presence is expressed through microposts, comments, short messages, leet-speak, tweets, status updates, Facebook social add-ons, and embedded applications. All these forms have elements of communicative discourse enabling users to get socially engaged through brief, non-formal messages that have meaning and within their context denote something: interaction, connected presence and fostering and maintaining connections.

Human relationships depend more and more on new technologies, such as computers, mobile phones and, most relevantly here, on their social network identities. These enable us to interact with others and human relationships in new interconnected virtual habitats become increasingly dependant on these objects. This “dependency” creates a new sociability pattern of being constantly online and present and of relationships becoming a fluid ever-changing continuum. These new technologies enable the exchange of communication practices that we call here ‘phatic expressions’: phatic posts that enable creating, fostering and sustaining relationships and social interaction through non formal conversations, online presence and intimacy. Some researchers like Licoppe and Smoreda (2005) indicated that non formal and non-dialogic means of interaction had helped the emergence of small communicative processes and gestures whose purpose at the first glance may appear to lack meaningful information, but in its substance those gestures and communication expressions foster sociability and maintain social connections. As we showed earlier in the paper, these are communicative processes Malinowski described as phatic communion. Phatic expressions in communication practices are very meaningful because they indicate and imply social recognition, online intimacy and sociability in online communities. Phatic posts potentially denote a lot more substance and weight to them than the content itself suggests.

Coming back to the phatic function postulated by Jakobson we can add a new function particularly present, on the social networks: conflict avoiding. On Facebook the two most popular forms of phatic communication on which we want to focus - besides status updates - are the concept of like and poke. This last form seems very interesting because Facebook has a "Like" button and not an “I don’t like” button. This is because it seems to be much easier to maintain balance in a community if one establishes relationships of mutual conformistic harmony with other people and it could create a conflictual relationship, reducing interaction (someone could be unfriended) and reducing the total number of the users. Iacchetti, Altafini and Iacono (2011: 1) have based their theory on the “Balance Theory” a motivational theory of attitude change, proposed by Fritz Heider, (1958) whose work was related to the Gestalt school. This theory tends to study the origin and the structure of tensions and conflicts in a network of individuals whose mutual relationships are characterized in terms of friendship and hostility. Furthermore this theory, using a mathematical model, shows how on a social network the users tend to be more conformist and that clearly shows how stressful situations from a social perspective tend to be avoided. In fact they show how a “balanced relation” is more valuable than an unbalanced relation that tends to generate frustration. Therefore, by using phatic function, such as keeping in touch or performing light conversations, we are avoiding contrast and conflict, and the social and communication tensions are weakening, excluding whose who would disturb the structure of the social network. In this way phatic communicative practices are useful, because they allow the members of the SNS to be involved in the discussion, sometimes without having anything to say, just by clicking on “Like” to say “I agree with you”. The Facebook feature “poke” offers the same situation. Facebook defines poke as a social utility that connects you with the people around you. Radovanovic (2008) indicated that in social networking terms, poke is contextual, and the context of poke is dependent upon the current level of familiarity between the ‘poker’ and the ‘pokee’. It usually denotes an expression such as: “Hey, what’s up?” or “Look at me!", saying "Hi" to someone you already know well: "Hey, I'm here, online!" followed usually by a message or email. There are numerous possible meanings and interpretations behind the poke and in the context of social networking technologies they can include: a) showing romantic interest for the other; b) a high visibility, low pressure way of getting attention; c) a lightweight interaction.

Following a feature that is typical of participative web applications, trending topics on Twitter provide an insight into the different types of communication dynamics and practices. Through web observation of trending topic tweets we identified four types of phatic posts:

a) the first type of phatic posts implies short nodding, approval or disapproval using expressions like: yes, right, uhm, hm, lol, <3, smileys here when they are used as a message or a (hash)tag, and many other signs and expressions from leet-speak and everyday communications similar to aizuchi in Japanese.

b) the second type of phatic posts implies information about mundane everyday life in order to start up the conversation. Some may call it a pointless conversation form without any value. But looking below those pointless phatic posts one would realize that they contain an information value that actually carries a specific message. For example, a person who is just eating an ice-cream informs their audience of the type of food they are eating. If that person is a micro-celebrity it brings even bigger value to this information-micropost.

c) the third type of phatic posts indicates a secret language or an internal language especially between teens. Teens and young adults use a lot of phatic when communicating among themselves. They use it to protect their privacy and publicly express themselves through these short messages and posts – of which only they know the meaning – so that way they keep adults from their world. danah boyd (2010) wrote on this – decoding the youth and their "secret" language.

d) finally the fourth type of phatic posts is to indicate online connected presence. Also we can see that the phatic process has the function of displaying the other person's online presence, i.e. expressing that one is still "there". This is very indicative to young people, (Radovanovic 2010), who post from their mobile phones status updates in the evening after school, and look for their peers online. This is the function of online presence – to know that someone is out there. Phatic communication and online connection to the other becomes significant and phatic dialogue enables relationship maintenance as well as connected presence in social networks. This way the relevance of the phatic function of microposts is emerging as a form of online intimacy and of social connections in social networks.

5. CONCLUSION

The concepts of phatic communion coming from Malinowski and the phatic function theorized by Jakobson, are both concepts of real importance in this moment, giving us a fundamental theoretical framework on which to move to better understanding, and revealing implications for development and applications in the future. The phatic function comprises: a) social function; b) communicative function and c) validation and recognition function; d) to provide indexical information for social categorization; e) to negotiate the relationship, in particular relative status, roles and affectivity; f) to reinforce social structure. Furthermore, in relation with the social networks, we have added another function: conflict-avoidance. This one helps the social network to keep a balance and harmony, and diminishes the damage caused by conflicts. By using and stimulating new application of phatic communication and small-talk, tensions are weakened and the social network in which it is applied could be positively influenced. We thus believe the role of phatic posts deserves further scrutiny. It is clearly important to the success of SNS, and has analogues in the underlying protocols used by communication technology. We expect there will be re-usable patterns that system designers can use to ensure channels for phatic communication are available. There is clearly much opportunity for further investigations and research, since we anticipate that the role of phatic communications is inherent in all human social communication and expect to find it implicated in any online communications system.

6. ACKNOWLEDGMENTS

Thanks to Dr Bernie Hogan, Oxford Internet Institute, Simon Phipps of the Open Source Initiative, and Dejan Petrovic, BMm http://dejanseo.com.au/dan/

7. REFERENCES

[1] boyd, danah. (2010). "Living Life in Public: Why American Teens Choose Publicity Over Privacy." AOIR 2010. Gothenburg, Sweden, October 23. URL: http://www.danah.org/papers/talks/2010/AOIR2010.html.

[2] Giuseppe Facchetti, Giovanni Iacono, Claudio Altarini, *Computing global structural balance in large-scale signed social networks*, National Academy of Sciences of the United States of America, http://people.sissa.it/~altafini/papers/FaIaAl11/FaIaAl11.pdf

[3] Heider, Fritz (1958). The Psychology of Interpersonal Relations. John Wiley & Sons.

[4] Hine, Christine. (Ed.) (2005), Virtual Methods: Issues in Social Research on the Internet, Oxford: Berg Publishers.

[5] Jakobson, Roman. (1981) "Linguistics and Poetics." Poetry of Grammar and Grammar of Poetry. Vol. 3 of Selected Writings. 7 Vols. The Hague: Mouton. 18-51.

[6] Licoppe, C., & Smoreda, Z. (2005). Are social networks technologically embedded? How networks are changing today with changes in communication technology. Social Networks, 27 (4), 317-335

[7] Miller, Vincent. (2008), New Media, Networking and Phatic Culture, Convergence: The International Journal of Research into New Media Technologies, Vol. 14, No. 4, 387-400

[8] Malinowski, B. 1923. 'The Problem of Meaning in primitive languages. In C. K. Ogden & I. A. Richards (Eds), The meaning of meaning (pp. 146-152). London: Routledge & Kegan Paul

[9] Malinowski, B. (1935). Coral gardens and their magic, 2 vols. Allen & Unwin.

[10] Radovanovic, Danica (2008). Digital Serendipities, blog post: "Poke me, poke you back: Facebook social networking context"http://www.danicar.org/2008/02/15/poke-me-poke-you-back-facebook-social-networking-context/

[11] Radovanovic, Danica, Интернет парадигма, структура и динамика онлајн друштвених мрежа: Фејсбук и млади у Србији (Internet Paradigm, Structure, and Dynamics of Online Social Networks: Facebook and Young Adults in Serbia) (June 20, 2010). Pancevacko citaliste, Vol. III, No.17, pp. 20-26, 2010. Available at SSRN: http://ssrn.com/abstract=1986066

[12] Riley, Philip (2007) Language, culture and identity: an ethnolinguistic perspective, London: Continuum.

Exploiting Twitter's Collective Knowledge for Music Recommendations[*]

Eva Zangerle, Wolfgang Gassler, Günther Specht
Databases and Information Systems, Institute of Computer Science
University of Innsbruck, Austria
{firstname.lastname}@uibk.ac.at

ABSTRACT

Twitter is the largest source of public opinion and also contains a vast amount of information about its users' music favors or listening behaviour. However, this source has not been exploited for the recommendation of music yet. In this paper, we present how Twitter can be facilitated for the creation of a data set upon which music recommendations can be computed. The data set is based on microposts which were automatically generated by music player software or posted by users and may also contain further information about audio tracks.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Performance, Human Factors, Experimentation

Keywords

Recommender Systems, Music Recommendation, Twitter

1. INTRODUCTION

Throughout the last years, music recommendation services have become very popular in both academia and industry. The goal of such services is the recommendation of suitable music for a certain user. This is traditionally accomplished by (i) either taking the user profile consisting of the tracks the user listened to in the past and (if available) the user's rating for songs into account or (ii) analysing the song itself and using the extracted features in order to find similar songs. For the recommendation of music, huge corpora and user profiles are required as there are millions of different audio tracks. There are some large services, such as last.fm[1], which own such big corpora. However, most of them are not publicly available. Especially for academic purposes, only few (mostly small) data sets for the evaluation of the proposed approaches are available, like e.g. the million song data set [4].

Twitter is a publicly available service, which holds huge amounts of data and is still growing tremendously. Twitter stated that there are about 140 million new messages a day. Such messages can also be exploited in the context of music recommendations. Many audio players offer the functionality of automatically posting a tweet containing the title and artist of the track the user currently is listening to. These tweets traditionally contain keywords like `nowplaying` or `listeningto`, like e.g. in the tweet "`#nowplaying Tom Waits-Temptation`". For users who frequently make use of such a service, the set of these tweets can be seen as a user profile in terms of her musical preferences and provide well suited data for e.g. a music recommendation corpus.

In this paper we present an approach for gathering such data and refining it such that the tweeted artists and tracks can directly be related to the free music databases FreeDB and MusicBrainz. As a use case scenario, we present the recommendation of music based on the data set.

This paper is structured as follows. Section 2 describes the processes underlying the creation of the proposed data set. Section 3 features the approach for the recommendation of suitable music tracks as a use case for the gathered data. Section 4 contains related work and Section 5 concludes the paper and discusses future work.

2. DATA SET CREATION

The goal of this approach is the creation of a corpus of music tracks gathered from tweets of users. These tweets contain tracks the user previously listened to and tweeted about (the so-called user stream). In particular, we propose to make use of tweets which have been posted by users or audio players and contain the title and artist of the music track currently played, like e.g. "`#NowPlaying Best Thing I Never Had by Beyonce`". The following sections describe the steps taken for the creation of the data set.

2.1 Crawling of Twitter Data Set and Analysis

The data set was crawled via the Twitter Streaming API-between July 2011 and February 2012. The only publicly available access method is the Spritzer access which only provides real-time access to about 1% of all posted Twitter

[*] This research was partially funded by the University of Innsbruck (Nachwuchsförderung 2011).

[1] http://www.last.fm

Published as part of the #MSM2012 Workshop proceedings,
available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

messages. Due to these restrictions, we crawled 4,734,014 tweets containing one of the keywords `nowplaying`, `listento` or `listeningto` posted by 864,736 different users. This implies an average of 5.5 tweets for each user. Within our data set, the distribution of tweets per user resembles a longtail distribution, as can be seen in Table 1. Such a distribution implies that considering the fact that recommendations can only be made if a user has posted about two or more tracks, a total of 457,675 users and the respective tweets can not be facilitated for our approach as only one tweet of these users is featured within the data set.

Tweets in stream	Users
1	457,675
> 3	196,422
> 5	126,783
> 10	63,017
> 100	3,190
$> 1{,}000$	253
$> 10{,}000$	5

Table 1: Population of User Streams

In total, 5,916,294 hashtags were used within the data set. Clearly due to our used search keywords the hashtags `#nowplaying` and `#listeningto` were the most prominent hashtags within the crawled data set. Also, general hashtags like e.g. `#music`, `#radio` or `#video` have been used frequently. Music streaming services or online radios also make use of hashtags when tweeting about the currently playing track (e.g. `#cityfm` or `#fizy`).

A total of 1,413,983 tweets (29.8% of the whole corpus) featured hyperlinks. An analysis of these URLs revealed that URLs are mostly used to point to music services like e.g. Youtube or Spotify, an online music streaming service. A large part of the hyperlinks lead to the website of the service which was used to post the track information on Twitter, like e.g. tweetmylast.fm or tinysong.com.

2.2 Resolution of Twittered Tracks

This task aims at parsing the gathered tweets and recognizing the artist name and track title mentioned in the tweet. Consider e.g. the tweet "`#NowPlaying Best Thing I Never Had by Beyonce`". For this tweet, we have to extract Beyonce as the artist and "Best Thing I Never Had" as the title of the audio track and match it with a reference music database. Most of the crawled messages are very noisy and consist of many terms which are not concerned with the music track itself. Considering e.g. the tweet "`listening to Hey Hey My My (Out Of The Blue) by Neil Young on @Grooveshark: #nowplaying #musicmonday http://t.co-/7os3eeA`" which contains further information about the online radio service, a URL and other information which are not related to the music track. Especially when dealing with such noisy tweets, the matching is a crucial task as the quality of the data resulting from this step significantly influences the quality of the resulting recommendations.

2.2.1 *Resolution Approach*

As a reference database for artists and the according tracks, we made use of the publicly available databases FreeDB[2] and MusicBrainz[3]. FreeDB contains information about more than 37 million audio tracks, roughly 3,000,000 discs and 766,909 different artists. MusicBrainz was also considered as a reference database as we expected it to be of higher quality than FreeDB. MusicBrainz contains about 8 million tracks of about 650,000 different artists.

The goal of this task is to assign each tweet a FreeDB and a MusicBrainz entry which represents the title and the according artist extracted from the tweet. We tackle this resolution task by making use of a Lucene fulltext index as it allows a simple matching of strings, namely the tweet and a certain FreeDB or MusicBrainz entry. The fulltext index is filled with a combined string containing both the artist and the title of all tracks within the reference databases.

In a next step, we query this fulltext index for each of the tweets within the data set in order to obtain the most suitable FreeDB/MusicBrainz candidates for the title and artist of the track. We then use the top-20 search results of Lucene as candidates for the assignment of tracks to the information mentioned in the according tweet. Lucene's ranking function is based on the term frequency/inverse document frequency measure (tf/idf). This measure is dependent on the length of the query which is not favourable in our approach as tweets contains a high degree of noise (e.g. URLs, feelings, smilies, etc.) which are not part of a track title but also part of the query (the tweet). Therefore, we implemented a bag-of-words similarity measure between the query and the documents contained within the Lucene index similar to the Jaccard similarity measure. Our proposed similarity measure is defined by the ratio between the size of the (term-) intersection of the query and the track and the number of terms contained in the track, as can be seen in Equation 1.

$$sim_{music}(tweet, track) = \frac{|tweet \cap track|}{|track|} \tag{1}$$

The advantage of such a measure is the independence of the length of the query and the reduced influence of the noise in tweets. Furthermore, as our goal is to find the best matching audio track for all given tweets, it is crucial that most terms within the track are matched. However, in the case of multiple search results having obtained an equal score, we still rely on the tf/idf values computed by Lucene. Our proposed score is used for a ranking of the Lucene search results. For each of the tweets, the track which obtained the highest score are assigned to the tweet. In order to be able to set a certain threshold for the scores of the matching entries later, we also store the computed sim_{music}-score.

2.2.2 *Evaluation of Resolution*

For the evaluation of the resolution and the comparison of FreeDB and MusicBrainz, we created a ground truth data set which consists of 100 tweets randomly chosen from the data set. Subsequently, we tried to assign matching tracks in the FreeDB and MusicBrainz databases manually. This task was done by the same person for both reference databases and also contains the resolution of abbreviations or mentions which link to the artist's Twitter account. For example the tweet `#nowplaying @Lloyd_YG ft. @LilTunechi - You` can be resolved to the two Twitter accounts *Lloyd-Young Goldie* and *Lil Wayne WEEZY F* and therefore to the MusicBrainz entry *Lil Wayne feat. Lloyd - You*. Having gathered all possible information from the tweet, the assigning person searched for matching tracks in the database.

[2] `http://www.FreeDB.org`

[3] `http://www.MusicBrainz.org`

If the artist or the title of the track were not directly recognizable in the tweet, single words are used to search the database and find matching artists or titles. We only considered tweets which were resolved to both the according track and artist. Tweets such as `Chris Duarte, famous blues musician - free videos here: http://t.co/UZMXaGQ #blues #guitar #music #roots #free #nowplaying #musicmonday` which only contain information about the artist were not counted as a match. However, such information is also very valuable as it describes the musical taste of a user. For our ground truth data set, we were able to manually assign 57 tracks of FreeDB and 59 tracks of MusicBrainz. This shows that the size of both data sets is similar, however the FreeDB data set is very noisy (typos, spelling errors and variations).

Subsequently we ran our automated Lucene based resolution process on the ground truth dataset using both reference databases (see details in Table 2). Considering a sim_{music}-score threshold of 0.8 we were able to resolve 73% of the ground truth correctly and had an error rate (false positives) of about 10% of all matched tracks. The high number of false positives using the FreeDB data set can be lead back to the noisy entries in FreeDB.

RefDB	Manually	Automated	False Pos.
MusicBrainz	59	43 (73%)	5 (10%)
FreeDB	57	31 (54%)	18 (36%)

Table 2: Resolution Ground Truth (100 tweets)

Due to these obtained results we used MusicBrainz for all further computations (e.g. music recommendations).

3. MUSIC RECOMMENDATIONS

As a use case, we implemented a music recommendation service on top of the data set. The necessary steps for a recommendation of music are described in the following.

The proposed approach for the recommendation of music titles relies on the co-occurrence of titles within a user stream. Based on the obtained tweets and the assigned tracks, we propose to use association rules [2] in order to be able to model the co-occurrence of items efficiently. In the case of the co-occurrence of tweeted music titles, an association rule $t_1 \rightarrow t_2$ describes that a particular user who tweeted about song t_1 also tweeted about song t_2. These rules are the basis for the further recommendation process and are stored as triples $r = (t_1, t_2, c)$, where t_1 and t_2 are tracks which have been tweeted by the same user. c is a variable holding the popularity of the rule. Hence, such a rule denotes that track t_1 and track t_2 both have been listened by c users.

3.1 Ranking of Recommendation Candidates

In this step, the computed association rules are analysed and so-called recommendation candidates are extracted. Based on the rules, the recommended tracks for a certain user are computed by selecting a subset $\mathcal{C} \subseteq \mathcal{T}$ of track recommendation candidates by determining all rules which feature tracks occurring on the user stream. The final step for the recommendation of tracks is the ranking of the recommendation candidates within the set $\mathcal{C}$. Therefore, we make use of the count value c describing the popularity of a certain track within all association rules matching the tracks of the input user stream. Hence, all recommendation candidates are ranked by the respective count values where a higher count value results in a higher rank for the candidate.

3.2 Offline Evaluation

As a first evaluation we performed an offline evaluation and compared the computed track recommendations with recommendations provided by the last.fm API[4] which lists tracks similar to a given track including a score stating the relevance of the song (matching score).

We made use of the MusicBrainz data set as it contains cleaner data than FreeDB. Firstly, we removed all tweets of users who contributed only one tweet and which were matched with a MusicBrainz track with $sim_{music} < 0.8$ to dismiss uncertain mappings. Hence our final data set consisted of 2.5 million tweets of 525,751 users. Based on this data set we computed the according association rules and obtained 500 million distinct rules. Due to computability reasons and API limitations, we chose a subset consisting of the most popular tracks and according rules which are present more than 10 times ($c > 10$). The final data set consisted of 15,000 unique tracks and 90 million distinct rules.

We called the last.fm API for all tracks and the API was able to recognize 13,138 out of 15,000 songs. The API returned 3.2 million similar tracks which we matched with our internal MusicBrainz database. In total, 83% of all tracks with a score > 0.8 were matched. We transformed the gathered last.fm data to association rules and computed the overlap of rules with our rule set. 19% of the last.fm rules are covered by the Twitter-based rules. If we consider only similar tracks of last.fm with a matching score (gathered via the last.fm API) higher than 0.6, the twitter-based rules cover 79% of all rules in the set. When comparing the top-10 recommendations on both sides the coverage is only about 1% of all rules. These low numbers can be lead back to the restrictions of the Twitter API and the resulting sparse data set. Especially the incomplete user profiles decrease the coverage. E.g. within the "taste" subset of the million song data set roughly 70% of the tracks were played more than 10 times. In contrast, in our data set only 5% of the tweets were contained more than 10 times. This fact strengthens the evidence that the crawled data set is not representative enough which can be lead back to the API limitation and uncertainties in the matching processes. Furthermore, due to the diversity of music tracks, such an offline evaluation may not reveal the full potential of the approach. Online evaluations may achieve better results for our proposed approach and are subject to future work.

4. RELATED WORK

Research related to the presented approach can be categorized into (i) approaches dealing with recommendations either for Twitter or based on tweets and (ii) approaches mainly dealing with the recommendation of music.

The utilization of a corpus of tweets for the recommendation of resources has been a popular research topic. For example the recommendation of suitable hashtags is discussed in [14]. Many approaches aim at the recommendation of users who might be interesting to follow, like e.g. in [7]. Such approaches are typically based on the social ties of a user (his followees and followers). There are also many ap-

[4] `http://www.last.fm/api`

proaches which exploit these ties to recommend resources, such as websites [6] or news [12].

As for the second category of related work, the recommendation of music, many different approaches have been presented. Celma [5] provides an overview about this topic. Within Recommender Systems, in principle two major approaches are distinguished [1]: content-based recommendations and collaborative filtering (CF) approaches. Content-based recommendation systems aim at recommending resources which are similar to the resources the user already consumed or showed interest in. Collaborative filtering approaches aim at finding users with a profile similar to the current user in order to recommend items which these similar users also were in favor of. This categorization also holds within music recommendations. Content-based methods for music titles typically rely on the extraction and analysis of audio features. The presented approach relies on the second type as the computation of association rules based on user profiles can be assigned to the class of CF approaches.

However, for music recommendations also a third important aspect is exploited for the computation of recommendations: context. The notion of context has e.g. been defined by Schmidt et al. as being threefold: physical environment, human factors and time [13]. These three factors have all been addressed by music recommendation research. As for the physical environment of a user, e.g. Kaminskas and Ricci presented a location-aware approach for music recommendations [8]. The mood of users has been incorporated for the computation of recommendations in [9] and Baltrunas et al. [3] considered temporal facts when recommending music.

Many approaches exploited user profiles in social networks to recommend resources. Mesnage et al. [10] showed that people prefer the music that their friends in the social network prefer. The Serendip.me project[5] provides its users with music which is selected solely based on the Twitter ties (the followees) of the user. The dbrec project [11] is concerned with recommending music based on the DBPedia data set. In particular, the authors developed a distance metric for resources within DBPedia which enables the authors to recommend similar artists.

However, to the best of our knowledge there are no approaches concerned with the recommendation of music based on an analysis of “nowplaying” user streams on Twitter.

5. CONCLUSION AND FUTURE WORK

In this paper we showed that tweets can be exploited to build a corpus for music recommendations. The comparison with the recommendation service of last.fm showed that despite the sparse corpus due to Twitter’s API limitations, the coverage of last.fm’s recommendations is up to 79%. The results are very promising although the approach has to be enhanced to be usable in real-world recommendation environments. A mayor improvement would be the expansion of the data set as currently the corpus is very sparse and the user profiles are incomplete. Also, the matching task of noisy tweets deteriorates the quality of recommendations. This is due to the fact that many uncertain matching results have to be dismissed and hence, the size of the usable data corpus decreases. Future work also comprises the enhancement of the matching process by using metadata such as location, URLs or further sentiment analysis. Additionally, applying CF techniques for the exploitation of the social ties of the user are subject to future work. In order to evaluate the approach from a user’s point-of-view, online user tests are also part of the future work.

[5] `http://serendip.me`

6. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. of the 20th Intl. Conf on Very Large Data Bases*, pages 487–499, 1994.

[3] L. Baltrunas and X. Amatriain. Towards Time-Dependant Recommendation based on Implicit Feedback. *Workshop on ContextAware Recommender Systems CARS 2009 in ACM Recsys*, 2009:1–5.

[4] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The Million Song Dataset. In *Proc. of the 12th Intl. Conf. on Music Information Retrieval*, 2011.

[5] Ò. Celma. *Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 2010.

[6] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and Tweet: Experiments on Recommending Content from Information Streams. In *Proc. of the 28th Intl. conference on Human Factors in Computing Systems*, pages 1185–1194. ACM, 2010.

[7] J. Hannon, M. Bennett, and B. Smyth. Recommending Twitter Users to Follow using Content and Collaborative Filtering Approaches. In *Proc. of the 4th ACM Conf. on Recommender Systems*, pages 199–206. ACM, 2010.

[8] M. Kaminskas and F. Ricci. Location-Adapted Music Recommendation Using Tags. In *User Modeling, Adaption and Personalization 2011, Girona, Spain, July 11-15, 2011*, volume 6787 of *LNCS*, pages 183–194. Springer, 2011.

[9] J. Lee and J. Lee. Context Awareness by Case-based Reasoning in a Music Recommendation System. In *Proc. of the 4th Intl. Conference on Ubiquitous Computing Systems*, pages 45–58. Springer, 2007.

[10] C. Mesnage, A. Rafiq, S. Dixon, and R. Brixtel. Music Discovery with Social Networks. In *Proc. of the Workshop on Music Recommendation and Discovery 2011 in conjunction with ACM RecSys*, volume 793, pages 1–6. CEUR-WS, 2011.

[11] A. Passant. dbrec - Music Recommendations Using DBpedia. *The Semantic Web–ISWC 2010*, pages 209–224, 2010.

[12] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to Recommend Real-Time Topical News. In *Proc. of the third ACM conference on Recommender systems*, RecSys ’09, pages 385–388, New York, NY, USA, 2009. ACM.

[13] A. Schmidt, M. Beigl, and H. Gellersen. There is more to Context than Location. *Computers & Graphics*, 23(6):893–901, 1999.

[14] E. Zangerle, W. Gassler, and G. Specht. Using Tag Recommendations to Homogenize Folksonomies in Microblogging Environments. In *Social Informatics*, volume 6984 of *LNCS*, pages 113–126. Springer, 2011.

Extracting Unambiguous Keywords from Microposts Using Web and Query Logs Data

Davi de Castro Reis
Google Engineering
Belo Horizonte, Brazil
davi@google.com

Felipe Goldstein
Google Engineering
Paris, France
felipeg@google.com

Frederico Quintao
Google Engineering
Belo Horizonte, Brazil
quintao@google.com

If a lion could talk, we could not understand him.
(Ludwig Wittgenstein)

ABSTRACT

In the recent years, a new form of content type has become ubiquitous in the web. These are small and noisy text snippets, created by users of social networks such as Twitter and Facebook. The full interpretation of those microposts by machines impose tremendous challenges, since they strongly rely on context. In this paper we propose a task which is much simpler than full interpretation of microposts: we aim to build classification systems to detect keywords that unambiguously refer to a single dominant concept, even when taken out of context. For example, in the context of this task, *apple* would be classified as ambiguous whereas *microsoft* would not. The contribution of this work is twofold. First, we formalize this novel classification task that can be directly applied for extracting information from microposts. Second, we show how high precision classifiers for this problem can be built out of Web data and search engine logs, combining traditional information retrieval metrics, such as inverted document frequency, and new ones derived from search query logs. Finally, we have proposed and evaluated relevant applications for these classifiers, which were able to meet precision $\geq 72\%$ and recall $\geq 56\%$ on unambiguous keyword extraction from microposts. We also compare those results with closely related systems, none of which could outperform those numbers.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic Processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

General Terms

Algorithms, Experimentation

Published as part of the #MSM2012 Workshop proceedings,
available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

Keywords

unambiguity, word sense disambiguation, query logs, web, advertising

1. INTRODUCTION

The availability of huge Web based corpora has spawned a series of important advancements in the Natural Language Processing (NLP) field recently [2, 10]. Moreover, the increase of computational power has made it possible to run simpler algorithms over much more data [7]. However, computer interpretation of natural language is still an unsolved challenge. On the other hand, a very successful set of free text controlled applications have blossomed in the last decade: the Web search engines. The natural language processing techniques employed by these systems are not enough to give users a natural speaking experience, but regardless of that, users type queries in sites like Yahoo!, Bing and Google on a daily basis. Instead of teaching the machine how to speak like us, we ended up learning a simple yet effective way of expressing our information needs.

In this work we start from the premise that full understanding of natural language cannot be achieved with the current technology. In fact, not even a human being is able to fully understand all communication due to either lack of cultural context or to inherent ambiguity in the language. This is especially true in the context of microposts, where both the media culture and technical constraints impose a limit on the size of the messages. Given these limitations, we have aimed to detect parts of natural language that can be unambiguously understood in the lack of any context. The key observation that allowed us to identify those parts of natural language is that they tend to be used as search engine queries on the Web [17].

For example, when using a search engine, if the user wants to know about Michael Jackson, the American singer, dancer, and entertainer, she expects to type these two words in thesearch box and get relevant results. On the other side, if she is looking for Michael Jackson, the social anthropologist from New Zealand, she knows that she will need to further qualify the query.

In this work we show that search engine query logs are a valuable source of data to build classifiers that can identify unambiguous concepts in a language. In our work, the features for such classifiers are calculated over a crawl of the Web and all queries issued in evenly spread days of a large commercial search engine. Classifiers like these seem especially suited to process short, noisy, conversational texts, which since recently have become widely available on the

Web. We show experimental results from shares and microposts from both Facebook[1] and Twitter[2]. We also propose two different applications to the classifiers built in this paper.

The rest of this paper is organized as follows. Section 2 discusses existing work that is relevant to our system. In Section 3 we discuss in detail how the classifiers proposed in this paper are built and in Section 4 we present numbers assessing their effectiveness. A discussion of potential applications for the system is shown in Section 5. Finally, Section 6 contains our conclusions about this work.

2. RELATED WORK

Krovetz and Croft observed that 75% of early information retrieval systems queries are unambiguous [8]. This observation has been later corroborated by a survey from Sanderson [17], where the impact of word sense disambiguation in information retrieval systems is studied. Although we do not rely only on that observation, one of the core hypothesis of this paper is that, to a lesser extent, this continues to be true for modern search engine queries, as long as the query does not show often in the query logs with further qualification. For example, the query *Washington* often needs to be refined as *George Washington* or *Washington (state)* or even *Washington, D.C.* while the query *Canada* often does not. This hypothesis is the central idea behind the metric described in Session 3.4.2, which has shown to be one of the strongest signals described in this paper.

The usage of the Web as an implicit training set for Natural Language Processing problems and ambiguity resolution in particular is presented in [15], where the author shows that the usage of simple algorithms and features extracted from large amounts of data yield competitive results with sophisticated unsupervised approaches and close results to that of supervised state of the art approaches.

Wacholder et al. studied the problem of disambiguation of proper names in [20]. Nadeu and Sekine conducted a survey [14] on the related field of Named Entity Recognition and Classification (NERC). Finally, the broader area of Word Sense Disambiguation is discussed in depth by Navigli [16].

The problem of extracting information from microposts has gained significant attention recently. In [4], Choudhury and Breslin have presented a classifier for Twitter posts able to detect players and associated micro-events in a sports match, achieving a f-measure of 87%. Using knowledge of the domain, such as sports jargon and names of players, they are able to disambiguate the Tweets. Li et al. proposed using a keyword extraction system for targeting ads to Facebook updates in [12], one of the applications we discuss in Section 4. Signals based on capitalization and document frequency are present in their work, but they did not explore any of the query log derived metrics.

Although the problems presented by the works discussed above share similarities with ours, none of their techniques can be directly applied. Word Sense Disambiguation is focused on finding senses of ambiguous terms in the local context, and does not discuss the properties of a given keyword outside its context. Also, traditional keyword extraction systems extract a set of keywords that characterize or summarize a given text, even if each of the individually extracted keywords might be ambiguous outside that set. Similarly, Named Entity Recognition systems look for entities that may or may not be unambiguous outside their context, such as *Ford* However, in our problem definition, only the keywords *Ford Motors*, *Henry Ford* or *Gerard Ford* should be extracted. Finally, we have no knowledge of the microposts being analyzed, preventing us from using domain specific features.

3. DETECTING UNAMBIGUITY

The ultimate goal of this work is to develop classifiers that detect unambiguous keywords in a language. As far as the knowledge of the authors goes, this is the first work proposing such classifiers. In order to formally present the problem, we will first introduce a few relevant concepts.

A common step previous to the processing of any corpus is the *segmentation* of the text in the documents. This is a complex problem which is beyond the scope of this paper and we assume that there is a state-of-the-art segmenter available[3]. The output of the segmentation process is a set of *keywords*. One keyword can be composed by one word or by a sequence of words – in the latter case we also refer to it as an n-gram or compound.

The Merriam-Webster dictionary[4] defines ambiguity as (1) *doubtful or uncertain especially from obscurity or indistinctness* or (2) *capable of being understood in two or more possible senses or ways*. However, there is a shortcoming in this definition, since it relies on human interpretation. One person can say that a given word is ambiguous while another could disagree. It turns that both could be right since the interpretation of the senses of a word can be done at different granularities.

Lapata and Keller [11] define ambiguity, or its complement, unambiguity, as function of a frequency of the senses that a given word or compound shows in a large corpus. We will instead use the terminology of the semiotics model by Ferdinand de Saussure [18], which yields a more intuitive definition for the scope of our work.

Definition 1. *Let a pair (f, c) be a sign, being f the form of the sign and c the concept it represents.*[5] *Let L be a language and S the set of all signs used in that language. Let the document frequency of a sign $df(f, c)$ in S be the number of documents the sign appear in a large corpus representative of the language L. We say that f is unambiguous if and only if $df(f, c)/\sum df(f, c') > \alpha$.*

In other words, we say that f is unambiguous if one of the concepts it may represent is α times more frequent in documents of the corpus than all the others combined. For our purposes, f is always a word or a compound word, and given that restriction, we will use Definition 1 as the basis for the problem being discussed through the rest of the paper.

3.1 Keyword Evaluation Methodology

Given a keyword q, an human evaluator can use Definition 1 to rate it as ambiguous or unambiguous. From this

[1] www.facebook.com

[2] www.twitter.com

[3] In this paper we use a segmenter developed internally at Google, Inc.

[4] www.merriam-webster.com

[5] In the original, form and concept are called signifier and significant, respectively.

definition, the evaluator should look at all the web documents containing q to understand the sense of the keyword in each of them. In practice, we do not need to go through all the documents in the corpus to find whether a keyword is unambiguous. Instead, we select, from the web, 16 random documents that contain q and manually check if all occurrences refer to the same concept, i.e., all positive occurrences. If yes, we say that the keyword q is unambiguous.

The choice of 16 random documents is legitimate. Since the evaluation of each random document that contains q is an experiment that has only two possible answers, we can assume it is a random sampling over a binomial distribution. Then, by using the Wilson method [21], we can calculate the binomial proportion confidence interval for a sample size of 16 with all of them positive occurrences, i.e. $\hat{p} = 1.0$, which result is $[0.8, 1.0]$ with center at 0.9. This interval gives us the α of Definition 1, which in this case will have a lower bound of 0.8 and an expected value of 0.9 with a 95% confidence. In other words: Given a keyword that was human-evaluated as unambiguous, we have 95% of chance that this keyword will refer to the same dominant concept in 80% of the corpus, but more likely it will be 90% of the corpus. This is the threshold we decided to use to assume a keyword is unambiguous in our human evaluations.

Using this methodology we have built a reference-set with 2634 ambiguous and 426 unambiguous keywords to be used in the analysis of the metrics presented in the next sections and as training-set input to the Machine Learning approach at Section 3.6.

3.2 Model Generation

There are two main source of signals for the unambiguous keywords classifiers presented here. The first is a sample with billions of documents of Google's search engine web collection. The second is as sample with billions of query entries from Google's query log corpus collected in evenly spread days.

The model generation is composed by a chain of off-line calculations of statistical properties of all signs in these two corpora and takes a few thousands of cpu-hours to complete. These properties will serve as the basis for the classification algorithms. This is an one-off work that only needs to be redone whenever there is an opportunity and/or the need of improving the performance of the system.

3.3 Web Collection Metrics

For every keyword resulting from the segmentation, we compute several properties using a large MapReduce-like system [5] visiting all the documents of the Web corpus. In the next sections we explain each property and give a histogram of its distribution among the keywords. Additionally to the plain histograms, we present two additional complementary histograms, one for the probability density of the metric among the ambiguous and another one for the unambiguous keywords of the reference-set defined in Section 3.1.

3.3.1 Inverse Document Frequency

The first metric, the Inverse Document Frequency (IDF), is the same found in the traditional information retrieval literature. It is computed over the Web document collection and is a proxy of the popularity of the keyword. It also serves as a good confidence indicator for all remaining metrics. The more popular a keyword is, the better the signal-to-noise ratio we have on it for all metrics. Figure 1 shows the IDF distribution for unigrams and keywords found on the Web collection.

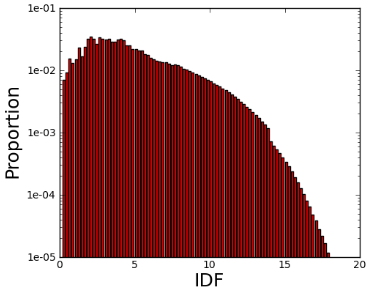


Figure 1: IDF distribution for the web collection.

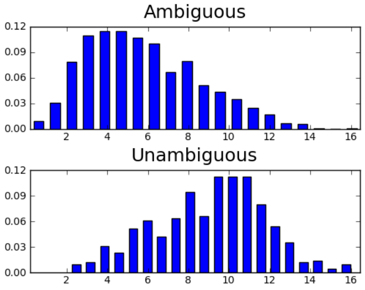


Figure 2: IDF distribution for the reference-set.

The histograms for IDF distribution among the reference-set is plotted in Figure 2. The top chart shows the histogram for ambiguous keywords while the bottom shows the unambiguous. This histogram shows that ambiguous keywords tends to have lower IDF values. The lowest ones are language constructions such as *just* and *this*. The misspellings and uncommon person names lies in the very high IDF range. While unambiguous keywords tends to have higher IDF values, there is a big overlap with lots of unambiguous ones in the mid-lower IDF range, such as *White House* and *Disney*. This overlap makes it hard for the IDF metric to be used to separate both sets. However, we can apply it for filtering language constructions and misspellings.

3.3.2 Caps First Ratio

Caps First Ratio (CFR) is the ratio that a given keyword shows up on the Web collection with the first letter capitalized and we interpret it as strong indicator of names. We implemented the techniques from [13] to detect capitalized keywords.

The CFR metric has the obvious property of detecting nouns, but it has another subtle interesting characteristic. Several noun compounds include, as an extra qualifier, subcompounds or unigrams that are unambiguous by themselves, for example *Justin Bieber* and *Bieber* are both unambiguous. In this case, we consider the occurrence of every capitalized word not only in the CFR calculation of the compound it belongs to – *Justin Bieber* – , but also in the CFR score of the sub-compounds and unigrams of that compound

– *Bieber*. This helps increasing the CFR score of nouns that act as unambiguous qualifiers. For example, for the *Bieber* unigram, using only the initials for legibility, we calculate:

$$CFR(B) = \frac{count(JB) + count(B)}{count(jb) + count(b) + count(JB) + count(B)}$$

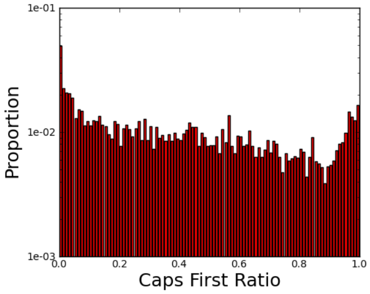


Figure 3: CFR distribution for the web collection.

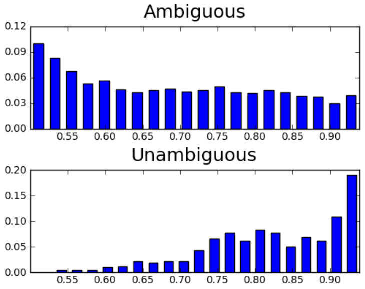


Figure 4: CFR distribution for the reference-set.

Figure 3 shows, the CFR distribution seen on Web documents. The reference-set histograms at Figure 4, are more heterogeneous than the IDF counterpart. The mid-low range of CFR values includes mostly only ambiguous keywords, while the unambiguous histogram has a sharp growth in the high values.

3.4 Query Log Metrics

Query logs have proved to be a valuable source of information for several fields of computer science [19]. In our work we collected data from three evenly spread days worth of queries in the logs of a large search engine. As with the Web corpus, we compute the following metrics for each keyword generated by the segmentation of the query log corpus.

3.4.1 Sessions Inverse Document Frequency

The Sessions Inverse Document Frequency (SIDF) is analogous to the Web metric of the same name, but it is calculated over the search engine query stream. Each session [19] is considered as a document. Figures 5 and 6 presents the distribution of this metric for the query stream and for the reference-set respectively. This signal has similar properties to its Web counterpart, but with a bias towards concepts and against intrinsic language characteristics. By comparing Figures 1 and 5, one can draw an interesting conclusion: stopwords and auxiliary language constructions appear much less often in the query stream. Because of that we can say it is safe to discard anything that is not popular in the query stream.

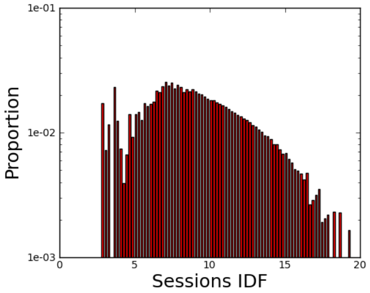


Figure 5: SIDF distribution for an infinite stream of web text.

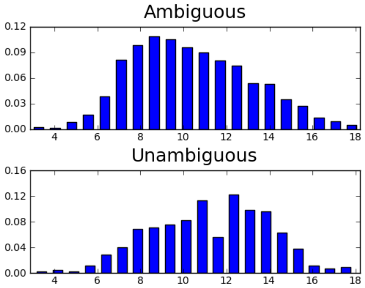


Figure 6: SIDF distribution for the reference-set.

3.4.2 Sessions Exact Ratio

A key metric that comes from the query stream analysis is the Sessions Exact Ratio (SER). It tells how often a given keyword shows up by itself in the search box. This is the strongest indicator that this keyword is unambiguous when taken out of context. Figures 7 and 8 shows the histogram for this metric on the Web collection and the reference-set respectively. As can be seen, the ambiguous and unambiguous reference-set is mostly separable. Some examples of unambiguous keywords in the very high range of the histogram are: *Tom Hicks*, *Madison Square Garden* and *Groupon*.

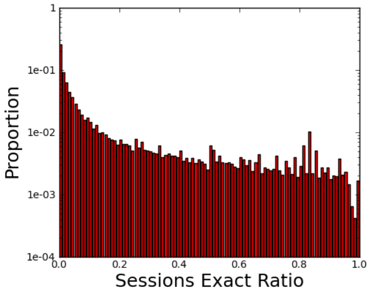


Figure 7: SER distribution for an infinite stream of web text.

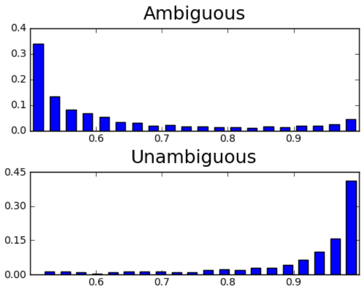


Figure 8: SER distribution for the reference-set.

3.4.3 Search Bias

The last metric, Search Bias (SB), is not directly derived from the query stream, but rather obtained through a combination of sessions and Web signals. Search Bias can be thought as the ratio of appearance of a keyword in the query logs corpus divided by the ratio of appearance of the same keyword on the Web corpus.

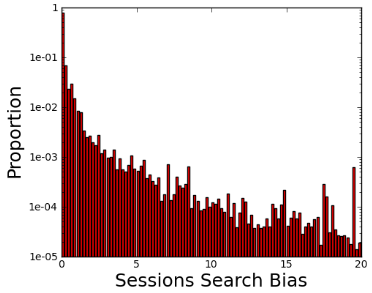


Figure 9: SB distribution for an infinite stream of web text.

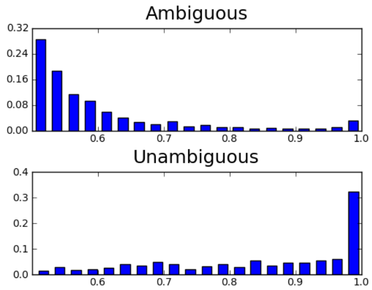


Figure 10: SB distribution for the reference-set.

The naive calculation of this number leads to a value with bad properties due to very common words on the Web corpus and the high frequency of compounds in the query log corpus. To avoid those issues, search bias is calculated taking into account only "noble" occurrences of a keyword in Web and query logs corpora. For the Web, we consider that only capitalized occurrences are noble, while from query logs we only consider those occurrences where the keyword appear by itself in the query. The distribution of this metric can be seen on Figures 9 and 10.

The histograms shown here are not just a tool to help visualize how each metric may be used to dissociate both sets, but more than that, it is an evidence that the metrics used here can succeed in building an effective classifier.

3.5 A hand-crafted classifier

In this section we present a simple hand-crafted algorithm. It was developed upon the discussions and histogram observations of above metrics, regarding the reference-set separation. We use this algorithm to expose the ideas without adding the complexity that inherently comes with traditional machine learning techniques, as well as to avoid hiding the interesting properties of the data under analysis. Later in this paper we present a Support Vector Machines (SVM) approach for the same classification task. Refer to Section 3.6 for more details.

Algorithm 1: The *IsUnambiguous* Algorithm.

```
 1 begin
 2     if $sidf > 15$ then return $false$;
 3     if $uni \wedge idf > 12 \wedge sidf > 12$ then return $false$;
 4     if $cfr < 0.4$ then return false;
 5     if $ser < 0.35$ then return false;
 6     if $sb < 0.01$ then return false;
 7     if $cfr + ser + sb < 1$ then return false;
 8     if $charcount < 3$ then return false;
 9     if $blacklisted$ then return false;
10 end
```

Algorithm 1 presents our hand-crafted approach for the unambiguity classification problem. Each line is a filter of ambiguous keywords. In Figure 11 one can see how many keywords are being discarded as the classifier is applied on top of the Web corpus. In the end, only 3.8% of all keywords occurrences are considered unambiguous.

The *Sessions IDF* funnel component corresponds to Line 2 of the algorithm. Its goal is to remove everything that is too rare, such as misspells. Usually, plain IDF is used for this type of cutting, but bigrams and larger keywords have a very high IDF on the Web corpus. In the query logs corpus, however, large keywords appear much more often and anything that is not too rare will not be filtered by this rule.

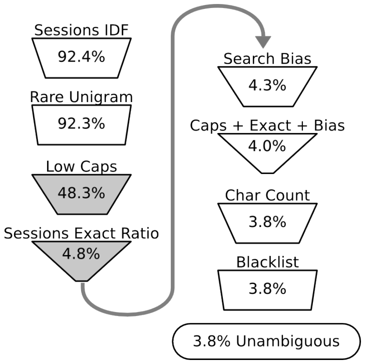


Figure 11: The percentage of the input text keywords each line of the algorithm filters. The most important filters are highlighted in gray.

In Line 3, the *Rare Unigrams* filter unigrams typos that are not rare enough to be discarded by Sessions IDF and also come with all types of capitalization. Since unigrams are more frequent than compounds, we can apply a more restrictive threshold.

The *Low Caps* filter comes from Line 4 of the algorithm. Not only it is responsible for restricting the classifier to nouns, it also rejects all the general nouns. For example, the noun *ball* has a very low caps first ratio, but *Wilson NCAA Reaction Basketball* is almost always typed all caps.

The most powerful feature for our classifier, the *Sessions Exact Ratio* (SER) filter, is used in Line 5. It reflects the key intuition that our work builds upon: users know that search engines have little context to understand their communication and because of that they formulate unambiguous queries.

The derived metric *Search Bias* is used in Line 6. Some keywords, like *Unfortunately*, tend to have both high CFR – because they are used to start phrases – and high SER – because they have low query volume. This filter detects those language constructions that are way more common in the Web corpus than in the search corpus and discards them.

The combined *Caps+Exact+Bias* filter in Line 7 is the most complex part of the algorithm. Its goal is allow us to reduce the thresholds of the individual filters applied before without incurring in a loss of precision. This filter will let keywords that score very high in any of the metrics combined pass, but will discard those that have a low average all around.

The *Character Count* is a simplistic filter as can be seen in Line 8. When dealing with two characters keywords, all the metrics have bad properties, and we simply discard all of them. In fact, a perfect English classifier limited to two character unigrams can be manually built by inspecting all the 626 possible combinations.

Finally, the last step of the algorithm is the *Blacklist* filter, in Line 9. Some keywords have very extreme metrics and tend to pass through all the filters, and we simply blacklist them. For English, we currently blacklist 40 greetings expressions, such as *Happy birthday* and *Hello* and some very common words like *Indeed*. In fact, the metrics for those keywords are so extreme that by looking at the top values for each metric one can easily spot them. We also blacklisted the Web sites names *Google*, *Facebook*, *Twitter* and *Gmail* because, although unambiguous, they are so common in our evaluation data that they would positively benefit our results with no interesting characteristics.

3.6 A Machine Learning approach

To challenge the hand-crafted algorithm presented in the previous section and test if its intuitions were correct, we employ a Support Vector Machines (SVM) algorithm to the same classification task. It is a well-known technique and there is a ready to use state-of-the-art implementation, namely libSVM [3]. The down-side of the machine learning approach is that labeled data acquisition for this novel classification task is challenging. The training set used was the reference-set explained before, with the 2634 ambiguous and 426 unambiguous manually classified keywords. Each metric shown in sections 3.3 and 3.4 were rescaled to the 0-1 range and used as SVM input features. We used only the Radial Basis Function kernel present in libSVM: $K(x_i, x_j) = \exp(-\gamma \times \|x_i - x_j\|^2), \gamma > 0$. By tuning the γ and C parameters we can control the trade-off between false-positive and false-negative errors. After doing a simple grid-search of both parameters using cross-validation, we picked a value around 0.05 for γ and 0.1 for C.

4. EXPERIMENTAL RESULTS

In this section we present experimental results obtained with the implementation of the classifier described in Section 3.5 and the SVM described in Section 3.6.

4.1 Test set definition for manual evaluation

The input of each classifier is a chunk of free form text, from now on referred to as a *micropost*, and the output is a list of keywords assumed by the classifier to represent unambiguous concepts from the input. We decided to use a micropost as a unit of evaluation, in opposition to a single keyword, because we believe the results from this analysis represent better the real world applications. In order to not favor our classifier, the rater is instructed to mark a micropost as False Positive if she finds one ambiguous keyword classified as unambiguous, even if the classifier also correctly extracted another unambiguous keyword from the same micropost.

We selected two different sources of microposts to feed the classifier: Twitter and Facebook. This data set and the reference-set that was used to train the SVM classifier are disjoint to avoid over-fit. Twitter is a social Web site where users can send and read text-messages with up to 140 characters, called tweets. We expect this kind of social Web site to have mostly conversational text and noise, which carry little information by themselves. To feed the classifiers, each tweet is considered independent from each other and its text is used as input to the classifier. Facebook is a social network site where users can create virtual connections with their friends. One Facebook feature is the status message updates. The status message is much like a tweet, but the 140 characters limit is not imposed. Since users can follow-up on their friends updates, the entire conversation is used as input for the classifiers.

4.2 Methodology for manual evaluation

The methodology used for manual evaluation consists of collecting a random set of microposts from each data source and feeding each one into the classifiers. The original micropost and the output of the classifier are then shown to three different raters. The output might be empty, in the case the classifier did not find any unambiguous keyword in the micropost. Otherwise it contains at least one keyword, which was classified as unambiguous. In case of discordance, it is discussed until consensus is reached. Regardless of the classifier output, raters must investigate the keywords present in the micropost. They use the methodology presented in Section 3.1 to rate each keyword q. Based on the output of the classifier and the inspection of the micropost content carried out by the rater, each micropost is rated as below:

True Positive (TP): There are unambiguous keywords in the micropost and the system has extracted at least one.

True Negative (TN): There are no unambiguous keywords and the system extracted nothing.

False Positive (FP): The system has extracted an ambiguous keyword.

False Negative (FN): There are unambiguous keywords in the micropost but the system has extracted none.

We use the output of the rating phase to compute the standard evaluation metrics in the Information Retrieval field, such as precision, recall, accuracy and F-score [1].

Both models – the hand-crafted and the SVM classifier – were built using the context available at the English Web, but it must be considered that people have an incomplete cultural context and sometimes it may not be obvious that a given keyword is unambiguous. For example, during the evaluation one rater could not recognize upfront the keyword *Doug Flutie*, which was extracted by the system. Even though this rater did not recognize this keyword, *Doug Flutie* is indeed an unambiguous keyword because every person with culture about American Football will recognize him as Douglas Richard "Doug" Flutie, a famous football quarterback who played professionally in the United States Football League and, more importantly, the name *Doug Flutie* is not used as an identifier in any other significant context besides that. Our precise definition of unambiguity prevents this problem, since the rater will learn the sense(s) of the keyword when looking at the 16 randomly sampled documents, as it was the case in this example.

4.3 Numerical results

Table 1 presents the output of the raters for Facebook and Twitter microposts.

	Hand Algorithm		**SVM**	
	Twitter	Facebook	Twitter	Facebook
TP	99	106	85	85
TN	494	480	511	510
FP	38	34	22	21
FN	74	64	87	68

Table 1: Break-down of metrics for Twitter and Facebook.

Twitter

Following the experimental methodology we analyzed a set of 705 tweets, which were randomly selected from a set of 170k tweets that were crawled by a fairly random walk in the Twitter graph. We used these tweets as input of both classifiers and presented the output to the raters. The hand-crafted classifier was able to reach precision of 72.26%, and sensitivity (recall) of 56.22%. The True Negative Rate (TNR, or specificity) is high (92.86%), upon what one can conclude that most tweets do not contain unambiguous keywords. For Twitter the system reached an accuracy of 84% with an F-Score of 0.64. The SVM model reached a precision of 79.43%, i.e., a performance almost 10% better than the achieved by the hand-crafted algorithm. The achieved recall is 49.42%, considerably worse than the recall reached by the hand-crafted algorithm. The TNR of the SVM model is 95.87%, and the system reached an accuracy of 85% with an F-Score of 0.61.

Facebook

Following a random selection strategy similar to Twitter, we collected 684 conversations that took place around Facebook status message updates. For this data set, the hand-crafted system reached a precision of 75.71% and recall of 62.36%. The TNR was of 93.39%, whereas the accuracy reached 85% with an F-score of 0.68. The SVM model got slightly better results. The precision is 80.19% (around 6% better), whereas the recall is 55.55%. Again, the True Negative Rate is really high, 96.05%. The classifier has an accuracy of 87% with an F-Score of 0.66. The high value for the True Negative Rate is a sign that most conversations in Social Networks like Facebook and Twitter are not proper for context-extraction systems such as content-targeted advertisement if used without any pre-processing.

To compare the results of the two classifiers presented above, we also evaluated two known systems: Yahoo! Term Extractor API – aimed to Keyword Extraction tasks – reached 18.98% of precision and 57.69% of recall for Facebook data, and 14.89% of precision and 77.7% of recall for Twitter data; and the Stanford Named Entity Recognizer [6] – aimed to Named Entity Recognition and Classification tasks – reached 35% of precision and 69.99% of recall for Facebook data, and 39.65% of precision and 74.19% of recall for Twitter data.

Both systems reached a higher recall, but for the real-world applications discussed in section 5 we cannot afford extracting a wrong keyword from a noisy text. In these applications precision is more important, and for both systems it is much lower than the two filters developed in this work. The high recall and low precision result is expected for these systems, since they were engineered for different tasks and do not perform well for the unambiguity detection task defined here.

5. APPLICATIONS

Given the properties of the classifiers presented in this paper, we believe they are suited for a set of different applications that are becoming more important given last developments on the Web industry.

5.1 Ad targeting in Social Networks

In Social Networks users keep updating their status messages (or tweets) with what they have in mind. The number of daily updates in the most prominent networks is huge[6], turning this channel into a potential candidate for input of content-targeted advertisement systems [12]. For instance, it is just fine to deliver an advertisement piece like *Buy tickets to the coming Jonas Brothers show!*, right next to a micropost where a user claims to be the biggest fan of this music group. However, the conversational text brings even more complexity for the already tough task [9] of delivering content-targeted ads. Feeding these systems with noisy text may lead them to return non-relevant ads. One can use the classifiers proposed in this paper as a filtering layer on top of current content-targeted advertisement systems. The filter would delegate calls to the ads systems only when it is possible to retrieve relevant content from the microposts being targeted.

5.2 Automatic reference in Social Networks

User profiles in Social Networks could also be classified as unambiguous by using the profile name for example. Whenever a micropost has the unambiguous keywords that matches a profile name, a link to that profile could be added, instead of just pure text. This could be done for celebrity profiles, for example, when a user posts *"I just watched the last Quentin Tarantino movie."*, a link to the *Quentin Tarantino* profile could be added.

[6] http://news.cnet.com/8301-13577_3-10378353-36.html

6. CONCLUSIONS

In this paper we presented a novel classification problem aimed at identifying the unambiguous keywords in a language, and formally defined it together with an evaluation methodology. We also have presented two different algorithms for the classification problem and the corresponding numerical results achieved by both of them. The proposed algorithms are built on top of traditional information retrieval metrics and novel metrics based on the query log corpus. The introduction of these metrics, Sessions IDF, Sessions Exact Ratio and Search Bias, is by itself an important contribution. We believe these metrics will be useful in other problem domains as well.

Our evaluation have shown that our classifiers were able to meet precision $\geq$ 72%, recall $\geq$ 49%, accuracy $\geq$ 84% and F-Score $\geq$ 0.61, even when the input is composed by the noisy microposts from Facebook and Twitter, two of the biggest sites in the world nowadays, outperforming two known systems from the traditional keyword extraction and Named Entity Recognition fields.

Another interesting aspect of the presented work is that it diverges from the bag-of-words analyses that dominate the research in the area. Instead, we have focused on directly finding the specific keyword that define a concept, avoiding the shortcomings that come from having a representation that cannot be understood by a human or does not meet the expectations of other systems. This leads immediatelly to our future work proposal of using the extracted keywords as beacons for further qualification of other keywords in the text. For example, the extracted keyword *Lionel Messi* can be used to anchor the word *goal* to the concept of scoring in the soccer sport, instead rather the more general idea of an objective to be achieved. We expect this inside-out approach for extracting semantics in microposts to perform better than traditional word collection approches.

More and more researchers have access to query logs and many may directly benefit from the metrics proposed here either to tackle the same classification problem or to innovate in their own domains. For the industry, we have shown a solution for extracting information from microposts, a type of content that has experienced tremendous growth on the Web in the recent past.

7. REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[2] M. Banko and E. Brill. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 26–33, Morristown, NJ, USA, 2001.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: a Library for Support Vector Machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

[4] S. Choudhury and J. Breslin. Extracting semantic entities and events from sports tweets. In *Making Sense of Microposts (#MSM2011)*, pages 22–32, 2011.

[5] J. Dean and S. Ghemawat. Mapreduce: Simplified Data Processing on Large Clusters. In *Sixth Symposium on Operating System Design and Implementation*, pages 137–150, 2004.

[6] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[7] A. Halevy, P. Norvig, and F. Pereira. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24:8–12, 2009.

[8] R. Krovetz and W. B. Croft. Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems*, 10:115–141, April 1992.

[9] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto. Learning to Advertise. In *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 549–556, New York, NY, USA, 2006.

[10] M. Lapata and F. Keller. Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing*, 2(1):3, 2005.

[11] M. Lapata and F. Keller. An Information Retrieval Approach to Sense Ranking. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 348–355, 2007.

[12] Z. Li, D. Zhou, Y.-F. Juan, and J. Han. Keyword Extraction for Social Snippets. In *Proceedings of the 19th ACM International Conference on World wide web*, pages 1143–1144, New York, NY, USA, 2010.

[13] A. Mikheev. A Knowledge-free Method for Capitalized Word Disambiguation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 159–166, Morristown, NJ, USA, 1999.

[14] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007.

[15] P. Nakov and M. Hearst. Using the Web as an Implicit Training Set: Application to Structural Ambiguity Resolution. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 835–842, Morristown, NJ, USA, 2005.

[16] R. Navigli. Word Sense Disambiguation: A Survey. *ACM Comput. Surv.*, 41(2):1–69, 2009.

[17] M. Sanderson. Retrieving with Good Sense. *Information Retrieval*, 2(1):49–69, 2000.

[18] F. D. Saussure. *Course in General Linguistics.* Open Court Pub Co, 1986.

[19] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(1):6–12, 1999.

[20] N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of Proper Names in Text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 202–208, Morristown, NJ, USA, 1997.

[21] E. B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):pp. 209–212, 1927.

Knowledge Discovery in distributed Social Web sharing activities

Simon Scerri, Keith Cortis, Ismael Rivera, Siegfried Handschuh
Digital Enterprise Research Institute
National University of Ireland, Galway
firstname.lastname@deri.org

ABSTRACT

Taking into consideration the steady shift towards information digitisation, an increasing number of approaches are targeting the unification of the user's digital "Personal Information Sphere" to increase user awareness, provide single-point management, and enable context-driven recommendation. The Personal Information Sphere refers to both conventional information such as semi/structured information on the user's personal devices and online accounts, but also in the form of more abstract personal information such as a user's presence and activities. Online activities constitute a rich source for mining this type of personal information, since they are usually the only means by which a typical user consciously puts effort into sharing their activities. In view of this opportunity, we present an approach to extract implicit presence knowledge embedded in multiple streams of heterogeneous online posts. Semantic Web technologies are applied on top of syntactic analysis to extract and map entities onto a personal knowledge base, itself integrated within the wider context of the Semantic Web. For the purpose, we introduce the DLPO ontology—a concise ontology that captures all facets of dynamic personal information shared through online posts, as well their various derived links to personal and global semantic data clouds. Based on this conceptualisation, we outline the information extraction techniques targeted by our approach and present an as yet theoretical use-case to substantiate it.

Keywords

Social Web, Microposts, Presence, Ontologies, Personal Information Management

1. INTRODUCTION

The di.me project[1] is one of many initiatives targeting the unification of a user's personal information sphere across heterogeneous sources, with the aim of providing an intelligent and proactive system—the di.me userware—that assists the user in their day-to-day activities.

The di.me userware generates and constantly updates a representation of the user's Personal Information Model (PIM), based on a comprehensive modelling framework that combines various ontologies provided by the OSCA Foundation (OSCAF)[2]. Thus, the PIM serves as an integrated personal Knowledge Base (KB) containing all available knowledge about the user (i.e. their devices, accounts, social relationships, resources, activities, etc.), as crawled and mined by the di.me userware. It is however not a self-contained knowledge model, also containing references to resources in open repositories. Knowledge stored in the PIM enables advanced features such as distributed personal information management, improved search and retrieval, context-awareness and context-dependant recommendation.

In the context of di.me, the term personal information sphere refers to not just conventional structured or semi-structured data (e.g. files, folder structures, contact lists, photo albums, status messages, etc.), as was the case in earlier initiatives such as the Social Semantic Desktop [22]. Di.me also covers unstructured, more abstract forms of personal information, including complex concepts such as user contexts, situations, physical and online presence. In order to elicit this type of personal information, di.me identifies two types of sources: sensors (presence information relayed by device-embedded sensors, user attention monitoring, etc.) and social sharing activities (serving as 'virtual sensors' [5]).

In this paper, we focus on the latter, as a novel and rich source for enriching the PIM with presence-related knowledge. Thus, the main motivation for this work is the extraction of information from multiple streams of heterogeneous online-post data[3], by both the users and their contacts, in order to generate valuable outputs. More specifically, we exploit online sharing activities in order to i) enrich the PIM with discovered personal and social knowledge (e.g. detecting a user's current activity, availability, learning who is in the same area, doing a similar activity, discussing the same topic, etc.), ii) semantically link post items across personal social networks (e.g. Facebook, Twitter, LinkedIn posts

[1] http://dime-project.eu/

[2] http://www.oscaf.org/ – OSCAF ontologies have been adopted by various initiatives, including di.me.

[3] Although at the conceptual level we consider all types and lengths of online posts, the technical approach is more focused on the shorter, so-called microposts.

Published as part of the #MSM2012 Workshop proceedings, available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

about the same topic, event, location, video, etc.), and iii) enable social-based recommendations (e.g. the user is shown contacts that have similar interests, are in the same area, are discussing similar topics, doing related activities, etc.).

In line with the above objectives, this paper provides two research contributions: the provision of a suitable model for the representation of online posts and their embedded knowledge, and the design of comprehensive semantic lifting techniques for the extraction of the knowledge itself. The first contribution consists of the LivePost Ontology (DLPO), an open knowledge representation standard[4] integrated within the existing PIM models, that provides rich conceptualisations of various types of online posts, covering more emerging Social Web sharing features than any of the available standards.

The second contribution is an overview of the techniques to be employed for the extraction of semantics from microposts. Online post items contain both easily-acquired semi/structured data (e.g. hyperlinks, creator, date and time, people tagged, nearby places, etc.) as well as hidden abstract data that requires the application of advanced linguistic techniques. The target of the semantic lifting process is to break down retrieved posts into one or more specific subtypes (e.g. message post, image post, checkin, etc.), and enrich them with clear semantics (including links to existing PIM and/or Semantic Web resources). The Information Extraction (IE) techniques employed include shallow parsing techniques such as Named-Entity Extraction (NEE), keyword extraction and hyperlink resolution, followed by more sophisticated analysis such as Named-Entity Resolution (NER) and co-reference resolution, topic extraction, and time window analysis.

After comparing related work (Section 2) and outlining the approach (Section 3), we provide examples of how our approach can result in a semantic representation that is integrated within the PIM (Section 4), before providing some concluding remarks and directions for future work (Section 5).

2. RELATED WORK

In the light of our requirements, we here outline and compare related approaches. The use of social data (in the form of microposts) as input data for providing some form of recommendations is not an entirely novel concept, and has been in fact applied by a number of other approaches [7], [6], [8], [1], [26]. In particular, Chang and Sun [8] analyse a dataset of Point of Interests (POIs) collected from Facebook Places to construct a prediction model for a user's future locations. The BOTTARI mobile app [7] exploits social context to provide items to the user, relevant to their location. Our approach is more similar to the latter, collecting information about a user's and their contacts' presence, such as to enable the discovery of information which would otherwise easily be missed, e.g., contacts discussing the same topics, travelling to the same city, etc. The collected information will also be used to provide context-aware suggestions (nearby POIs that are recommended by trusted contacts) and warnings (untrusted persons in the vicinity).

Given our emphasis on knowledge representation, Table 1 compares existing vocabularies (or a combination of) modelling the required domains—User Presence and Social Web sharing through posts—against the DLPO. As is evident, most approaches are targeted towards one domain, with only a few supporting cross-domain modelling. The Semantically-Interlinked Online Communities (SIOC) Ontology for instance, is more oriented towards Social Web sharing, as its original aim was to interlink online communities [3]. Although it caters for online posts (denoted by a full circle), it does not attempt to make any sort of link between user posts and user presence, which is a missed opportunity given that a large number of microposts are linked to physical and online user activities/experiences[5] (i.e. user presence[6]). A proposed combination of the SIOC(T), Friend of a Friend (FOAF) [4] and Simple Knowledge Organization System (SKOS) [17] ontologies[7] is also unable to represent any form of user presence in relation to the posts. The Bottari Ontology [7] is another SIOC extension which supports relationships between posts, locations and user sentiment, as extracted from tweets. Another vocabulary that provides for representations of online posts in the context of user presence is the Online Presence Ontology (OPO) [23], which models a user's current activities on online services. But since the OPO does not cover all physical presence aspects, it is insufficient for our representation needs. The PreSense Ontology [5] reuses OPO vocabulary to effectively cater for the representation of both physical and online presence. It also makes a connection between a user's presence and their online status streams, which can serve as 'virtual' presence sensors. However, without the re-use of additional OPO vocabulary, PreSense remains unable to provide the comprehensive modelling of online posts that we need.

Another requirement is to be able to decompose a post into multiple concurrent sub-types (e.g. into a status message, an image photo upload, and a check-in). Post decomposition has the advantage of improving retrieval (e.g. user later looks for all items posted in an area, thus showing only the meta-data of specific posts) and deletion (e.g. it's much easier to remove any posts related to a deleted PIM concept, since any related PIM concepts are directly linked to their corresponding posts, unlike in SIOCT where relations between a PIM concept and a resource are not directly known). Additionally, vocabularies such as the Privacy Preference Ontology (PPO) [21] could be employed to enable a user to restrict or allow access to only some types of subposts (e.g. share all types of posts with a contact group, except for place check-ins). Although most approaches in Table 1 provide for multiple post types (denoted by a half-full circle), the DLPO provides the best representation of concurrent posts. The original SIOC vocabulary did not even distinguish between different kinds of posts, a limitation which

[4] `http://www.semanticdesktop.org/ontologies/dlpo/`— a candidate OSCAF submission

[5] 77.7% of tweets consist of 'conversations' and 'pointless babble', both of which are considered a source of user presence information: `http://mashable.com/2009/08/12/twitter-analysis/`

[6] By the term 'presence' we refer to both online (activities e.g. check-ins, posts, liking; interactions e.g. playing a game, chatting; availability, visibility, etc.) and physical (activities e.g. travelling, walking, working; current location, nearby people and places, etc.) user experiences

[7] `http://sioc-project.org/node/341`

Table 1: Knowledge Models

	SIOC	SIOCT	SIOC(T)+FOAF+SKOS	Bottari	OPO	PreSense	DLPO
Online Posts (General)	●	●	●	●	●	○	●
Post Multi/Sub-typing	○	◐	◐	◐	◐	○	●
Microposts	○	●	●	●	●	●	●
Online Presence	○	○	○	○	●	●	●
Physical Presence	○	○	○	◐	◐	●	●
Online Sharing Practices	◐	◐	◐	◐	○	○	●

was addressed by a later extension (SIOCT[8]) that provides additional sub-types such as *sioc:MicroBlogPost*.

A last quality on which we base our comparison is the modelling support for emerging Social Web sharing practices and interactions such as: post item replies, resharing (or retweeting), endorsement ('liking', starring, favouriting), and general time-awareness (i.e. timestamps, succession of posts, etc.). Neither of the surveyed vocabularies provide for all of the above features, in contrast to the DLPO. At this stage it is necessary to point out that the DLPO does not ignore existing established vocabularies such as SIOC, and in fact, fully re-uses some of its elements. Apart from providing all qualities listed in Table 1, the DLPO stands out for another unmatched characteristic—it is integrated within an entire framework of ontologies targeting the representation of a user's entire personal information sphere. Thus, instances of the DLPO automatically become part of, and are tightly integrated within, a representation for the user's entire PIM.

After considering alternate knowledge models, we now compare the di.me approach to related approaches, against the IE and semantic lifting techniques they employ (Table 2). Some form of linguistic analysis (keyword/topic extraction, NEE for various entity types) is performed by all. The tools provided by [15] and [20], present improved NEE techniques based on informal communication—noisy, informal and insufficient information—of microposts such as tweets. On the other hand, NER is only performed by Zoltan and Johann [26], in their approach to construct user profiles from knowledge extracted from microposts. In addition, extracted entities are linked to specific concepts within the Linked Open Data (LOD)[9] Cloud. This constitutes what we refer to as 'semantic lifting', whereby structured/unstructured data is lifted onto standard knowledge models such as ontologies. However, our approach is unique because, apart from community KBs (such as LOD), we also enrich a personal KB—the user's PIM. This has the obvious advantage that it consists only of personal data, making it easier to determine equivalence between entities in microposts and PIM items. In addition to the generic techniques listed in Table 2, we also employ techniques such as hyperlink resolution and time window analysis.

3. APPROACH

In this section we present both the conceptual and the practical aspects of our approach. The former consists of an exercise in knowledge modelling, with the resulting ontology then serving as a standard for data integration across multiple and heterogeneous online data sources. The practical side of our approach targets the semantic lifting of data from these sources, including straightforward lifting of structured data, before moving on to the more challenging extraction of semantics from semi-structured and unstructured data.

3.1 MODELLING ONLINE POSTS

The motivation for our approach outlines four major requirements for this modelling task:

i) to support and re-use existing standards—particularly the W3C submission for SIOC [3]

ii) integration within established PIM knowledge models

iii) semantic decomposition into independent sub-posts

iv) representation of emerging online Social Web practices

The DLPO, of which an overview is given by Fig. 1, satisfies all of the above requirements. To adhere to the first, the superclass *dlpo:LivePost* is itself an extension of the generic *sioc:Post*, inheriting all SIOC properties that apply (e.g. sioc:has_creator, sioc:hasTopic). DLPO also introduces two subproperties of the core SIOC properties *sioc:has_reply* and *sioc:reply_Of*, for use within the DLPO context: *dlpo:hasReply*, *dlpo:replyOf*.

The second requirement ensures that the information represented by DLPO is firmly integrated within the wider context of distributed personal information modelling. For the purpose, embracing the Information Element (NIE) ontology, *dlpo:LivePost* is also a subclass of *nie:InformationElement*. This means that a post instance will inherit all properties that apply, e.g. *nie:created* and its DLPO extension *dlpo:timestamp*, both of which are subproperties of *dcterms:created*. The purpose of the NIE ontologies[10] is to provide a vocabulary for describing information elements which are commonly present on a source hosting information belonging to a user. The Personal Information Model Ontology (PIMO)[11] is then used to generate a representation of the user's mental model, by abstracting multiple information element occurrences (e.g. different contacts for a person, formats for a document, etc.) onto a unique and integrated model. Sources for populating the PIM include personal devices and online accounts. The Account Ontology (DAO) enables the representation of online sources (e.g. Facebook, LinkedIn), such that personal information elicited from within can retain the link to

[8] http://rdfs.org/sioc/types

[9] http://richard.cyganiak.de/2007/10/lod/

[10] http://www.semanticdesktop.org/ontologies/nie/

[11] http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/

Table 2: General Approaches

	[7]	[24]	[6]	[9]	[8]	[1]	[26]	[18]	[15]	[20]	di.me
Keyword Extraction	○	○	●	●	○	○	●	○	○	○	●
Topic Extraction	○	○	●	○	●	●	●	●	○	●	●
NEE (Events)	○	◐	○	●	◐	●	○	○	○	◐	●
NEE (People)	○	◐	●	●	●	●	●	○	●	●	●
NEE (Activities)	○	◐	○	○	◐	○	○	○	○	○	●
NEE (Locations)	◐	◐	●	○	●	●	●	●	●	●	●
NER	○	○	○	○	○	○	●	○	○	○	●
Semantic Lifting	○	○	○	○	○	○	●	○	○	○	●

each source. *dao:Account* is a subclass of *sioc:Container*, and *dao:source* a subproperty of *sioc:hasContainer*. Another relevant OSCAF ontology is the Annotation Ontology (NAO)[12], which provides a vocabulary for defining generic, domain-independent relationships between related resources. In the DLPO, the NAO is applied to define the online post's creator *nao:creator*, a defining image or symbol *nao:prefSymbol*, topics *nao:hasTopic*, labels *nao:prefLabel*, tags *nao:hasTag*, descriptions *nao:description* and pointers to the post's unique identifier on the source account *nao:externalIdentifier*. The DLPO also extends two NAO properties, the *nao:isRelated* with *dlpo:relatedResource*, and *nao:hasSuperResource* with *dlpo:definingResource*, to create two types of generic relationships between online posts, or their sub-types, to items in the PIM. *dlpo:relatedResource* can be used to create a semantic link between a livepost and the PIM items which it is about (e.g. a post about people, topics, images, events, places, etc. that are known and represented in the PIM). The *dlpo:definingResource* goes one step further, defining a direct relationship between a post subtype and a PIM item (e.g. linking an EventPost to the actual Event which it describes).

Sub-typing online posts is related to the third requirement, that of decomposing a post into semantically distinguishable subposts. Online service accounts commonly distinguish between different types of posts. The DLPO supports these distinctions, and also the fact that post sub-types can either occur individually or, also in conjunction (e.g. a composite Status Message that contains text, a photo, a nearby location, people tagged, etc.). By definition, a *dlpo:LivePost* may consist of multiple subposts, which relationship can be represented through the use of the *dlpo:isComposedOf* property (a subproperty of *nao:hasSubResource*). The DLPO differs between four categories of sub-posts:

1. **Message** - Represents the text-based subpart of online posts. If a message is not in-reply to a previous message (denoted by *dlpo:replyOf*) then it is of subtype *dlpo:Status*, otherwise it is of subtype *dlpo:Comment*.
2. **MultimediaPost** - Represents subposts containing links to multimedia items that are either available online, or that have been uploaded to the online account. This category of subposts is further refined into *dlpo:VideoPost*, *dlpo:ImagePost* and *dlpo:AudioPost*.
3. **WebDocumentPost** - Represents subposts containing links to online text-based containers. Examples range from a note (*dlpo:NotePost*) or blog entry (*dlpo:BlogPost* as a subclass of *sioct:BlogPost*) to other unresolved non-multimedia links (e.g. online article, web pages, etc.).
4. **PresencePost** - Represents subposts relating to a user's presence. This can refer to not only online presence (*dlpo:AvailabilityPost*) but also physical (*dlpo:ActivityPost*, *dlpo:EventPost* or *dlpo:Checkin*).

The use of the *dlpo:definingResource* property is crucial to achieve the required integration of DLPO instance within

[12] `http://www.semanticdesktop.org/ontologies/nao/`

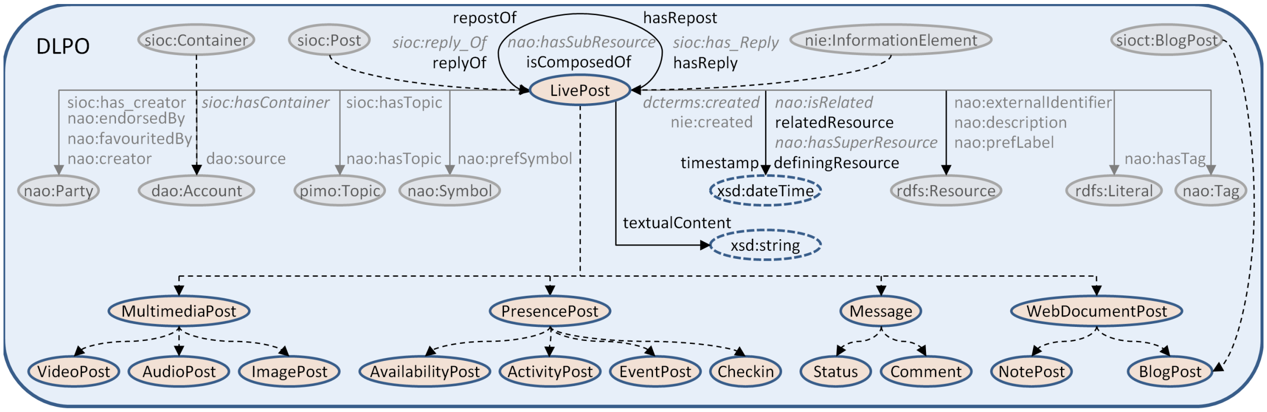


Figure 1: The LivePost Ontology

the PIM. In fact, subtypes are only defined for an online post if it has been determined that it is describing existing PIM items. These items have clearly-defined semantics, and therefore specific subposts will always have specific items as their defining resource, as defined by one of the OSCAF domain ontologies. For example, *dlpo:EventPost* will be related to an *ncal:Event* instance from the Calendar Ontology (NCAL)[13], a *dlpo:Checkin* to a *pimo:Location*, a *dlpo:Video/Image/AudioPost* to a respective instance from the File Ontology (NFO)[14], and an Availability/Activity-Post to one of the available instances provided by the Presence Ontology[15] (DPO) e.g. busy, available, etc. and travelling, working, sleeping, etc.

The fourth modelling requirement is tackled by the introduction and/or adoption of a number of useful properties. These currently include the *dlpo:repostOf* and *dlpo:hasRepost* pair of properties to represent the functionality provided by many online accounts to reshare a (personal or a contact's) post with different social accounts. Additionally, we also reuse the *nao:endorsedBy* and its subproperty *nao:favouritedBy* to represent Social Web functions of 'liking' and favouriting/starring online posts, respectively.

3.2 SEMANTIC LIFTING

This section first explains the transformation from semi-structured posts to a DLPO instance (Sect. 3.2.1), before detailing the syntactic and semantic analysis performed to extract richer information from both semi-structured post metadata and unstructured text. The analysis is decomposed into several tasks, and grouped into one or more semantic annotation pipelines defined using the General Architecture for Text Engineering (GATE) [11] and the ANNIE IE system [10], which includes a variety of algorithms for sentence splitting, gazetteer lookup, etc.

3.2.1 Lifting Semi-Structured data onto the PIM

This phase focuses on the integration of extracted data within the PIM. At this stage, only explicit micropost knowledge is captured and transformed into a DLPO instance. The required XML to RDF[16] transformation is carried out through XSPARQL[17]. The latter combines XQuery[18] and SPARQL[19], thus also enabling the use of external SPARQL endpoints to further enrich the results (e.g. to resolve a location-name based on its coordinates).

Listing 1 is an excerpt of the XSPARQL query created for transforming tweets into *dlpo:StatusMessage* instances. At a glance, it traverses status messages (tweets) contained in an XML document. For each, a `construct` clause creates a new *dlpo:StatusMessage*, and populates it with metadata. The tweet author is created in the nested `construct` clause, and linked to the status message using the *nao:creator* property.

```
prefix xsd : <http://www.w3.org/2001/XMLSchema#>
prefix ...

let $doc := "<xml><statuses>...</statuses></xml>"
let $statuses := $doc/statuses/status

return
  for $status in $statuses
    let $id := $status/id
    let $time := $status/created_at
    let $text := $status/text
    let $user := $status/user
    construct
    { _:stm{data($id)} a dlpo:StatusMessage;
      nao:externalIdentifier {data($id)};
      dlpo:timestamp {data($time)}^^xsd:dateTime;
      dlpo:textualContent {data($text)};
      nao:creator _:c{data($id)}.
      {   let $userId := $user/id
        let $name := $user/name
        let $photoUrl := $user/profile_image_url
        let $description := $user/description
        construct {
          _:c{data($id)} a nco:PersonContact;
            nao:externalIdentifier {data($userId)};
            nco:photo {data($photoUrl)};
            nao:description {data($description)};
            nco:fullname {data($name)}. }}}
```

Listing 1: XSPARQL query for Twitter

3.2.2 Preprocessing Unstructured Data

To maximize the quality of results of micropost analysis, our pipeline starts off with the following operations: emoticons removal, abbreviations substitution/removal (using `noslang.com` as an abbreviations dictionary), part-of-speech (POS) tagging, stop words removal and stemming. These tasks execute in that specific order, since e.g. stop words are necessary for our POS tagger, even though they are usually considered as noise by IE algorithms.

3.2.3 Keyword extraction

The majority of research on keyword/keyphrase extraction concentrates on large collections of formal documents (e.g. research papers, news articles), using techniques requiring domain-specific training. Micropost keyword extraction is more challenging due to various reasons; citing their length, the informality of the language, and the higher diversity of topics as examples. Two algorithms suitable for short text (e.g. tweets) are the TF-IDF (term frequency–inverse document frequency) and TextRank [16]. For our approach, we selected TextRank due to i) a better performance compared to classic TF-IDF [25], ii) the use of POS annotations generated by the pipeline to double the accuracy of the output [14], and iii) a design which may reduce the overhead of dynamically adding new documents for the analysis, since convergence is usually achieved after fewer iterations.

3.2.4 Topic extraction

In our context, topics are one level up in abstraction in comparison to keywords. When extracting topics for a resource, they need not explicitly be in the text. Although they may overlap, a topic is also distinct from a category. Whereas the latter are meant for structuring items under a more-or-less fixed taxonomy, the former are intended to function as high-level markers (or tags). In fact, the envisaged di.me userware will allow the user to extend a pre-defined set of topics (instances of *pimo:Topic*).

[13] `http://www.semanticdesktop.org/ontologies/ncal/`

[14] `http://www.semanticdesktop.org/ontologies/nfo/`

[15] `http://www.semanticdesktop.org/ontologies/dlpo/`— candidate OSCAF vocabulary for the representation of recurring (online and physical) user presence components.

[16] http://www.w3.org/RDF/

[17] http://xsparql.deri.org/

[18] http://www.w3.org/TR/xquery/

[19] http://www.w3.org/TR/sparql11-query/

Techniques for topic discovery from corpora include Latent Dirichlet Allocation (LDA) [2] and Latent Semantic Indexing (LSI) [12]. Both extract topics by clustering words or keyphrases found within. These methods share one important difficulty: finding a meaningful label for the topics. Topic labels are a requirement for the di.me userware—they need to make sense to the user. To overcome this issue, our approach is to involve the user for assigning explicit topics to items in their PIM (including microposts), and recommend topic candidates by finding frequent keywords that co-occur using FP-Growth [13], an algorithm for keyword pattern mining. This algorithm was selected because it offers good performance with limited amount of data, takes minimum support as an argument, leaving the possibility to change how it behaves as the data grows; and is currently one of the fastest approaches to frequent item set mining.

3.2.5 Named-entity extraction

Our NEE task is special since extracted entities are to be mapped onto the user's PIM, which can be seen as a personalised set of entities. Core concepts of the PIMO ontology are also very similar to the generic entities typically extracted by named-entity recognition algorithms (e.g. people, organisations, locations), but they also include more personal (or group) entities (e.g. projects, events, tasks). After extraction, entities will be matched against resources in the PIM, and if that doesn't return any, against the LOD cloud. The matching is syntax-based, comparing named-entity and resource (*rdfs:label*) labels. If several matches are returned, a confidence value will be calculated for each resource, based on keywords extracted from microposts and from descriptive metadata about the resource (e.g. *rdfs:comment*).

NEE taggers based on gazetteers such as [19], [15] or [11] are a good fit for entity extraction where a personal KB may feedback the algorithm with new entities created either directly by the user, or as the result of integrating data from an external KB.

3.2.6 Named-entity & co-reference resolution

To determine orthographic co-reference between terms in natural language text, we use the *orthomatcher* module in GATE. An existing hand-crafted set of rules will be extended for additional entity types handled for di.me, e.g. *pimo:Project*. Due to the short lengths of text, co-reference resolution in microposts is hard. However, grouping them into 'conversations' through (e.g. replies and retweets), the algorithms will have more contextual information to work with. Until entity co-reference is combined with resolution, microposts are not able to enrich the PIM and vice-versa. The fact that "John Doe" and "Mr. Doe" are in fact the same person, is not directly exploitable in semantic lifting. Here, we employ the *Large KB Gazetteer* GATE module to query the PIM in order to create a gazetteer for entity lookup. The results of a similarity measure (1) involving the Levenshtein distance:

$$s_e = \frac{Levenshtein\,(e_i) - l_e}{l_e} \qquad (1)$$

—where s_e is the similarity score for an entity e_i, and l_e is its length—are sorted such that only those with the highest similarity are considered. Resource pairs with a score of above 0.9 are automatically interlinked, whereas pairs between 0.7 and 0.9 will require confirmation through the user interface, or discarded if ignored.

3.2.7 Hyperlink resolution

We make use of regular expressions to determine the type of the resource underneath hyperlinks embedded in microposts, i.e. image, video, audio, document, etc. Their type can also be determined by their content-type (e.g. *image/jpeg* generates an instance of an *dlpo:ImagePost*. In case the content-type is *text/html*, all boilerplate and template/presentation markup around the main textual content is detected and removed using *boilerpipe*. Depending on the type, the resource is then passed on to one of the pipelines for content analysis. The same keyword, topic and NE extraction techniques will here be applied to extract and annotate the relevant post subtype (e.g. *dlpo:ImagePost*, *dlpo:WebDocumentPost*) with topics and tags, as well as link them to matched resources in the PIM or the LOD cloud.

3.2.8 Time window analysis

A user's personal events are crawled from calendaring tools and services and integrated within the PIM as instances of *ncal:Event*. Events are an important source of information, since they usually occur at a specific time and have a limited duration. This conforms to what we refer to as a 'time window', which provides the analysis tasks with context information. In this sense, we have identified two tasks to which this extra context is especially beneficial: NER and co-reference resolution.

In microposts such as *"In the MSM workshop in Lyon. David is giving a great presentation, amazing work!"*, 'David' may be recognised as a named-entity (Person). To determine which specific *pimo:Person* instance this refers to without having more information is hard, if not impossible. Event instances in the PIM may refer to a list of attendees[20]. By taking into consideration a workshop event (instance of *ncal:Event*) that is known to be taking place now, and in which a particular David is also an attendee, will help disambiguate the micropost named-entity for David.

Microposts are usually informal and incomplete, e.g. *"great performance! I love them :)"*. Although no named-entities can be extracted from this text, by adding contextual information and performing co-reference resolution, the results may improve significantly. For example, the author of the above post is known to be attending an event described as "Chemical Brothers concert", where 'Chemical Brothers' has been recognised as a named-entity and disambiguated to the LOD resource `http://dbpedia.org/resource/The_Chemical_Brothers`. Time windowing analysis allows for co-reference resolution to be applied on both the post and the event description, such that 'them' in the former may be resolved to the entity in the latter, resulting in an additional semantic link between two items in the PIM.

4. USE CASES FOR APPLICATION

[20] As extracted from structured or unstructured calendar data—which is not discussed in this paper.

In this section we demonstrate the novelty of our approach through a simple use-case. Fig. 2 depicts how two items (larger rectangles), posted online by user Juan on two different accounts—Facebook and Twitter, are represented as DLPO instances, and integrated within the rest of the PIM ontologies. Once the post from Twitter (left-hand side) is obtained and transformed to RDF, core metadata is immediately generated for the *dlpo:LivePost* instance, consisting of relationships to the source (i.e. Twitter, as a known instance of *dao:Account*/*sioc:Container* in the PIM), to the creator (i.e. Juan Martinez, a known *pimo:Person*) and of other easily-obtained data, such as a timestamp. Since the post contains textual content that does not form part of a URI, an instance of *dlpo:Status* is generated as a subpost, through *dlpo:isComposedOf*. This subpost type only points to a string containing the textual content. Following a successful execution of the next stage in the semantic lifting process, the reference to "@aalford" is matched to the known PIM item 'Anna Alford' (*pimo:Person*). Similarly, the reference to '#Lyon' is matched to another existing item. This time however, the item is not found in the user's PIM, but in an external open KB (e.g. DBPedia). Since not enough information about the semantics of the relationship between the post and these items can be extracted (without as yet resorting to natural language processing techniques to determine e.g. that an arrival to a place amounts to a check-in), the items are loosely linked to the superpost through *dlpo:relatedResource*. Once the hyperlink from the post to the external Lyon representation is established, the same resource is adopted to the PIM as an external instance of *pimo:City*, thus virtually also becoming part of the PIM. In addition, the lemma for 'arrived', together with the reference to a *pimo:City*, are matched to two of the items (keyword "arrive", any location instance) assigned to a pre-defined *pimo:Topic* - Travel. As a result, this post is assigned this topic through *nao:hasTopic*.

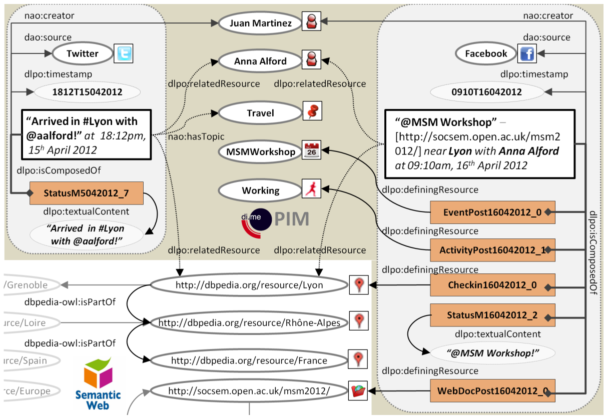


Figure 2: Semantic Integration of Post Concepts

The post retrieved from Facebook (right-hand side) generates another *dlpo:LivePost* instance. The relationship between the post and the 'Anna' and 'Lyon' PIM items is in this case easier to extract, since the information on this source is more structured (Anna is tagged, and Lyon is shown as a nearby location). The improved structure also translates into the generation of a specific *dlpo:Checkin* subpost instance for the superpost, whose defining resource is the same representation for Lyon linked to both external posts through *dlpo:relatedResource*. As the post on Facebook also includes a link, hyperlink resolution is employed to determine that it is not of any specific type (multimedia, blogpost, URI describing a resource, etc.). Thus, an instance of *dlpo:WebDocumentPost* is generated as a subpost, with its defining resource being the URL for the web document itself. It is also determined that the superpost is composed of two other subposts - an *dlpo:EventPost* and an *dlpo:ActivityPost*. The former is discovered after the 'MSM Workshop' named entity is matched to the label of an existing *pimo:Event* instance, thus automatically becoming the subpost's defining resource. Similarly, lemmas from the textual content of the superpost are matched to keywords attached to an existing system-defined instance of the *dpo:Activity* class in the DPO—'Working'. As a result, it is established that the user is posting about a 'Working' activity that is currently in progress. As shown in Fig. 2, our approach enables posts from different sources to be mapped to unique representations of items in both the PIM and the LOD cloud.

5. FUTURE WORK AND CONCLUSIONS

In this paper we have presented an approach for analysing and extracting presence-related information from Social Web sharing activities, in order to enrich a user's integrated Personal Information Model, so as to improve the user's experience in using a proposed intelligent personal information management system—the di.me userware. Apart from being the basis for providing context-aware recommendations, aggregated user presence-related information also becomes more readily available for Social Web sharing.

Our main contribution is the DLPO ontology, which successfully combines aspects from both user presence and online posting domains in a concise ontology, itself integrated within established PIM Knowledge Models. Although most of the discussed techniques for semantic lifting have already been implemented, some in a more advanced state (keyword/topic extraction, hyperlink resolution) than others (NEE, NER & co-reference resolution), the overall approach is still being improved and extended. Apart from discussed advanced features such as time window analysis, next in line for investigation are techniques for a richer semantic analysis of posts, either by additional natural language processing techniques (e.g. to discover implicit user actions in text) or through the exploitation of metadata that is already attached to shared items (e.g. location coordinates from a posted image), through the use of vocabularies such as Rich Snippets, Open Graph protocol, schema.org, RDFa, Microdata and Microformats. This metadata is much more easily-obtained, since it already contains explicit references to entities in the LOD Cloud. Once all of the envisaged functionality is in place and the di.me userware is deployed for use, we will perform an adequate evaluation of our entire approach as we propose it. Evaluation will be performed on three different aspects — i) success of online post decomposition compared against a manual approach, ii) determination of PIM concepts that online posts are mostly/rarely/never linked to, and iii) decomposition of online posts within different social networks, to find out which subtypes are mostly useful for users e.g. checkins - users can see the current location of their contacts in a graph format, event posts - are automatically stored in a personal calendar, activity posts - graph representation showing the activities that a user normally does. In the meantime, we plan to evaluate individual aspects of the overall technique and improve it accordingly.

6. ACKNOWLEDGMENTS

This work is supported in part by the European Commission under the Seventh Framework Program FP7/2007-2013 (*digital.me* – ICT-257787) and in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (*Líon-2*).

7. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *Proceedings of the International Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 1–12, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards semantically-interlinked online communities. In *European Semantic Web Conference (ESWC)*, Lecture Notes on Computer Science, pages 500–514. Springer, 2005.

[4] D. Brickley and L. Miller. Foaf vocabulary specification 0.98, 2010.

[5] A. E. Cano, A.-S. Dadzie, V. S. Uren, and F. Ciravegna. Sensing presence (presense) ontology: User modelling in the semantic sensor web. In *Proceedings of the ESWC Workshop on User Profile Data on the Social Semantic Web (UWeb 2011)*, 2011.

[6] A. E. Cano, S. Tucker, and F. Ciravegna. Capturing entity-based semantics emerging from personal awareness streams. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, pages 33–44, 2011.

[7] I. Celino, D. Dell'Aglio, E. D. Valle, Y. Huang, T. Lee, S. Park, and V. Tresp. Making sense of location based micro-posts using stream reasoning. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, pages 13–18, 2011.

[8] J. Chang and E. Sun. Location: How users share and respond to location-based data on social networking sites. In *Proceedings of the AAAI Conference on Weblogs and Social Media*, pages 74–80, 2011.

[9] S. Choudhury and J. Breslin. Extracting semantic entities and events from sports tweets. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, pages 22–32, 2011.

[10] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[11] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (V.6)*. 2011.

[12] S. Deerwester. Improving Information Retrieval with Latent Semantic Indexing. In *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, Atlanta, Georgia, 1988. American Society for Information Science.

[13] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1):53–87, Jan. 2004.

[14] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[15] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[17] A. Miles and S. Bechhofer. Skos simple knowledge organization system reference, 2009.

[18] A. Passant, U. Bojars, J. Breslin, T. Hastrup, M. Stankovic, and P. Laublet. An overview of smob 2: Open, semantic and distributed microblogging. pages 303–306, 2010.

[19] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[20] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named Entity Recognition in Tweets: An Experimental Study. In *EMNLP*, 2011.

[21] O. Sacco and A. Passant. A privacy preference ontology (ppo) for linked data. In *Proceedings of the Linked Data on the Web Workshop (LDOW)*, 2011.

[22] M. Sintek, S. Handschuh, S. Scerri, and L. van Elst. Technologies for the social semantic desktop. In *Reasoning Web. Semantic Technologies for Information Systems*, Lecture Notes in Computer Science, pages 222–254. Springer-Verlag, 2009.

[23] M. Stankovic and J. Jovanovic. Online presence in social networks. In *Proceedings of the W3C Workshop on the Future of Social Networking*, 2009.

[24] T. Steiner, A. Brousseau, and R. Troncy. A tweet consumers' look at twitter trends. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, 2011.

[25] W. Wu, B. Zhang, and M. Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 689–692, Stroudsburg, PA, USA, 2010.

[26] K. Zoltan and S. Johann. Semantic analysis of microposts for efficient people to people interactions. In *Proceedings of the 10th Roedunet International Conference (RoEduNet), 2011*, pages 1–4, 2011.

Visualizing Contextual and Dynamic Features of Micropost Streams

Alexander Hubmann-Haidvogel, Adrian M.P. Braşoveanu,
Arno Scharl, Marta Sabou, Stefan Gindl
MODUL University Vienna
Department of New Media Technology
Am Kahlenberg 1
1190 Vienna, Austria

{alexander.hubmann, adrian.brasoveanu, arno.scharl,
marta.sabou, stefan.gindl}@modul.ac.at

ABSTRACT

Visual techniques provide an intuitive way of making sense of the large amounts of microposts available from social media sources, particularly in the case of emerging topics of interest to a global audience, which often raise controversy among key stakeholders. Micropost streams are context-dependent and highly dynamic in nature. We describe a visual analytics platform to handle high-volume micropost streams from multiple social media channels. For each post we extract key contextual features such as location, topic and sentiment, and subsequently render the resulting multi-dimensional information space using a suite of coordinated views that support a variety of complex information seeking behaviors. We also describe three new visualization techniques that extend the original platform to account for the dynamic nature of micropost streams through dynamic topography information landscapes, news flow diagrams and longitudinal cross-media analyses.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces – interaction styles. I.3.6. [Computer Graphics]: Methodology and Techniques – Interaction Technique.

General Terms

Algorithms, Measurement, Design, Human Factors

Keywords

Social Media Analytics, Microposts, Contextual Features, News Flow, Dynamic Visualization, Information Landscape

1. INTRODUCTION

The ease of using social media channels has enabled people from around the world to express their opinions and propagate local or global news about virtually every imaginable topic. In doing so, they most often make use of short text messages (tweets, status updates) that we collectively refer to as microposts. Given the high volume, diversity and complex interdependency of social media-specific micropost streams, visual techniques play an increasingly important role in making sense of these novel data sources. Visual techniques can support analysts, journalists and marketing managers alike in taking the pulse of public opinion, in understanding the perceptions and preferences of key stakeholders, in detecting controversies, and in measuring the impact and diffusion of public communications. This is particularly true for domains that pose challenges through their global reach, the competing interests of many different stakeholders, and the dynamic and often conflicting nature of relevant evidence sources (e.g., environmental issues, political campaigns, financial markets).

To support such scenarios across application domains, we have developed a (social) media monitoring platform with a particular focus on visual analytics (Hubmann, 2009). The platform enables detecting and tracking topics that are frequently mentioned in a given data sample (e.g., a collection of Web documents crawled from relevant sources). The advanced data mining techniques underlying the platform extract a variety of contextual features from the document space. A visual interface based on multiple coordinated views allows exploring the evolution of the document space along the dimensions defined by these contextual features (temporal, geographic, semantic, and affective), and subsequent drill-down functionalities to analyze details of the data itself. In essence, the platform has the key characteristics of a decision support system, namely: 1) it aggregates data from many diverse sources; 2) it offers an easy to use visual dashboard for observing global trends; 3) it allows both a quick drill-down and complex analyses along the dimensions of the extracted contextual features. We briefly report on the overall platform in Section 3.

The platform has been originally designed to analyze traditional news media, but from early 2011 we have adapted it to support micropost analysis, taking advantage of the robust infrastructure for crawling, analyzing and visualizing Web sources. The multi-dimensional analysis enabled by the original design of the portal is well suited for analyzing contextual features of microposts. However, the visualization metaphors did not properly capture the highly dynamic nature of micropost streams, nor did they allow cross-comparison between social and traditional media sources. Our latest research therefore focuses on novel methods to support temporal analysis and cross-media visualizations. In sections 4, 5 and 6 we describe these novel visualizations.

2. RELATED WORK

With the rise of the social networks (Heer, 2005), understanding large-scale events through visualization emerged as an important research topic. Various visual interfaces have been designed for inspecting news or social media streams in diverse domains such as sports (Marcus, 2011), politics, (Diakopoulos, 2010; Shamma, 2009; Shamma, 2010), and climate change (Hubmann, 2009).

Published as part of the #MSM2012 Workshop proceedings, available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

Researchers have emphasized different aspects of extracting useful information including (sub-)events (Adams, 2011), topics (Hubmann, 2009), and video fragments (Diakopoulos, 2011). *Vox Civitas*, for example, is a visual analytic tool that aims to support journalists in getting useful information from social media streams related to televised debates and speeches (Diakopoulos, 2010). In terms of the number and type of social media channels that are visualized, most approaches focus primarily on Twitter, while streams from Facebook and YouTube are visualized to a lesser extent (Marcus, 2010). We regard these three channels as equally important and visualize their combined content.

To reflect the dynamic nature of social media channels, some visualizations provide real-time updates displaying messages as they are published, and also projecting them onto a map – e.g., *TwitterVision.com* or *AWorldofTweets.frogdesign.com*. Given the computational overhead, however, real-time visualizations are the exception rather than the norm, since most projects rely on update times anywhere between a few minutes and a few days.

Visual techniques render microposts along dimensions derived from their contextual features. Most frequently, visualization rely on temporal and geographic features, but increasingly they exploit more complex characteristics such as the sentiment of the micropost, its content (e.g., expressed through relevant keywords), or characteristics of its author. Indeed, user clustering as seen in *ThemeCrowds* (Archambault, 2011) or geographical maps (e.g., (Marcus, 2011), *TwitterReporter* (Meyer, 2011)) are must-have features for every system that aims to understand local news and correlate them with global trends. Commercial services such as *SocialMention.com* and *AlertRank.com* use visualizations to track sentiment across tweets. During the 2010 U.S. Midterm Elections, sentiment visualizations have been present in all major media outlets from *New York Times* to *Huffington Post* (Peters, 2010).

Fully utilizing contextual features requires the use of appropriate visual metaphors. In general, social media visualizations rely on one of the following three visual metaphors:

- *Multiple Coordinated Views*, also known as linked or tightly coupled views in the literature (Scharl, 2001), (Hubmann, 2009), ensures that a change in one of the views triggers an immediate update within the others. For example, the interface of *Vox Civitas* uses coordinated views to synchronize a timeline, a color-coded sentiment bar, a Twitter flow and a video window which helps linking parts of the video to relevant tweets (Diakopoulos, 2010). Additionally, (Marcus, 2011) use the multiple coordinated views in their system geared towards Twitter events and offer capabilities to drill down into sub-events and explore them based on geographic location, sentiment and link popularity.
- *Visual Backchannels* (Dork, 2010) represent interactive interfaces synchronizing a topic stream (e.g., a video) with real-time social media streams and additional visualizations. This concept has evolved from the earlier concept of *digital backchannel*, referring to news media outlets supplementing their breaking news coverages with relevant tweets – e.g., during political debates or sport games (Shamma, 2010). However, additionally to the methods described in *Hack the Debate* (Shamma, 2009) and *Statler* (Shamma, 2010), tools that use the visual backchannel metaphor display not only the Twitter flow that corresponds to certain media events such as debates, but also a wealth of graphics and statistics.
- *Timelines* follow the metaphor with the longest tradition, well suited for displaying the evolution of topics over time. Aigner et al. present an extensive collection of commented timelines (Aigner, 2011). The work by Adams et al. (Adams, 2011) is similar to our approach as it combines a color-coded sentiment display with interactive tooltips.

Beyond understanding micropost streams, a challenging research avenue compares the content of social media coverage with that of traditional news outlets. Cross-media analysis based on social sources is a relatively new field, but promising results have been published recently. In most cases comparisons are made between two sources such as Twitter and New York Times (Zhao, 2011), or Twitter and Yahoo! News (Hong, 2011). (Zhao, 2011) compares a Twitter corpus with a New York Times corpus to detect trending topics. For the New York Times, they apply a direct Latent Dirichlet Allocation (LDA), while for Twitter they use a modified LDA under the assumption that most tweets refer to a single topic. They use metrics including the distribution of categories, breadth of topics coverage, opinion topic and the spread of topics through re-tweets, and show that most Twitter topics are not covered appropriately by traditional news media channels. They conclude that for spreading breaking world news, Twitter seems to be a better platform than a traditional medium such as New York Times. Hong et al. compare Twitter with Yahoo! News to understand temporal dynamics of news topics (Hong, 2011). They show that local topics do not appear as often in Twitter, and they go on to compare the performance of different models (LDA, Temporal Collection, etc). (Lin, 2011) conducts a study on media biasing on both social networks and news media outlets, but is focused only on the quantity of mentions. While these studies highlight differences between social and news media, they typically lack visual support for monitoring diverse news sources.

3. ACQUISITION AND AGGREGATION OF CLIMATE CHANGE MICROPOSTS

Climate Change is a global issue characterized by diverse opinions of different stakeholders. Understanding the key topics in this area, their global reach and the opinions voiced by different parties is a complex task that requires investigating how these dimensions relate to each other. The *Media Watch on Climate Change* portal (www.ecoresearch.net/climate) addresses this task by providing advanced analytical and visual methods to support different types of information seeking behavior such as browsing, trend monitoring, analysis and search.

The underlying technologies have originally been developed for monitoring traditional news media (Hubmann, 2009) and have recently been adapted for use with social media sources, in particular micropost content harvested from *Twitter, YouTube* and *Facebook*. Between April 2011 and March 2012, the system has collected and analyzed an estimated 165,000 microposts from these channels. To support a detailed analysis of the collected microposts, we use a variety of visual metaphors to interact with contextual features along a number of dimensions: temporal, geographic, semantic and attitudinal. A key strength of the interface is the rapid synchronization of *multiple coordinated views*. It allows selecting the relevant data sources and provides trend charts, a document viewing panel as well as just-in-time information retrieval agents to retrieve similar documents in terms of either topic or geographic location. The right side of the interface contains a total of four different visualizations (two of which are being shown in Figure 1), which capture global views on the dataset. In addition to the shown semantic map (= information landscape; see Section 4) and tag cloud, users can also select a geographic map and an ontology graph. Any of these views can be closed, maximized, or opened in a separate pop-up window to allow a more thorough inspection (the views remain synchronized even when placed in different windows). While the *Media Watch on Climate Change* focuses on environmental coverage, the same technologies are currently being used for other domains as well, for example, for the Web intelligence platforms of the National Oceanic and Atmospheric Administration (NOAA), the National Cancer Institute (NCI), and the Vienna Chamber of Commerce and Industry (see www.weblyzard.com).

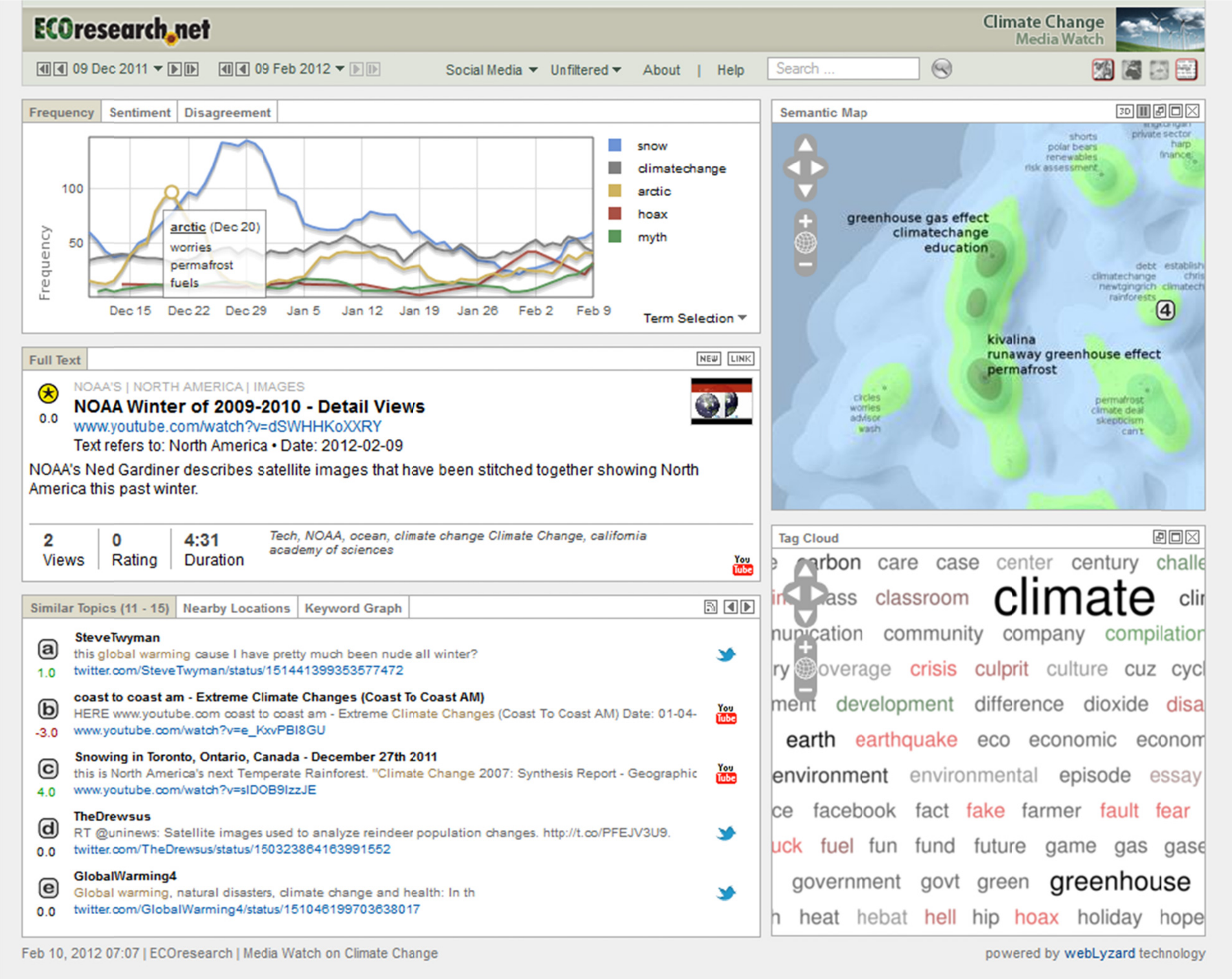
The portal's visualizations provide a good starting point for analyzing microposts along a variety of contextual features, in particular in the area of climate change. The system does not reflect the dynamic character of these micropost streams, however, and therefore misses a key benefit of social media – that of capturing events as they unfold. To overcome this limitation, we are currently developing the following set of novel visualizations that focus on the longitudinal and temporal analysis of micropost streams:

1. The dynamic topography information landscapes are an extension of the information landscapes paradigm. Instead of capturing the state of the information space at discrete moments in time, the topography is continuously updated as new microposts are being published (Section 4).
2. The news flow diagrams visualize microposts from multiple social media channels in real time, and reveal correlations in terms of the topics that they mention (Section 5).
3. The cross-media analysis charts allow longitudinal analyses of frequency and sentiment for any given topic and across data sources (e.g., between social media, news media, and the blogosphere; see Section 6).

4. DYNAMIC TOPOGRAPHY INFORMATION LANDSCAPE

Information Landscapes represent a powerful visualization technique for conveying topical relatedness in large document repositories (Krishnan, 2007). Yet, the traditional concept of information landscapes only allows for visualizing static conditions. We have made use of such static landscapes when visualizing traditional news media, which were less dynamic than social media sources and where it was sufficient to recompute the information landscape at weekly intervals.

For visualizing highly dynamic micropost streams, however, this is not a satisfying solution. What is required instead is a visual representation such as *ThemeRiver* (Havre, 2002) that conveys changes in topical clusters. Unfortunately, most of these representations lack the means to express complex topical relations and are therefore no substitute for the information landscape metaphor.



NOAA'S | NORTH AMERICA | IMAGES

NOAA Winter of 2009-2010 - Detail Views

0.0

www.youtube.com/watch?v=dSWHHKoXXRY

Text refers to: North America • Date: 2012-02-09

NOAA's Ned Gardiner describes satellite images that have been stitched together showing North America this past winter.

2 Views | 0 Rating | 4:31 Duration

Tech, NOAA, ocean, climate change Climate Change, california academy of sciences

(a) 1.0 **SteveTwyman**
this global warming cause I have pretty much been nude all winter?
twitter.com/SteveTwyman/status/151441399353577472

(b) -3.0 **coast to coast am - Extreme Climate Changes (Coast To Coast AM)**
HERE www.youtube.com coast to coast am - Extreme Climate Changes (Coast To Coast AM) Date: 01-04-
www.youtube.com/watch?v=e_KxvPBI8GU

(c) 4.0 **Snowing in Toronto, Ontario, Canada - December 27th 2011**
this is North America's next Temperate Rainforest. "Climate Change 2007: Synthesis Report - Geographic
www.youtube.com/watch?v=sIDOB9IzzJE

(d) 0.0 **TheDrewsus**
RT @uninews: Satellite images used to analyze reindeer population changes. http://t.co/PFEJV3U9.
twitter.com/TheDrewsus/status/150323864163991552

(e) 0.0 **GlobalWarming4**
Global warming, natural disasters, climate change and health: In th
twitter.com/GlobalWarming4/status/151046199703638017

Feb 10, 2012 07:07 | ECOresearch | Media Watch on Climate Change

Figure 1. Screenshot of the Media Watch on Climate Change (www.ecoresearch.net/climate)

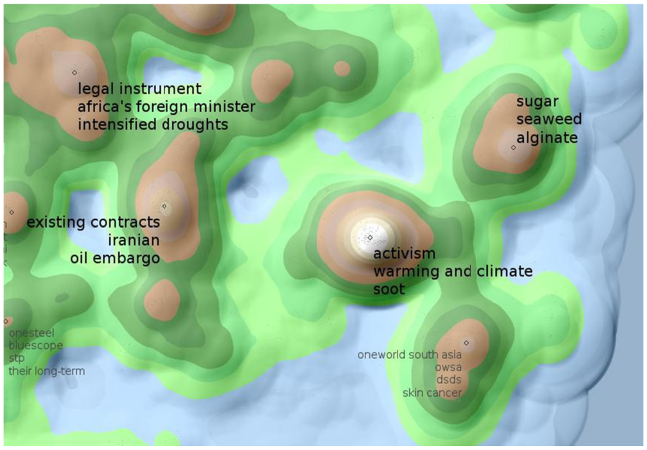


Figure 2. Information landscape on climate change coverage

In previous research, we have introduced dynamic topography information landscapes (Sabol et al., 2010) to address both topical relatedness and rapidly changing data. Dynamic topography information landscapes are visual representations based on a geographic map metaphor where topical relatedness is conveyed through spatial proximity in the visualization space with hills representing agglomerations (clusters) of topically similar documents. As shown in Figure 2, the hills are labeled with sets of dominant keyword labels (n-grams) extracted from the underlying documents to facilitate the users' orientation.

Micropost streams are characterized by the rapid emergence and decay of topics. The topical structure changes with each new posting. Dynamic information landscapes convey these changes as tectonic processes which modify the landscape topography accordingly. Rising hills indicate the emergence of new topics; shrinking hills a fading of existing ones. In the process of generating information landscapes, high-dimensional data is projected into a lower-dimensional space.

5. NEWS FLOW DIAGRAM

While the dynamic topography information landscape metaphor depicts the evolution of topic clusters within a collection of microposts without differentiating their origin, some scenarios require a comparative analysis of individual micropost streams. The two key problems related to the visualization of microposts originating from multiple social media sources is to show their provenance as well as the dynamic changes of topical associations between them. The *News Flow Diagram* concept addresses these issues by integrating several visual metaphors into a single display (see screenshot in Figure 3):

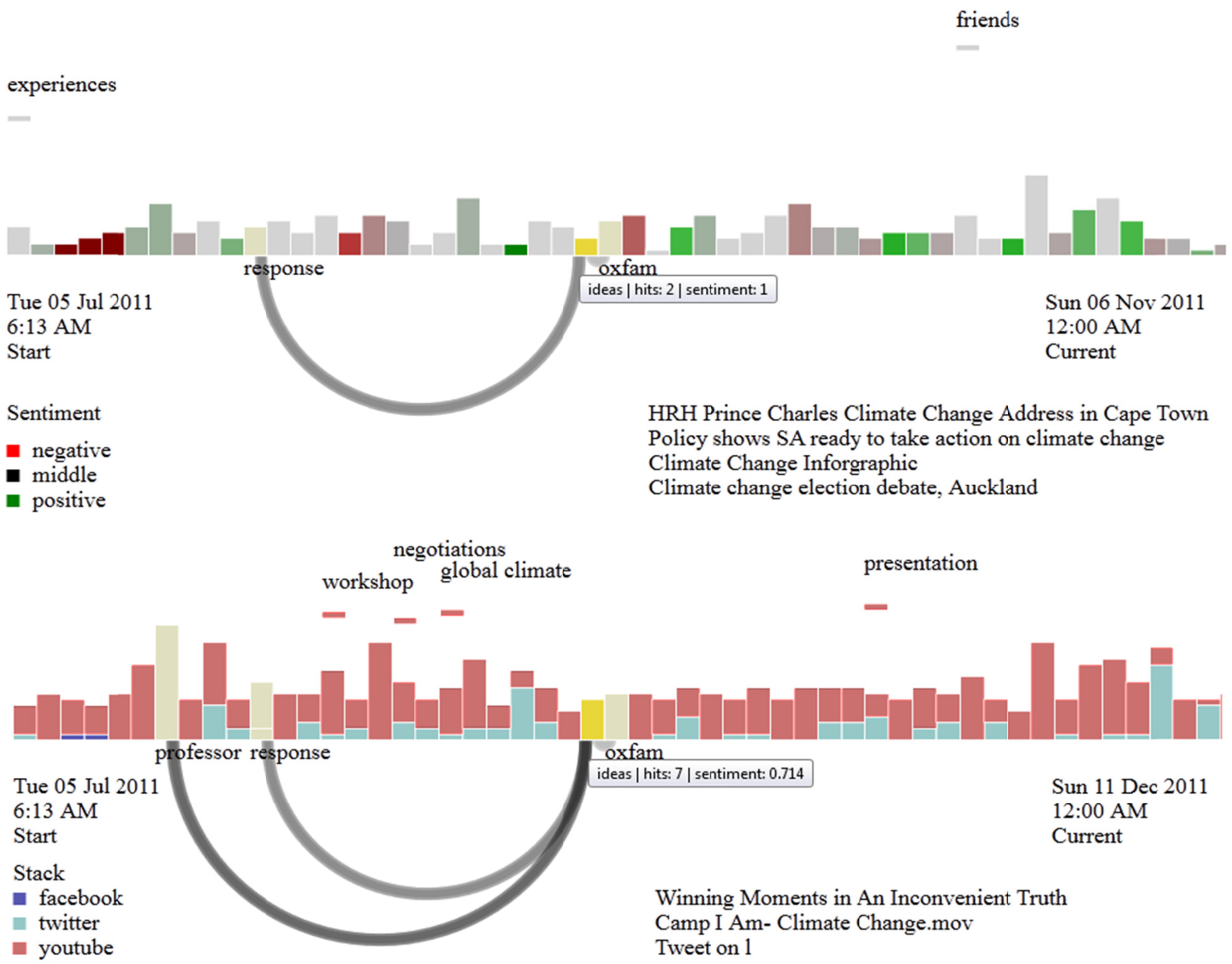
1. *Falling bar graphs* – Each micropost (Twitter message, Facebook status update, YouTube message) is represented internally through the title of the post, its time of publishing, its content, and a list of associated keywords. When a new micropost is posted we visualize, in real-time, its respective associated keywords through falling words. One document generates one falling word for each mentioned topic. The falling words will "hit" the lower part of the visualization and dissolve into the corresponding keyword bar, which increases in size accordingly. Figure 3 depicts how topics fall towards their respective keyword bars (e.g., “experiences” and “friends” in the upper diagram). A keyword bar collects all mentions of a certain topic in microposts from different social media channels and, therefore, its height correlates with the popularity of the topic in the social media outlets that are visualized. The falling bar metaphor was quite popular a few years ago due to the success of the *Digg Stack* visualization [Baer 2008].
2. *Multi-source stacked bars and color-coded sentiment bars*. Each falling word is color-coded to represent either its provenance or its associated sentiment value. Figure 3, for example, uses the color of the falling words to reflect their origin (Twitter: gray, Facebook: blue marine, YouTube: red). This color-coding is maintained in the keyword bars, each bar showing through its diversely colored portions the percentage of mentions of the corresponding keyword within the individual media sources. This allows inferring the most and least mentioned topics across sources. The same metaphor can be used to show sentiment values instead of provenance (not shown in Figure 3).
3. *Threaded arcs*. We use an adaptation of the threaded arcs display to convey associations between the keywords that appear in the same document. Dynamic link patterns conveyed through shifting arcs allow us to understand how the associations, initially displayed through falling bars, modify over time. Related keywords are highlighted to quickly notice them. Figure 3b shows the topic “ideas”, which has appeared seven times, co-occurs most frequently with the two topics: “professor” and “response”. These threaded arcs are only displayed when we click on a keyword bar.

Color-coding is an important part of this visualization. We use it to highlight various aspects of the data:

- *Sentiment coloring* – the color of the bars can represent the sentiment of a certain topic (see Figure 3a);
- *Source coloring* – words can also be displayed as stacked bars with specific colors that represent the sources in the stacked layout (provenance information); Figure 3b, for example, shows a situation where we have three sources (Facebook, Twitter, YouTube) and keywords from one source (YouTube) falling;
- *Arc coloring* – we use darker shades of gray for stronger relations between the terms (i.e., they co-occur more frequently), connecting the most related terms.

To demonstrate how associations evolve over time, we show the same word (“ideas”) in both diagrams: Figure 3a uses color coding for sentiment information, Figure 3b for distinguishing the source (users can easily switch between both modes). In Figure 3a, “ideas” has a stronger connection with “response” than with “oxfam”, but the word has only two hits. Figure 3b shows that after seven hits, “ideas” has a stronger connection with “professor”, than with “response”, and also the same weak link with “oxfam”. Future versions of the visualization module will include information related to these connections in the tooltips, emphasizing the importance of interactivity and revealing the evolution of connections over time.

This visualization showcases the powerful mechanism of combining various visual metaphors with color-coding. We use the news flow diagrams to identify key topics (we only show the 50 most important terms), to describe the relations between them (co-occurrence of terms in a micropost are displayed through the falling bars), and to show the evolution of social media coverage over time (dynamic changes in the distribution of keywords/topics across various social media sources is represented through the lower arcs).



HRH Prince Charles Climate Change Address in Cape Town
Policy shows SA ready to take action on climate change
Climate Change Inforgraphic
Climate change election debate, Auckland

Winning Moments in An Inconvenient Truth
Camp I Am- Climate Change.mov
Tweet on 1

Figure 3. News Flow Diagram with color coding for showing sentiment (Figure 3a, above) and source (Figure 3b, below).

6. CROSS-MEDIA ANALYSIS

The Media Watch on Climate Change offers longitudinal analysis (i.e., monitoring over time) in terms of topic frequency, sentiment associated to a topic and disagreement over a topic. However, these trend charts are only plotted over a single data source (e.g., either news media or social media) and are available only for a set of pre-computed topics. Therefore they are neither suitable for social media streams where new topics emerge rapidly, nor do they allow comparing across different media sources.

To overcome these limitations, we are developing the new visualization shown in Figure 4, which allows (a) defining a topic to be monitored over time and (b) monitoring this topic across different media sources selected by clicking the appropriate check-boxes in the interface (e.g., traditional news media outlets, blogs, social networks such as Twitter, YouTube and Facebook). The visualization makes use of the data collection and charting frameworks of the portal to plot both frequency and sentiment related charts.

By plotting topic frequency (i.e., number of documents per day that mention that topic) over time, this visualization shows the impact of a topic on different media sources. For example, the screenshot in Figure 4 depicts a query for the topic "durban" and compares the amount of news coverage about the *17th Conference of the Parties to the United Nations Framework Convention on Climate Change* (COP17) held in Durban, South Africa, from 01 Nov to 31 Dec 2011 in traditional news media, Twitter postings, blogs, and NGOs. Coinciding with the beginning of the conference on the 28th of November, both samples show an increase in the coverage of this topic. The frequency then declines sharply after the end of the event, which is an effect more pronounced in the news media coverage. It also shows that coverage of the conference has been far more intense in news media than in micropost streams, except a short period of time in December.

In addition to frequency charts, we also visualize the sentiment of the documents mentioning a specific topic. A set of charts shows either positive or negative documents, average sentiment of documents for a day or the standard deviation of the sentiment over time. These charts help to understand the attitude expressed in different media outlets, e.g., which outlet has the most negative or positive documents, which outlet is characterized by the most controversies? A comparison of the average sentiment towards "COP17" in social media and news media channels showed a gradual shift from positive to negative in microposts, while news media sentiment remained positive during the entire duration of the event (see screenshot in Figure 4).

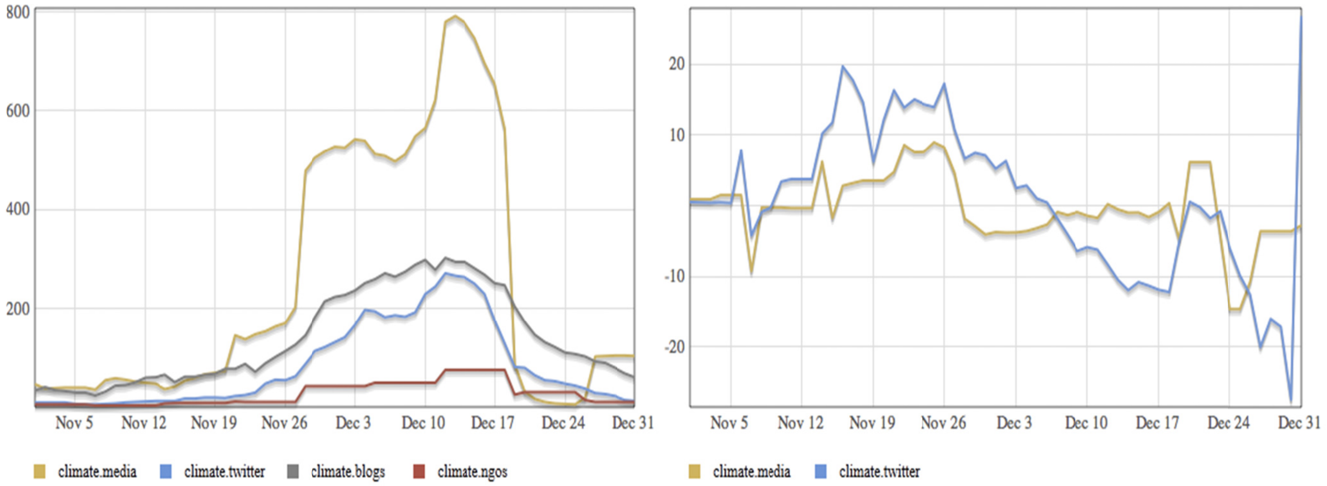


Figure 4. Cross-media analysis for November – December 2011; term frequency distribution for “Durban” (left) and average sentiment towards “COP17” (right).

An important issue of visualizing sentiment across media outlets is the meaningful computation of sentiment values for disparate documents. The sentiment detection algorithm cumulatively adds up the sentiment values of individual words in a document to compute an overall sentiment value for the document, which is then normalized based on the total number of tokens in the document. This allows the comparison of documents of different lengths, such as news articles and microposts.

The visualization can not only track user-specified topics, but can also assist the user in finding similar topics by providing a list of top terms associated with the query term. These associated terms are calculated using a combination of significant phrases detection and co-occurrence analysis on the document set (Hubmann, 2009), and are aggregated and ranked depending on documents matching the query term. A query for "COP17", for example, yields the terms "Durban", "UNFCC" and "Climate Change" as associated terms in Twitter microposts. Additional query term disambiguation is not required in this case, as the documents collected are already pre-filtered based on their relevance to the climate change domain.

7. CONCLUSION AND OUTLOOK

In this paper we describe recent work on making sense of microposts through visual means. Our earlier work on the *Media Watch on Climate Change* portal (www.ecoresearch.net/climate) focused on visual analytics over traditional news media and relied on extracting and visualizing a wealth of context features. This characteristic of the portal proved essential when adapting it to visualizing micropost streams from three main social media channels, as it enabled complex analysis along temporal, geographic, semantic and attitudinal dimensions in the challenging domain of climate change. Unlike many other social media visualizations, the presented approach relies on a robust infrastructure and combines data from multiple social media outlets.

While the contextual nature of the microposts has been fully capitalized upon, the existing visualizations fell short of conveying another key characteristic of microposts, namely their dynamic nature. This initiated research into the new visualizations described in this paper, including: (i) dynamic topographic information landscapes, which show through tectonic changes how major topic clusters evolve; (ii) the news flow diagrams which enable a fine-grained, comparative analysis across micropost streams, showing key topics being discussed and how they relate to each other; and (iii) cross-media analysis based on longitudinal datasets containing frequency and sentiment information.

Future work will focus on feature extraction from microposts and visualizations to depict contextual and dynamic characteristics of microposts. We are currently working on more robust methods for extracting contextual features from microposts, by further adapting our current methods to the particularities of these texts. Some of the features that we intend to introduce in future releases are related to interactive timelines and time series analysis.

Future research will also allow comparing timelines across topics and related to specific events. We will use timelines as a starting point for narrative visualizations (e.g., replaying the history of an event or a chain of events; identifying visual patterns that best describe a chain of events on social media). We will compare various media channels since our datasets and graphical tools are well suited for such an analysis. Finally, we will investigate novel ways of incorporating these individual visualizations to support the complex analytical scenarios of decision making tools.

8. ACKNOWLEDGMENTS

Key components of the system presented in this paper were developed within the DIVINE (www.weblyzard.com/divine) and Triple-C (www.ecoresearch.net/triple-c) research projects, funded by *FIT-IT Semantic Systems* of the Austrian Research Promotion Agency and the *Austrian Climate Research Program* of the Austrian Climate and Energy Fund (www.ffg.at; klimafonds.gv.at).

Adrian Braşoveanu was partially supported by the strategic grant POSDRU/88/1.5/S/60370 (2009) on "Doctoral Scholarships" of the Ministry of Labor, Family and Social Protection, Romania, co-financed by the European Social Fund – Investing in People.

9. REFERENCES

[Adams, 2011] Brett Adams, Dinh Phung, Svetha Venkatesh. 2011. Eventscapes: visualizing events over times with emotive facets. In *MM '11 Proceedings of the 19th ACM International Conference on Multimedia*, Scottsdale, AZ, USA (November 28 - December 01, 2011), 1477-1480.

[Aigner, 2011] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann Christian Tominski. 2011. *Visualization of Time-Oriented Data.* Springer, 2011.

[Archambault, 2011] D. Archambault, D. Greene, P. Cunningham, and N. Hurley. ThemeCrowds: Multiresolution Summaries of Twitter Usage. In Proc. of the 3rd Workshop on Search and Mining User-generated Contents, Glasgow, UK, October 2011.

[Baer, 2008] K. Baer. Information Design Workbook. Graphic Approaches, Solutions and Inspiration + 30 Case Studies. Rockport Publishers, 2008.

[Diakopoulos, 2010] N. Diakopoulos, M. Naaman, F. Kivranswain. 2010. Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry, *IEEE Symposium on Visual Analytics Science and Technology (VAST)*.

[Dork, 2010] Marian Dork, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. A visual backchannel for large-scale events. TVCG: Transactions on Visualization and Computer Graphics, 16(6):1129-1138, Nov/Dec 2010.

[Havre, 2002] Havre, S., Hetzler, E., Whitney, P., and Nowell, L.: ThemeRiver: Visualizing Thematic Changes in Large Document Collections. IEEE Transactions on Visualization and Computer Graphics, 8(1):9–20, 2002.

[Heer, 2005] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization*, 2005.

[Hong, 2011] Liangjie Hong, Byron Dom, Siva Gurumurthy, Kostas Tsioutsiouliklis. 2011. A Time-Dependent Topic Model for Multiple Text Streams. *KDD'11*, August 21–24, 2011, San Diego, California, USA.

[Hubmann, 2009] Alexander Hubmann-Haidvogel, Arno Scharl, and Albert Weichselbraun. 2009. Multiple coordinated views for searching and navigating Web content repositories. *Inf. Sci.* 179, 12 (May 2009), 1813-1821.

[Krishnan, 2007] Krishnan, M., Bohn, S., Cowley, W., Crow, V., and Nieplocha, J. 2007. Scalable visual analytics of massive textual datasets. 21st IEEE International Parallel and Distributed Processing Symposium. IEEE Computer Society.

[Lin, 2011] Yu-Ru Lin, James P. Bagrow, David Lazer. 2011. More Voices Than Ever? Quantifying Media Bias in Networks. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[Marcus, 2011] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI '11). ACM, New York, NY, USA, 227-236.

[Meyer, 2011] B. Meyer, K. Bryan, Y. Santos, Beomjin Kim. 2011. TwitterReporter: Breaking News Detection and Visualization through the Geo-Tagged Twitter Network. In: *Proceedings of the ISCA 26th International Conference on Computers and Their Applications*, March 23-15, 2011, Holiday Inn Downtown-Superdome, New Orleans, Louisiana, USA. ISCA 2011, 84-89.

[Peters, 2010] M. Peters. 2010. Four Ways to Visualize voter Sentiment for the Midterm Elections. *Mashable Social Media.* http://mashable.com/2010/10/29/elections-data-visualizations/.

[Sabol, 2010] Sabol, V., Syed, K.A.A., et al. 2010. Incremental Computation of Information Landscapes for Dynamic Web Interfaces. 10th Brazilian Symposium on Human Factors in Computer Systems (IHC-2010). M.S. Silveira et al. Belo Horizonte, Brazil: Brazilian Computing Society: 205-208.

[Scharl, 2001] Scharl, A. 2001. Explanation and Exploration: Visualizing the Topology of Web Information Systems, *International Journal of Human-Computer Studies*, 55(3): 239-258.

[Shamma, 2009] David A. Shamma, Lyndon Kennedy, Elizabeth F. Churchill. 2009. Tweet the Debates: Understanding Community Annotation of Uncollected Sources. In *The first ACM SIGMM Workshop on Social Media (WSM 2009)*, October 23, 2009, Beijing, China.

[Shamma, 2010] David A. Shamma, Lyndon Kennedy, Elizabeth F. Churchill. 2010. Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? In *ACM Conference on Computer Supported Cooperative Work (CSCW 2010)*, February 6-10, 2010, Savannah, Georgia, USA.

[Zhao, 2011] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan and Xiaoming Li. 2011. Comparing Twitter and Traditional Media using Topic Models. in *Advances in Information Retrieval - 33rd European Conference on IR Research*, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings. Lecture Notes in Computer Science 6611, Springer 2011, 338-349.

When social bots attack: Modeling susceptibility of users in online social networks

Claudia Wagner
Institute for Information and
Communication Technologies
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Silvia Mitter
Knowledge Management
Institute
Graz University of Technology
Graz, Austria
smitter@student.tugraz.at

Christian Körner
Knowledge Management
Institute
Graz University of Technology
Graz, Austria
christian.koerner@tugraz.at

Markus Strohmaier
Knowledge Management
Institute and Know-Center
Graz University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Social bots are automatic or semi-automatic computer programs that mimic humans and/or human behavior in online social networks. Social bots can attack users (targets) in online social networks to pursue a variety of latent goals, such as to spread information or to influence targets. Without a deep understanding of the nature of such attacks or the susceptibility of users, the potential of social media as an instrument for facilitating discourse or democratic processes is in jeopardy. In this paper, we study data from the Social Bot Challenge 2011 - an experiment conducted by the WebEcologyProject during 2011 - in which three teams implemented a number of social bots that aimed to influence user behavior on Twitter. Using this data, we aim to develop models to (i) identify susceptible users among a set of targets and (ii) predict users' level of susceptibility. We explore the predictiveness of three different groups of features (network, behavioral and linguistic features) for these tasks. Our results suggest that susceptible users tend to use Twitter for a conversational purpose and tend to be more open and social since they communicate with many different users, use more social words and show more affection than non-susceptible users.

Keywords

social bots, infection, user models

1. INTRODUCTION

Online social networks (OSN) like Twitter or Facebook are powerful instruments since they allow reaching millions of users online. However, in the wrong hands they can also be used to spread misinformation and propaganda, as one could for example see during the US political elections [9]. Recently a new breed of computer programs so-called *social media robots* (short *social bots* or *bots*) emerged in OSN. Social bots are automatic or semi-automatic computer programs that mimic humans and/or human behavior in OSN. Social bots can be directed to attack users (targets) to pursue a variety of latent goals, such as to spread information or to influence users [7]. Recent research [1] highlights the danger of social bots and shows that Facebook can be infiltrated by social bots sending friend requests to users. The average reported acceptance rate of such friend requests was 59.1% which also depended on how many mutual friends the social bots had with the infiltrated users, and could be up to 80%. This study clearly demonstrates that modern security defenses, such as the Facebook Immune System, are not prepared for detecting or stopping a large-scale infiltration caused by social bots.

We believe that modern social media security defenses need to advance in order to be able to detect social bot attacks. While identifying social bots is crucial, identifying users who are susceptible to such attacks - and implementing means to protect against them - is important in order to protect the effectiveness and utility of social media. In this paper, we define a *target* to represent a user who has been singled out by a social bot attack, and a *susceptible user* as a user who has been infected by a social bot (i.e. the user has in some way cooperated with the agenda of a social bot). This work sets out to identify factors which help detecting users who are susceptible to social bot attacks. To gain insights into these factors, we use data from the Social Bot Challenge 2011 and introduce three different groups of features: network features, behavioral features and linguistic features. In total, we use 97 different features to first *predict infections* by training various classifiers and second aim to *predict users' level of susceptibility* by using regression models.

Thus, unlike previous research, our work *does not focus on detecting social bots in OSN, but on detecting users who are susceptible to their attacks.* To the best of our knowledge,

Published as part of the #MSM2012 Workshop proceedings,
available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

this represents a novel task that has not been proposed or tackled previously. Our work is relevant for researchers interested in social engineering, trust and reputation in the context of OSN.

2. RELATED WORK

Social bots represent a rather new phenomenon that has received only little attention so far. For example, Chu et al. [3] use machine learning to identify three types of Twitter user accounts: users, bots and cyborgs (users assisted by bots). They show that features such as entropy of posts over time, external URL ratio and Twitter devices (usage of external Twitter applications) give good indications for differentiating between distinct types of user accounts [1]. Work by [6] describes how honeypots can be used to identify spam profiles in OSN. They present a long term study where 60 honeypots were able to harvest about 36.000 candidate content polluters over a period of 7 months. Based on the collected data they trained a classification model using features based on User Demographics, User Friendship Networks, User Content and User History. Their results and show that features which were most useful for differentiating between content polluters and legitimate users were User Friendship Network based features, like the standard deviation of followees and followers, the change rate of the number of followees and the number of followees. In the context of the goals of this paper, related work on spam detection in OSN is as well relevant. For example, Wang et al. [14] propose a general purpose framework for spam detection across multiple social networks. Unlike previous research, our work does not focus on detecting spammers or social bots in OSN, but on detecting users who are susceptible to their attacks.

Research about users' online behavior in general represents another field that is closely related to our research on user susceptibility. Predicting users' interaction behavior (i.e., who replies to whom, who friends whom) in online media has been previously studied in the context of email communications [12] and more recently in the context of social media applications. For example, Cheng et al. [2] consider the problem of reciprocity prediction and study this problem in a communication network extracted from Twitter. The authors aim to predict whether a user A will reply to a message of user B by exploring various features which characterize user pairs and show that features that approximate the relative status of two nodes are good indicators of reciprocity. Work described in [10] considers the task of predicting discussions on Twitter, and found that certain features were associated with increased discussion activity - i.e., the greater the broadcast spectrum of the user, characterized by in-degree and list-degree levels, the greater the discussion activity. The work of Hopcroft et al. [4] explores follow-back-behavior of Twitter users and find strong evidence for the existence of the structural balance among reciprocal relationships. In addition, their findings suggest that different types of users reveal interesting differences in their follow-back behavior: the likelihood of two elite users creating a reciprocal relationships is nearly 8 times higher than the likelihood of two ordinary users. Our work differs from the related work discussed above by focusing on modeling and predicting the behavior of users who are currently attacked by social bots.

3. THE SOCIAL BOT CHALLENGE

The Social Bot Challenge was a competition organized by Tim Hwang (and the WebEcologyProject). The competition took place between January and February 2011. The aim was to have a set of competing teams developing social bots that persuade targets to interact with them - i.e., reply to them, mention them in their tweets, retweet them or follow them. The group of targets consisted of 500 unsuspecting Twitter users which were selected semi-randomly: all users had an interest in or tweeted about *cats*. The majority of targets exhibited a high activity level, that means they tweeted more than once a day. We define a *susceptible user* as a target that interacted (i.e., replied, mentioned, retweeted or followed) at least once with a social bot.

3.1 Rules

Each team was allowed to create one lead bot (the only bot allowed to score points) and an arbitrary number of support bots. The participating teams got points for every successful interaction between their lead bot and any target. One point was awarded for any target who started following a lead bot and three points were awarded for any target who replied to, mentioned or retweeted a lead bot.

The following rules were announced for the game:

- No humans are allowed during the game. That means bots need to act in a completely automated way.
- Teams were not allowed to report other teams as spam or bots to Twitter, but other countermeasures and strategies to harm the opponents are allowed.
- The existence of the game needs to remain a secret. That means bots are not allowed to inform others about the game.
- The code needs to be published as open source under the MIT license.
- Teams are allowed to collaborate. That means they are allowed to talk to each other and exchange their code.

There was a period of 14 days during which teams were allowed to develop their social bots. Afterwards the game started on the Jan 23rd 2011 (day 1) and ended Feb 5th 2011 (day 14). During this period, bots were autonomously active for the first 7 days. At the 30th of January (day 8) the teams were allowed to update their codebase and change strategies. After this optional update, the bots continued to be autonomously active for the remaining time of the challenge

3.2 Participants and Challenge Outcome

The following three teams competed in the challenge.

- **Team A - @sarahbalham** The lead bot *sarahbalham* claims to be a young woman who grew up on the countryside and just moved to the city. This team didn't construct a bot-network,but only used one lead bot. This lead bot created 143 tweets, which is rather low in comparison to the other teams, and used only a few @replies and hashtags. Despite low activity level this team could reach the highest number of mutual connections, which is 119 connections. Overall the team only collected 170 points, since only 17 interactions with targets were counted.

- **Team B - @ninjzz** The woman impersonated by this bot - *ninjzz* - doesn't provide much personal information, only that she is a bit shy and looking for friends on Twitter. Ninjzz was supported by 10 other bots, which also created some tweets. This bot was rather defensive in the first round of the challenge, but changed the strategy on day 8 and acted in a much more aggressive way in the second part of the challenge. Overall this team created 99 mutual connections and 28 interactions, and therefore collected 183 points.

- **Team C - @JamesMTitus** The bot *JamesMTitus* claims to be a 24 old guy from New Zealand, who is new on Twitter, and a real cat enthusiast. Team C with their bot *JamesMTitus* won the game by collecting 701 points, with 107 mutual connections and 198 interactions. This team had five support bots, who only created social connections but did not tweet at all. The team picked a very aggressive strategy, tweeted a lot and also made extensively use of @replies, retweets and hashtags.

4. DATASET

The authors of this paper were not involved in nor did they participate in the design, setup or execution of this challenge. The dataset used for this analysis was provided by the WebEcologyProject after the challenge took place. Table 1 provides a basic description of this dataset. Figure 1 shows infections over time - i.e., it depicts on which day of the challenge targets interacted with social bots for the first time. One can see from this figure that at the beginning of the challenge - on day 2 - already 87 users became infected. One possible explanation for this might be the usage of auto-following features which some of the targets might have used. One can see from Figure 2 that for the users who became infected at an early stage of the challenge, we do not have many tweets in our dataset. This is a limitation of the dataset we use, which includes only tweets authored between the 23th of January and the 5th of February and social relations which where existent at the this point in time or created during this time period. Since most of our features require a certain amount of tweets a user authored in order to contain meaningful information about the user, we decided to remove all users who became susceptible before day 7. While this means we loose 133 susceptible users as samples for our experiments, we believe (i) that the remaining 76 susceptible users and 298 non-susceptible users are sufficient to train and test our classifiers and regression models and (ii) that eliminating those users that might have used an auto-follow feature is a good since they are less interesting to study from a susceptibility viewpoint.

5. FEATURE ENGINEERING

We adopt a two-stage approach to modeling targets' susceptibility to social bot attacks: (i) We aim to identify infected users via a binary classification task, and (ii) we aim to predict the level of susceptibility per infected user. To this end we explore three distinct feature sets that can be leveraged to describe the susceptibility of users: *linguistic features*, *behavioral features* and *network features*.

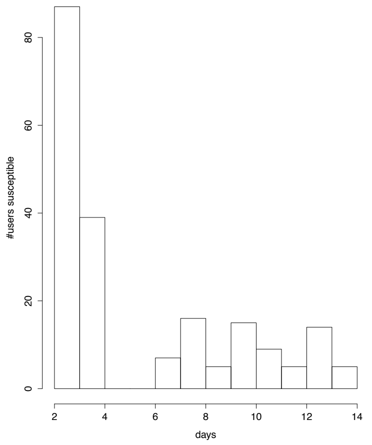


Figure 1: This figure shows for each day of the challenge the number of users who were infected - i.e., they interacted with a social bot for the first time.

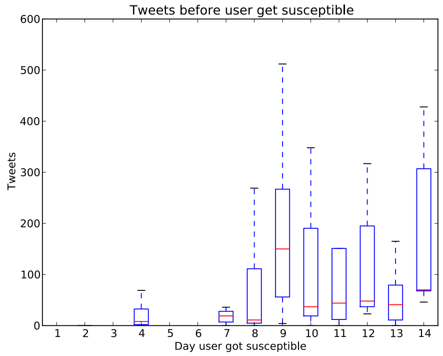


Figure 2: This figure shows when users were infected and how many tweets they have published before - i.e. between the start of the challenge and the day they were infected.

Table 1: Description of the Social Bot Challenge dataset

Susceptible	202
Non-Susceptible	298
Mean Tweets per User	146.49
Mean Nr of Follower/Followees per User	8.5

For all targets, we computed the features by taking all tweets they authored (up to the point in time where they become infected) and a snapshot of the targets' follow network which was as recorded at the 26th of January (day 4). Since we only study susceptible users who became infected on day 7 or later, this follow network snapshot does not contain any future information (such as tweets or social relations which were created after a user became infected) which could bias our prediction results. Based on this aggregation of tweets, we constructed the interaction and retweet network of each user by analyzing their reply and retweet interactions.

5.1 Linguistic Features

Previous research has established that physical and psychological functioning are associated with the content of writing [8]. In order to analyze such content in an objective and quantifiable manner, Pennebaker and colleagues developed a computer based text-analysis program, known as the Linguistic Inquiry and Word Count (short LIWC) [11]. LIWC uses a word count strategy searching for over 2300 words or word stems within any given text. The search words have previously been categorized by independent judges into over 70 linguistic dimensions. These dimensions include standard language categories (e.g., articles, prepositions, pronouns including first person singular, first person plural, etc.), psychological processes (e.g., positive and negative emotion categories, cognitive processes such as use of causation words, self-discrepancies), relativity-related words (e.g., time, verb tense, motion, space), and traditional content dimensions (e.g., sex, death, home, occupation).

In this work we use those 70 linguistic dimensions[1] as linguistic features and compute them based on the aggregation of tweets authored by each target. Due to space limits we do not describe all 70 features in detail, but explain those which seem to be relevant for modeling the susceptibility of users in the result section.

5.2 Network Features

To study the predictiveness of network theoretic features we constructed the following three directed networks from the data. In each of the networks nodes correspond to targets, while edges are constructed differently.

- *User-Follower* - A network representing the target - follower structure in Twitter. There exists an directed edge from user A to user B if the user A is followed by B.
- *Retweet* - A network representing the retweet behavior of targets. In this network there exists an edge from A to B if user A retweeted a message from B.
- *Interaction* - The third network captures the general interaction behavior of targets. There exists an edge from user A to user B if user A either mentioned, replied, or retweeted user B.

For each point in time, we constructed a retweet and interaction network by analyzing all tweets users published before that timestamp. The follower-network is based on a snapshot which was as recorded at the 26th of January (day 4).

5.2.1 Hub and Authority Score

Using Kleinberg's *HITS* algorithm [5], we calculated the authority as well as the hub score for all targets in our networks. A high authority-score indicates that a node (i.e., a user) has many incoming edges from nodes with a high hub score, while a high hub-score indicates that a node has many outgoing edges to nodes with high authority scores. For example, in the retweet network a high authority score indicates that a user is retweeted by many other users who retweeted many users, while a high hub score indicates that the user retweets many others who are as well retweeted by many others.

5.2.2 In- and Out-Degree

A high in-degree indicates that a node (i.e., a user) has many incoming edges, while a high out-degree indicates that a node has many outgoing edges. For example, in the interaction network a high in-degree means that a user is retweeted, replied, mentioned and/or followed by many other users, while a high out-degree indicates that the user retweets, replies, follows and/or mentions many other users.

5.2.3 Clustering Coefficient

The clustering coefficient is defined as the number of actual links between the neighbors of a node divided by the number of possible links between the neighbors of that node. A high clustering coefficient of a node indicates that a node has a central position in the network. For example, in the follow network a high clustering coefficient indicates that the users a user follows or is followed by, are also well connected via follow relations.

5.3 Behavioral Features

In our own previous work [13], we introduced a number of behavioral or structural measures that can be used to characterize user streams and reveal structural differences between them. In the following, we describe some of those measures and elaborate how we use them to gauge the susceptibility of targets.

5.3.1 Conversational Variety

The conversational variety per message $CVpm$ represents the mean number of different users mentioned in one message of a stream and is defined as follows:

$$CVpm = \frac{|U_m|}{|M|} \tag{1}$$

To measure the number of users being mentioned in a stream (e.g., via @replies or slashtags), we introduce $|U_m|$ for $u_m \in U_m$. A high conversational variety indicates that a user talks with many different users.

5.3.2 Conversational Balance

To quantify the conversational balance of a stream, we define an entropy-based measures, which indicates how evenly balanced the communication efforts of a user is distributed across his communication partners. We define the conversational balance of a stream as follows:

$$CB = -\sum_{u \in U_m} P(m|u) * log(P(m|u)) \tag{2}$$

[1] http://www.liwc.net/descriptiontable1.php

A high conversational balance indicates that the user talks equally much with a large set of users, i.e. the distribution of conversational messages per user is even. Therefore a high score indicates that it is hard to predict with whom a user will talk next.

5.3.3 Conversational Coverage

From the number of conversational messages $|M_c|$ - i.e., messages which contain an @reply - and the total number of messages of a stream $|M|$, we can compute the conversational coverage of a user stream, which is defined as follows:

$$CC = \frac{|M_c|}{|M|} \tag{3}$$

A high conversational coverage indicates that a user is using Twitter mainly for a conversational purpose.

5.3.4 Lexical Variety

To measure the vocabulary size of a stream, we introduce $|R_k|$, which captures the number of unique keywords $r_k \in R_k$ in a stream. For normalization purposes, we include the stream size ($|M|$). The lexical variety per message $LVpm$ represents the mean vocabulary size per message and is defined as follows:

$$LVpm = \frac{|R_k|}{|M|} \tag{4}$$

5.3.5 Lexical Balance

The lexical balance LB of a stream can be defined, in the same way as the conversational balance, via an entropy-based measure which quantifies how predictable a keyword is on a certain stream.

5.3.6 Topical Variety

To compute the topical variety of a stream, we can use arbitrary surrogate measures for topics, such as the result of automatic topic detection or manual labeling methods. In the case of Twitter we use the number of unique hashtags $r_h \in R_h$ as surrogate measure for topics. The topical variety per message $TVpm$ represents the mean number of topics per message and is defined as follows:

$$TVpm = \frac{|R_h|}{|M|} \tag{5}$$

5.3.7 Topical Balance

The topical balance TB can, in the same way as the conversational balance, be defined as an entropy-based measure which quantifies how predictable a hashtag is on a certain stream. A high topical balance indicates that a user talks about many different topics to similar extents. That means the user has no topical focus and it is difficult to predict about which topic he/she will talk next.

5.3.8 Informational Variety

In the case of Twitter we define informational messages to contain one or more links. To measure the informational variety of a stream, we can compute the number of unique links in messages of a stream $|R_l|$ for $r_l \in R_l$. The informational variety per message $IVpm$ is defined as follows:

$$IVpm = \frac{|R_l|}{|M|} \tag{6}$$

5.3.9 Informational Balance

The informational balance IB can, in the same way as the conversational balance, be defined as an entropy-based measures which quantifies how predictable a link is on a certain stream. A high informational balance indicates that a user posts many different links as part of her tweeting behavior.

5.3.10 Informational Coverage

From the number of informational messages $|M_i|$ and the total number of messages of a stream $|M|$ we can compute the informational coverage of a stream which is defined as follows:

$$IC = \frac{|M_i|}{|M|} \tag{7}$$

A high informational coverage indicates that a user is using Twitter mainly to spread links.

5.3.11 Temporal Variety

The temporal variety per message $TPVpm$ of a stream is defined via the number of unique timestamps of messages $|TP|$ (where timestamps are defined to be unique on an hourly basis), and the number of messages $|M|$ in a stream. The temporal variety is defined as follows:

$$TPVpm = \frac{|TP|}{|M|} \tag{8}$$

5.3.12 Temporal Balance

The temporal balance TPB can, in the same way as the social balance, be defined as an entropy-based measure which quantifies how balanced messages are distributed across these message-publication-timestamps. A high temporal balance indicates that a user is tweeting regularly.

5.3.13 Question Coverage

From the number of questions $|Q|$ and the total number of messages of a stream $|M|$ per stream we can compute the question coverage of a stream which is defined as follows:

$$QRpm = \frac{|Q|}{|M|} \tag{9}$$

A high question coverage indicates that a user is using Twitter mainly for gathering information and asking questions.

6. EXPERIMENTS

In the following, we attempt to develop models that (i) identify susceptible users (whether a user becomes infected or not) and (ii) predict their level of susceptibility (the extent to which a user interacts with a social bot). We begin by explaining our experimental setup before discussing our findings.

6.1 Experimental Setup

For our experiments, we considered all targets of the Social Bot Challenge, and divided them into those who were not infected (*non-susceptible users*) and those who were infected, i.e. started interacting with a bot within day 7 or later (*susceptible users*). For each of those targets we constructed the features as described in section 5 and normalized them.

Identifying the most susceptible users in a given community is often hindered by including users that are not susceptible at all. We alleviate this problem by first aiming to model the differences between susceptible and non-susceptible users in a binary classification task. Once susceptible users have been identified, we can then attempt to predict the level of susceptibility for each infected user. Therefore we performed the following two experiments.

1. *Predicting Infections* The first experiment sought to identify the factors that are associated with infections. To this end, we performed a binary classification task using 6 different classifier, partial least square regression (pls), generalized boosted regression (gbm), k-nearest neighbor (knn), elastic-net regularized generalized linear models (glmnet), random forest (rf) and regression trees (rpart). We divided our dataset into a balanced training and test set - i.e. in each training and test split we had the same number of susceptible and non-susceptible users. We performed a 10-cross-fold validation and selected the best classifier to further explore the most predictive features, and plotted ROC curves for each feature. The ROC curve is a method to visualize the prediction accuracy of ranking functions showing the number of true positives in the results plotted against the number of results returned. We use the area under the ROC curve (AUC) as the measure of feature importance.

2. *Predicting Levels of Susceptibility* After identifying susceptible users, it is interesting to rank them according to their probability of being susceptible for a bot attack, because one usually wants to identify the most susceptible users, i.e. those who are most in need for security measures and protection. In this experiment we aim to predict the susceptibility level of infected users and identify key features which are correlated with users' susceptibility levels. We define the susceptibility level of an infected user as the number of times a user followed, mentioned, retweeted or replied to a bot.

   We divided our dataset (consisting of infected users only) into a 75/25% split, fit a regression model using the former split and applied it to the latter. We used regression trees to model the susceptibility level of infected users, since they can handle strongly nonlinear relationships with high order interactions and different variable types. The resulting model can be interpreted as a tree structure providing a compact and intuitive representation.

7. RESULTS & EVALUATION

7.1 Predicting Infections

As a first step, we would like to compare the performance of different classifiers for this task and compare them with a random baseline classifier. We used all features and trained six different classifiers: partial least square regression (pls), generalized boosted regression (gbm), k-nearest neighbor (knn), elastic-net regularized generalized linear models (glmnet), random forests (rf) and regression trees (rpart). One can see from table 2 that generalized boosted regression models (gbm) perform best, since they have the highest accuracy.

Table 2: Comparison of classifiers' performance

	Susceptible			Non-Susceptible			
Model	**F1**	**Rec**	**Prec**	**F1**	**Rec**	**Prec**	**Overall**
random	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*	*0.5*
gbm	**0.71**	**0.70**	**0.74**	**0.70**	**0.74**	**0.68**	**0.71**
glmnet	0.69	0.75	0.67	0.73	0.72	0.77	0.71
rpart	0.64	0.56	0.78	0.44	0.60	0.36	0.54
pls	0.67	0.69	0.68	0.68	0.71	0.70	0.68
knn	0.70	0.71	0.71	0.72	0.75	0.71	0.71
rf	0.68	0.72	0.66	0.70	0.70	0.74	0.69

To understand which features are most predictive, we explore the importance of different features by using our best performing model. Table 2 shows the importance ranking of features using the area under the ROC curve as a ranking criterion.

One can see from Table 3 that the most important features for differentiating susceptible and non-susceptible is the out-degree of a user node in the interaction network. Figure 3 shows that susceptible users tend to actively interact (i.e., retweet, mention, follow or reply to a user) with more users than non-susceptible users do on average. That means, susceptible users tend to have a larger social network and/or communication network. One possible explanation for that is that susceptible users tend to be more active and open and therefore easily create new relations with users. Our results also show that susceptible users also tend to have a high in-degree in the interaction network, which indicates that most of their interaction efforts are successful (i.e., they are followed back by users they follow and/or get replies/mentions/retweets from users they reply/mention/retweet).

Further, susceptible users tend to use more verbs (especially present tense verbs, but also past tense verbs and auxiliary verbs) and use more personal pronouns (especially first person singular but also third person singular in their tweets. This suggest that susceptible users tend to use Twitter to report about what they are currently doing.

Interestingly, our results also show that susceptible users have a higher conversational variety and coverage than non-susceptible users, which means that susceptible users tend to talk to many different users on Twitter and that most of their messages have a conversational purpose. This indicates that susceptible users tend to use Twitter mainly for a conversational purpose rather than an informational purpose. Further, susceptible users also have a higher conversational balance which indicates that they do not focus on few conversation partners (i.e., heavily communicate with a small circle of friends) but spend an equal amount of time in communicating with a large variety of users. Its suggests again that susceptible users are more open to communicate with others, also if they are not in their closed circle of friends.

Our results further suggest that susceptible users show more affection - i.e. they use more affection words (e.g., *happy*, *cry*), especially words which expose positive emotions (e.g., *love*, *nice*) - and use more social words (e.g., *mate*, *friend*) than non-susceptible users, which might explain why they are more open to interact with social bots. Susceptible users also tend to use more motion words (e.g., *go*, *car*), adverbs (e.g., *really*, *very*), exclusive words (e.g., *but*, *without*) and negation words (e.g., *no*, *not*, *never*) in their tweets than non-susceptible users. It indicates again that susceptible users tend to use Twitter to talk about their activities and emotionally communicate.

To summarize, our results suggest that susceptible users tend to use Twitter mainly for a conversational purpose (high conversational coverage) and tend to be more open and social since they communicate with many different users (high out-degree and in-degree in the interaction network and high conversational balance and variety), use more social words and show more affection (especially positive emotions) than non-susceptible users.

Table 3: Importance ranking of the top features using the area under the ROC curve (AUC) is used as ranking criterion. The importance value is proportional to the most important feature which has an importance value of 100%.

Feature	Importance
out-degree (interaction network)	100.00
verb	98.01
conversational variety	96.93
conversational coverage	96.65
present	94.66
affect	90.15
personal pronoun	89.71
first person singular	89.27
conversational balance	87.28
motion	87.28
past	86.56
adverb	86.20
pronoun	84.41
negate	84.33
positive emotions	83.25
third person singular	82.38
social	82.02
exclusive	81.86
auxiliary verb	81.70
in-degree (interaction network)	81.66

7.2 Predicting Levels of Susceptibility

To model the susceptibility level of users, we use regression trees and aim to identify features which correlate with users' susceptibility levels. To gain insights into the factors which correlate with high or low susceptibility levels of a user, we inspect the regression tree model which was trained on 75% of our data. One can see from Figure 4 that users who use more negation words (e.g. not, never, no) tend to interact more often with bots, which means they have a higher susceptibility level. Further, users who tweet more regularly (i.e. have a high temporal balance) and users who use more words related with the topic death (e.g. bury, coffin, kill) tend to interact more often with bots than other susceptible users.

One can see from Figure 4 that the structure of the learned tree is very simple which means that our features only allow differentiating between rather lower and rather high susceptibility scores. For a more finer-grained susceptibility level prediction our approach is of limited utility. Also the rank correlation of users given their real susceptibility level and their predicted susceptibility level and the goodness of fit of the model is rather low. One potential reason for that is that our dataset is too small for fitting the model (we only have 76 samples and 97 features). Another potential reason is that our features do not correlate with susceptibility scores of users. We leave the task of elaborating on this problem to future work.

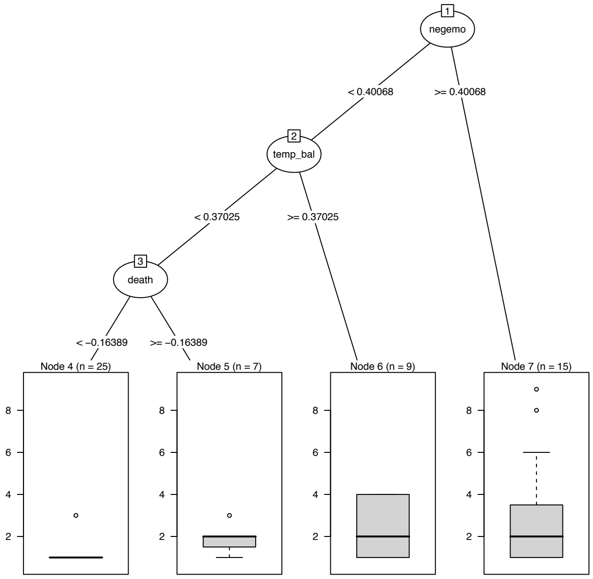


Figure 4: Regression tree model fitted to the susceptibility scores of our training split users. The tree-structure shows based on which features and thresholds the model selects branches and the box plots indicate the distribution of the susceptibility scores of users in each branch of the tree.

8. CONCLUSIONS AND OUTLOOK

In this work, we studied susceptibility of users who are under attack from social bots. To this end, we used data collected by the Social Bots Challenge 2011 organized by the WebEcologyProject. Our analysis aimed at (i) identifying susceptible users and (ii) predicting the level of susceptibility of infected users. We implemented and compared a number of classification approaches that demonstrated the capability of a classifier to outperform a random baseline.

Our analysis revealed that susceptible users tend to use Twitter mainly for a conversational purpose (high conversational coverage) and tend to be more open and social since they communicate with many different users (high out- and in-degree in the interaction network and high conversational balance), use more social words and show more affection (especially positive emotions) than non-susceptible users. Although finding that active users are also more susceptible for social bot attacks does not seem to be too surprising, it is an intriguing finding in itself as one would assume that users who are more active socially would develop some kind of social skills or capabilities to distinguish human users from social bots. This is obviously not the case and suggests that attacks of social bots can be effective even in cases where users have experience with social media and are highly active.

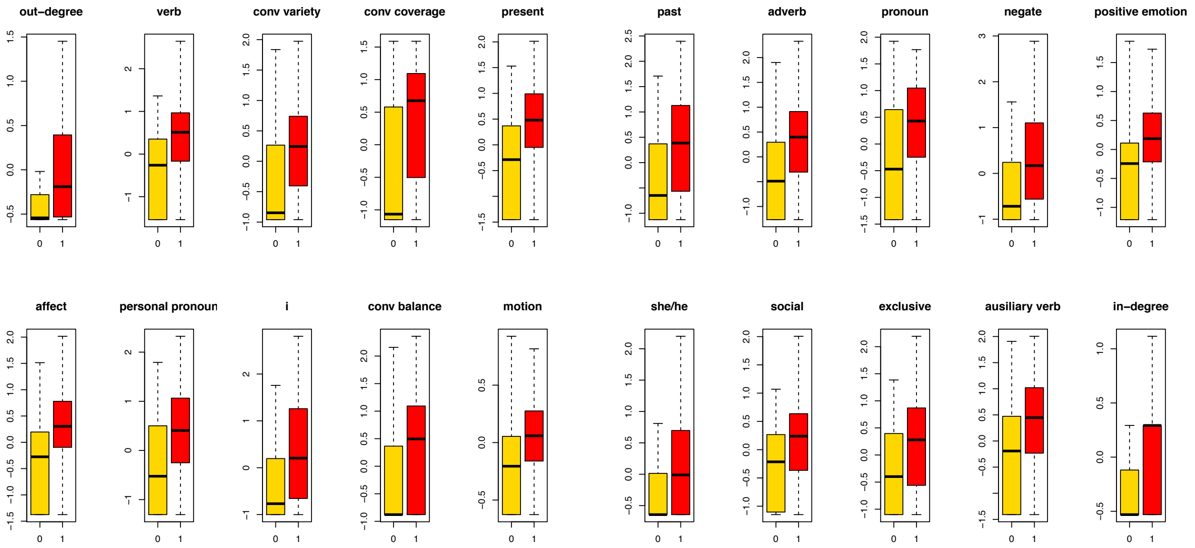


Figure 3: Box plots for the top 20 features according to the area under the ROC curve (AUC). Yellow boxes (class 0, left) represent non-susceptible users, red boxes (class 1, right) represent susceptible users. Differences between susceptible and non-susceptible users can be observed.

While our work presents promising results with regard to the identification of susceptible users, identifying the level of susceptibility is a harder task that warrants more research in the future. In general, the results reported in this work are limited to one specific domain (cats). In addition, all our features are corpus-based and therefore the size and structure of our dataset can have an influence on our results. In conclusion, our work represents a first important step towards modeling susceptibility of users in OSN. We hope that our work contributes to the development of tools that help protect users of OSN from social bot attacks, and that our exploratory work stimulates more research in this direction.

Acknowledgments

We want to thank members of the WebEcology project, especially Tim Hwang for sharing the dataset and Ian Pierce for technical support. Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science. This research is partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760.

9. REFERENCES

[1] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network. In *Proceedings of the 27th Annual Computer Security Applications Conference*, page 93. ACM Press, Dec 2011.

[2] J. Cheng, D. Romero, B. Meeder, and J. Kleinberg. Predicting reciprocity in social networks. In *he Third IEEE International Conference on Social Computing (SocialCom2011)*, 2011.

[3] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter. In *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC10*, page 21. ACM Press, Dec 2010.

[4] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: reciprocal relationship prediction. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1137–1146, New York, NY, USA, 2011. ACM.

[5] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In H. J. Karloff, editor, *SODA*, pages 668–677. ACM/SIAM, 1998.

[6] K. Lee, J. Caverlee, and S. Webb. *Uncovering Social Spammers : Social Honeypots + Machine Learning*, pages 435–442. Number i. ACM, 2010.

[7] D. Misener. Rise of the socialbots: They could be inuencing you online. web, March 2011.

[8] J. Pennebaker, M. Mehl, and K. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[9] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *CoRR*, abs/1011.3768, 2010.

[10] M. Rowe, S. Angeletou, and H. Alani. Predicting discussions on the social semantic web. In *Extended Semantic Web Conference*, Heraklion, Crete, 2011.

[11] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. 2010.

[12] J. R. Tyler and J. C. Tang. When can i expect an email response? a study of rhythms in email usage. In *Proceedings of the eighth conference on European Conference on Computer Supported Cooperative Work*, pages 239–258, Norwell, MA, USA, 2003. Kluwer Academic Publishers.

[13] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.

[14] D. Wang, D. Irani, and C. Pu. A social-spam detection framework. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference on*, pages 46–54. ACM Press, Sep 2011.

What makes a tweet relevant for a topic?

Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben
Web Information Systems, TU Delft
PO Box 5031, 2600 GA Delft, the Netherlands
{k.tao, f.abel, c.hauff, g.j.p.m.houben}@tudelft.nl

ABSTRACT

Users who rely on microblogging search (MS) engines to find relevant microposts for their queries usually follow their interests and rationale when deciding whether a retrieved post is of interest to them or not. While today's MS engines commonly rely on keyword-based retrieval strategies, we investigate if there exist additional micropost characteristics that are more predictive of a post's relevance and interestingness than its keyword-based similarity with the query. In this paper, we experiment with a corpus of Twitter messages and investigate sixteen features along two dimensions: topic-dependent and topic-independent features. Our in-depth analysis compares the importance of the different types of features and reveals that semantic features and therefore an understanding of the semantic meaning of the tweets plays a major role in determining the relevance of a tweet with respect to a query. We evaluate our findings in a relevance classification experiment and show that by combining different features, we can achieve a precision and recall of more than 35% and 45% respectively.

1. INTRODUCTION

Microblogging services such as Twitter[1] or Sina Weibo[2] have become a valuable source of information particularly for exploring, monitoring and discussing news-related information [7]. Searching for relevant information in such services is challenging as the number of posts published per day can exceed several hundred millions[3].

Moreover, users who search for microposts about a certain topic typically perform a keyword search. Teevan et al. [11] found that keyword queries on Twitter are significantly shorter than those issued for Web search: on Twitter people typically use 1.64 words (or 12.0 characters) to search while on the Web they use, on average, 3.08 words (or 18.8 characters). This can be explained by the length of Twitter messages which is limited to 140 characters so that long queries easily become too restrictive. Short queries on the other hand may result in a large (or too large) number of matching microposts.

For these reasons, building search algorithms that are capable of identifying interesting and relevant microposts for a given topic is a non-trivial and crucial research challenge. In order to take a first step towards solving this challenge, in this paper, we present an analysis of the following question: is a keyword-based retrieval strategy sufficient or can we identify features that are more predictive of a tweet's relevance and interestingness? To investigate this question, we took advantage of last year's TREC[4] 2011 Microblog Track[5], where for the first time an openly accessible search & retrieval Twitter data set with about 16 million tweets was published.

In the context of TREC, the ad-hoc search task on Twitter is defined as follows: given a topic (identified by a title) and a point in time *pt*, retrieve all *interesting and relevant* microposts from the corpus that were posted no later than *pt*. A subset of the tweets that were retrieved by the research groups participating in the benchmark were then judged by human assessors as either relevant to the topic or as non-relevant. For example, "Obama birth certificate" is one of the topics that is part of the TREC corpus. Given the temporal context, one can infer that this topic title refers the discussions about Barack Obama's birth certificate: people were questioning whether Barack Obama was truly born in the United States.

We rely on the judged tweets for our analysis and investigate topic-dependent as well as topic-independent features. Examples of topic-dependent features are the retrieval score derived from retrieval strategies that are based on document and corpus statistics as well as the semantic overlap score which determines the extent of overlap between the semantic meaning of a search topic and a tweet. In addition to these topic-dependent features, we also studied a number of topic-independent features: syntactical features (such as the presence of URLs or hashtags in a tweet), semantic features (such as the diversity of the semantic concepts mentioned in a tweet) and social context features (such as the authority of the user who published the tweet).

The main contributions of our work can be summarized as follows:

- We present a set of strategies for the extraction of fea-

[1] http://twitter.com/
[2] http://www.weibo.com/
[3] http://blog.twitter.com/2011/06/200-million-tweets-per-day.html
[4] http://trec.nist.gov/
[5] http://sites.google.com/site/trecmicroblogtrack/

Published as part of the #MSM2012 Workshop proceedings, available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

tures from Twitter messages that allow us to predict the relevance of a post for a given topic.

- Given a set of more than 38,000 tweets that were manually labeled as relevant or not relevant for a set of 49 topics, we analyze the features and characteristics of relevant and interesting tweets.
- We evaluate the effectiveness of the different features for predicting the relevance of tweets for a topic and investigate the impact of the different features on the quality of the relevance classification. We also study to what extent the success of the classification depends on the type of topics (e.g. topics of short-term vs. topics of long-term interest) for which relevant tweets should be identified.

2. RELATED WORK

Since its launch in 2006 Twitter attracted a lot of attention, both in the general public as well as in the research community. Researchers started studying microblogging phenomena to find out what kind of information is discussed on Twitter [7], how trends evolve on Twitter [8], or how one detects influential users on Twitter [12]. Applications have been researched that utilize microblogging data to enrich traditional news media with information from Twitter [6], to detect and manage emergency situations such as earthquakes [10] or to enhance search and ranking of Web sites which possibly have not been indexed yet by Web search engines.

So far, search on Twitter or other microblogging platforms such as Sina Weibo has not been studied extensively. Teevan et al. [11] compared the search behavior on Twitter with traditional Web search behavior. It was found that keyword queries that people issue to retrieve information from Twitter are, on average, significantly shorter than queries submitted to traditional Web search engines (1.64 words vs. 3.08 words). This finding indicates that there is a demand to investigate new algorithms and strategies for retrieving relevant information from microblogging streams.

Bernstein et al. [2] proposed an interface that allows for exploring tweets by means of tag clouds. However, their interface is targeted towards browsing the tweets that have been published by the people whom a user is following and not for searching the entire Twitter corpus. Jadhav et al. [6] developed an engine that enriches the semantics of Twitter messages and allows for issuing SPARQL queries on Twitter streams. In previous work, we followed such a semantic enrichment strategy to provide faceted search capabilities on Twitter [1]. Duan et al. [5] investigated features such as Okapi BM25 relevance scores or Twitter specific features (length of a tweet, presence or absence of a URL or hashtag, etc.) in combination with RankSVM to learn a ranking model for tweets (learning to rank). In an empirical study, they found that the length of a tweet and information about the presence of a URL in a tweet are important features to rank relevant tweets. In this paper, we re-visit some of the features proposed by Duan et al. [5] and introduce novel semantic measures that allow us to estimate whether a micropost is relevant to a given topic or not.

3. FEATURES OF MICROPOSTS

In this section, we provide an overview of the different features that we analyze to estimate the relevance of a Twitter message to a given topic. We present topic-sensitive features that measure the relevance with respect to the topic (keyword-based and semantic-based relevance) and topic-insensitive measures that do not consider the actual topic but solely exploit syntactical or semantic tweet characteristics. Finally, we also consider contextual features that, for example, characterize the creator of a tweet.

3.1 Keyword-based Relevance Features

keyword-based relevance score (Indri-based query relevance): To calculate the retrieval score for pair of (topic, tweet), we employ the language modeling approach to information retrieval [13]. A language model θ_t is derived for each document (tweet). Given a query Q with terms $Q = \{q_1, ..., q_n\}$ the document language models are ranked with respect to the probability $P(\theta_t|Q)$, which according to the Bayes theorem can be expressed as:

$$P(\theta_t|Q) = \frac{P(Q|\theta_t)P(\theta_t)}{P(Q)} \tag{1}$$

$$\propto P(\theta_t) \prod_{q_i \in Q} P(q_i|\theta_t). \tag{2}$$

This is the standard query likelihood based language modeling setup which assumes term independence. Usually, the prior probability of a tweet $P(\theta_t)$ is considered to be uniform, that is, each tweet in the corpus is equally likely. The language models are multinomial probability distributions over the terms occurring in the tweets. Since a maximum likelihood estimate of $P(q_i|\theta_t)$ would result in a zero probability of any tweet that misses one or more of the query terms in Q, the estimate is usually smoothed with a background language model, generated over all tweets in the corpus. We employed Dirichlet smoothing [13]:

$$P(q_i|\theta_t) = \frac{c(q_i,t) + \mu P(q_i|\theta_C)}{|t| + \mu}. \tag{3}$$

Here, μ is the smoothing parameter, $c(q_i, t)$ is the count of term q_i in t and $|t|$ is the length of the tweet. The probability $P(q_i|\theta_C)$ is the maximum likelihood probability of term q_i occurring in the collection language model θ_C (derived by concatenating all tweets in the corpus).

Due to the very small probabilities of $P(Q|\theta_t)$, we utilize $\log(P(Q|\theta_t))$ as feature scores. Note that this score is always negative. The greater the score (that is, the less negative), the more relevant the tweet is to the query.

3.2 Semantic-based Relevance Features

semantic-based relevance score This feature is also a retrieval score calculated according to Section 3.1 though with a different set of queries. Since the average length of search queries submitted to microblog search engines is lower than in traditional Web search, it is necessary to understand the information need behind the query. The search topics provided as part of the TREC data set contain abbreviations, part of names, and nicknames. One example (cf. Table 1) is the first name "Jintao" (in the query: "Jintao visit US") which refers to the President of the People's Republic of China. However, in tweets he is also referred to as "President Hu", "Chinese President", etc. If these semantic variants of a person's name and titles would be considered when deriving an expanded query, a wider variety of potentially relevant tweets could be found. We utilize the well-known Named-Entity-Recognition (NER) service DBPedia

Query	Jintao visits US	
Entity	**Annotated Text**	**Possible Concepts**
Hu Jintao	Jintao	Hu, Jintao, Hu Jintao

Table 1: Example of entity recognition and possible concepts in the query

Spotlight[6] to identify names and their synonyms in the original query. We merge the found concepts into an expanded query which is then used as input to the retrieval approach described earlier.

isSemanticallyRelated It is a boolean value that shows whether there is a semantic overlap between the topic and the tweet. This requires us to employ DBpedia Spotlight on the topic as well as the tweets. If there is an overlap in the identified DBpedia concepts, the value of this feature is *true*, otherwise it is *false*.

3.3 Syntactical Features

Syntactical features describe elements that are mentioned in a Twitter message. We analyze the following properties:

hasHashtag This is a boolean property which indicates whether a given tweet contains at least one hashtag or not. Twitter users typically apply hashtags in order to facilitate the retrieval of the tweet. For example, by using a hashtag people can join a discussion on a topic that is represented via that hashtag. Users, who monitor the hashtag, will retrieve all tweets that contain it. Teevan et al. [11] showed that such monitoring behavior is a common practice on Twitter to retrieve relevant Twitter messages. Therefore, we investigate whether the occurrence of hashtags (possibly without any obvious relevance to the topic) is an indicator for the relevance and interestingness of a tweet.
Hypothesis H1: tweets that contain hashtags are more likely to be relevant than tweets that do not contain hashtags.

hasURL Dong et al. [4] showed that people often exchange URLs via Twitter so that information about trending URLs can be exploited to improve Web search and particularly the ranking of recently discussed URLs. Therefore, the presence of a URL (boolean property) can be an indicator for the relevance of a tweet.
Hypothesis H2: tweets that contain a URL are more likely to be relevant than tweets that do not contain a URL.

isReply On Twitter, users can reply to the tweets of other people. This type of communication can, for example, be used to comment on a certain message, to answer a question or to chat with other people. Chen et al. [3] studied the characteristics of reply chains and discovered that one can distinguish between users who are merely interested in news-related information and users who are also interested in social chatter. For deciding whether a tweet is relevant for a news-related topic, we therefore assume that the boolean *isReply* feature, which indicates whether a tweet is a reply to another tweet, can be a valuable signal.
Hypothesis H3: tweets that are formulated as a reply to another tweet are less likely to be relevant than other tweets.

length The length of a tweet—measured in the number of characters—may also be an indicator for the relevance or interestingness. We hypothesize that the length of a Twitter message correlates with the amount of information that is conveyed in the message.
Hypothesis H4: the longer a tweet, the more likely it is to be relevant and interesting.

The values of boolean properties are set to 0 (false) and 1 (true) while the length of a Twitter message is measured by the number of characters divided by 140 which is the maximum length of a Twitter message.

There are further syntactical features that can be explored such as the mentioning of certain character sequences including emoticons, question marks, exclamation marks, etc. In line with the *isReply* feature, one could also utilize knowledge about the re-tweet history of a tweet, e.g. a boolean property that indicates whether the tweet is a copy from another tweet or a numeric property that counts the number of users who re-tweeted the message. However, in this paper we are merely interested in original messages that have not been re-tweeted yet[7] and therefore also merely in features which do not require any knowledge about the history of a tweet. This allows us to estimate the relevance of a message as soon as it is published.

3.4 Semantic Features

In addition to the semantic relevance scores described in Section 3.2, one can also analyze the semantics of a Twitter message independently from the topic of interest. We therefore utilize again the DBpedia entity extraction provided by DBpedia Spotlight to extract the following features:

#entities The number of DBpedia entities that are mentioned in a Twitter message may give further evidence about the potential relevance and interestingness of a tweet. We assume that the more entities can be extracted from a tweet, the more information it contains and the more valuable it is. For example, in the context of the discussion about birth certificates we find the following two tweets in our dataset:

t_1: *"Despite what her birth certificate says, my lady is actually only 27"*
t_2: *"Hawaii (Democratic) lawmakers want release of Obama's birth certificate"*

When reading the two tweets, without having a particular topic or information need in mind, it seems that t_2 has a higher likelihood to be relevant for some topic for the majority of the Twitter users than t_1 as it conveys more entities that are known to the public and available on Wikipedia and DBpedia respectively. In fact, the entity extractor is able to detect one entity, *db:Birth_certificate*, for tweet t_1 while it detects three additional entities for t_2: *db:Hawaii*, *db:Legislator* and *db:Barack_Obama*.
Hypothesis H5: the more entities a tweet mentions, the more likely it is to be relevant and interesting.

#entities(type) Similarly to counting the number of entities that occur in a Twitter message, we also count the number of entities of specific types. The rationale behind this feature being that some types of entities might be a stronger indicator for relevance than others. The importance of a specific entity type may also depend on the topic.

[6] DBpedia Spotlight, http://spotlight.dbpedia.org/

[7] This is in line with the relevance judgments provided by TREC which did not consider re-tweeted messages.

For example, when searching for Twitter messages that report about wild fires in a specific area, location-related entities may be more interesting than product-related entities. In this paper, we count the number of entity occurrences in a Twitter message for five different types: locations, persons, organizations, artifacts and species (plants and animals).
Hypothesis H6: different types of entities are of different importance for estimating the relevance of a tweet.

diversity The diversity of semantic concepts mentioned in a Twitter message can also be exploited as an indicator for the potential relevance and interestingness of a tweet. We therefore count the number of distinct types of entities that are mentioned in a Twitter message. For example, for the two tweets t_1 and t_2 mentioned earlier, the diversity score would be 1 and 4 respectively as for t_1 only one type of entity is detected (*yago:PersonalDocuments*) while for t_2 also instances of *db:Person* (person), *db:Place* (location) and *owl:Thing* (the role *db:Legislator* is not further classified) are detected.
Hypothesis H7: the greater the diversity of concepts mentioned in a tweet, the more likely it is to be interesting and relevant.

sentiment Naveed et al. [9] showed that tweets which contain negative emoticons are more likely to be re-tweeted than tweets which feature positive emoticons. The sentiment of a tweet may thus impact the perceived relevance of a tweet. Therefore, we classify the the semantic polarity of a tweet into positive, negative or neutral using *Twitter Sentiment*[8].
Hypothesis H8: the likelihood of a tweet's relevance is influenced by its sentiment polarity.

3.5 Contextual Features

In addition to the aforementioned features, which describe characteristics of the Twitter messages, we also investigate features that describe the context in which a tweet was publish. In our analysis, we investigate the social and temporal context:

social context The social context describes the creator of a Twitter message. Different characteristics of the message creator may increase or decrease the likelihood of her tweets being relevant and interesting such as the number of followers or the number of tweets from this user that have been re-tweeted. In this paper, we apply a light-weight measure to characterize the creator of a message: we count the number of tweets which the user has published.
Hypothesis H9: the higher the number of tweets that have been published by the creator of a tweet, the more likely it is that the tweet is relevant.

temporal context The temporal context describes *when* a tweet was published. The creation time can be specified with respect to the time when a user is requesting tweets about a certain topic (query time) or it can be independent of the query time. For example, one could specify at which hour during the day the tweet was published or whether it was created during the weekend. In our analysis, we utilize the temporal distance (in seconds) between the query time and the creation time of the tweet. *Hypothesis H10: the lower the temporal distance between the query time and the creation time of a tweet, the more likely is the tweet relevant to the topic.*

Contextual features may also refer to characteristics of Web pages that are linked from a Twitter message. For example, one could exploit the PageRank scores of the referenced Web sites to estimate the relevance of a tweet or one could categorize the linked Web pages to discover the types of Web sites that usually attract attention on Twitter. We leave the investigation of such additional contextual features for future work.

4. FEATURE ANALYSIS

In this section, we describe and characterize the Twitter corpus with respect to the features that we presented in the previous section.

4.1 Dataset Characteristics

We use the Twitter corpus which was used in the microblog track of TREC 2011[9]. The original corpus consists of approximately 16 million tweets, posted over a period of 2 weeks (January 24 until February 8th, inclusive). We utilized an existing language detection library[10] to identify English tweets and found that 4,766,901 tweets were classified as English. Employing NER on the English tweets resulted in a total over six million named entities among which we found approximately 0.14 million distinct entities. Besides the tweets, 49 topics were given as the targets of retrieval. TREC assessors judged the relevance of 40,855 topic-tweet pairs which we use as ground truth in our experiments. 2,825 tweets were judged as relevant for a given topic while the majority of the tweet-topic pairs (37,349) were marked as non-relevant.

4.2 Feature Characteristics

In Table 2 we list the average values and the standard deviations of the features and the percentages of true instances for boolean features respectively. It shows that relevant and non-relevant tweets show, on average, different characteristics for several features.

As expected, the average keyword-based relevance score of tweets, which are judged as relevant for a given topic, is much higher than the one for non-relevant tweets: -10.709 in comparison to -14.408 (the higher the value the better, see Section 3.1). Similarly, the semantic-based relevance score, which exploits the semantic concepts mentioned in the tweets (see Section 3.2) while calculating the retrieval rankings, shows the same characteristic. The isSemanticallyReleated feature, which is a binary measure of the overlap between the semantic concepts mentioned in the query and the respective tweets, is also higher for relevant tweets than for non-relevant tweets. Hence, when we consider the topic-dependent features (keyword-based and semantic-based), we find first indicators that the hypotheses behind these features hold.

For the syntactical features we observe that, regardless of whether the tweets are relevant to a topic or not, the ratios of tweets that contain hashtags are almost the same (about 19%). Hence, it seems that the presence of a hashtag is not necessarily an indicator for relevance. However, the presence of a URL is potentially a very good indicator: 81.9% of the relevant tweets feature a URL whereas only 54.1% of the non-relevant tweets contain a URL. A possible explana-

[8] http://twittersentiment.appspot.com/

[9] http://trec.nist.gov/data/tweets/

[10] Language detection, http://code.google.com/p/language-detection/

Category	Feature	Relevant	Standard deviation	Non-relevant	Standard deviation
keyword relevance	keyword-based	-10.709	3.5860	-14.408	2.6442
semantic relevance	semantic-based	-10.308	3.7363	-14.264	3.1872
	isSemanticallyRelated	25.3%	43.5%	4.6%	22.6%
syntactical	hasHashtag	19.1%	39.2%	19.3%	39.9%
	hasURL	81.9%	38.5%	54.1%	49.5%
	isReply	3.4%	18.0%	14.2%	34.5%
	length (in characters)	90.323	30.81	87.797	36.17
semantics	#entities	2.367	1.605	1.880	1.777
	#entities(person)	0.276	0.566	0.188	0.491
	#entities(organization)	0.316	0.589	0.181	0.573
	#entities(location)	0.177	0.484	0.116	0.444
	#entities(artifact)	0.188	0.471	0.245	0.609
	#entities(species)	0.005	0.094	0.012	0.070
	diversity	0.795	0.788	0.597	0.802
	sentiment (-1=neg, 1=pos)	-0.025	0.269	0.042	0.395
contextual	social context (#tweets by creator)	12.287	19.069	12.226	20.027
	temporal context (time distance in days)	4.85	4.48	3.98	5.09

Table 2: The comparison of features between relevant tweets and non-relevant tweets

tion for this difference is that the tweets containing URLs tend to feature also an attractive short title, especially for breaking news, in order to attract people to follow the link. Moreover, the actual content of the linked Web site may also stipulate users when assessing the relevance of a tweet. In Hypothesis 3 (see Section 3.3), we speculate that messages which are replies to other tweets are less likely to be relevant than other tweets. The results listed in Table 2 support this hypothesis: only 3.4% of the relevant tweets are replies in contrast to 14.2% of the non-relevant tweets. The length of the tweets that are judged as relevant is, on average, 90.3 characters, which is slightly longer than for the non-relevant ones (87.8 characters).

The comparison of the topic-independent semantic features also reveals some differences between relevant and non-relevant tweets. Overall, relevant tweets contain more entities (2.4) than non-relevant tweets (1.9). Among the five most frequently mentioned types of entities, persons, organizations, and locations occur more often in relevant tweets than in non-relevant ones. On average, messages are therefore considered as more likely to be relevant or interesting for users if they contain information about people, involved organizations, or places. Artifacts (e.g. tangible things, software) and species (e.g. plants, animals) are more frequent in non-relevant tweets. However, counting the number of entities of type species seems to be a less promising feature since the fraction of tweets which mention a species is fairly low.

The diversity of content mentioned in a Twitter message—i.e. the number of distinct types (only person, organization, location, artifact, and species are considered)—is potentially a good feature: the semantic diversity is higher for the relevant tweets (0.8) than for the non-relevant ones (0.6). In addition to the entities that are mentioned in the tweets, we also conducted a sentiment analysis of the tweets (see Section 3.4). Although most of the tweets are neutral (sentiment score = 0), the average sentiment score for relevant tweets is negative (-0.025). This observation is in line with the finding made by Naveed et al. [9] who found that negative tweets are more likely to be re-tweeted.

Finally, we also attempted to determine the relationship between a tweet's likelihood of relevance and its context. With respect to the social context, we however do not observe a significant difference between relevant an non-relevant tweets: users who publish relevant tweets are, on average, not more active than publishers of non-relevant tweets (12.3 vs. 12.2). For the temporal context, the average distance between the time when a user requests tweets about a topic and the creation time of tweets is 4.85 days for relevant tweets and 3.98 for non-relevant tweets. However, the standard deviations of these scores is with 4.53 days (relevant) and 4.39 days (non-relevant) fairly high. This indicates that the temporal context is not a reliable feature for our dataset. Preliminary experiments indeed confirmed the low utility of the temporal feature. However, this observation seems to be strongly influenced by the TREC dataset itself which was collected within a short time period of time (two weeks). In our evaluations, we therefore do not consider the temporal context and leave an analysis of the temporal features for future work.

5. EVALUATION OF FEATURES FOR RELEVANCE PREDICTION

Having analyzed the dataset and the proposed features, we now evaluate the quality of the features for predicting the relevance of tweets for a given topic. We first outline the experimental setup before we present our results and analyze the influence of the different features on the performance for the different types of topics.

5.1 Experimental Setup

We employ logistic regression to classify tweets as relevant or non-relevant to a given topic. Due to the small size of the topic set (49 topics), we use 5-fold cross validation to evaluate the learned classification models. For the final setup, 16 features were used as predictor variables (all features listed in Table 2 except for the temporal context). To conduct our experiments, we rely on the machine learning toolkit Weka[11]. As the number of relevant tweets is considerably smaller than the number of non-relevant tweets, we employed a cost-sensitive classification setup to prevent the classifier from following a best match strategy where simply all tweets are marked as non-relevant. As the estimation for the negative class achieves a precision and recall both over 90%, we focus on the precision and recall of the relevance classification (the positive class) in our evaluation as we aim to investigate the characteristics that make tweets relevant to a given topic.

[11] http://www.cs.waikato.ac.nz/ml/weka/

Features	Precision	Recall	F-Measure
keyword relevance	0.3040	0.2924	0.2981
semantic relevance	0.3053	0.2931	0.2991
topic-sensitive	0.3017	0.3419	0.3206
topic-insensitive	0.1294	0.0170	0.0300
without semantics	0.3363	0.4828	0.3965
all features	0.3674	0.4736	0.4138

Table 3: Performance results of relevance predictions for different sets of features.

Feature Category	Feature	Coefficient
keyword-based	keyword-based	0.1701
semantic-based	semantic-based	0.1046
	isSemanticallyRelated	0.9177
syntactical	hasHashtag	0.0946
	hasURL	1.2431
	isReply	-0.5662
	length	0.0004
semantics	#entities	0.0339
	#entities(person)	-0.0725
	#entities(organization)	-0.0890
	#entities(location)	-0.0927
	#entities(artifact)	-0.3404
	#entities(species)	-0.5914
	diversity	0.2006
	sentiment	-0.5220
contextual	social context	-0.0042

Table 4: The feature coefficients were determined across all topics. The total number of topics is 49. The three features with the highest absolute coefficient are underlined.

5.2 Influence of Features on Relevance Prediction

Table 3 shows the performances of estimating the relevance of tweets based on different sets of features. Learning the classification model solely based on the keyword-based or semantic-based relevance scoring features leads to an F-Measure of 0.2981 and 0.2991 respectively. There is thus no notable difference between the two topic-sensitive features. However, by combining both features (see topic-sensitive in Table 3) the F-Measure increases which is caused by a higher recall, increasing from 0.29 to 0.34. It appears that the keyword-based and semantic-based relevance scores complement each other.

As expected, when solely learning the classification model based on the topic-independent features—i.e. without measuring the relevance to the given topic—the quality of the relevance prediction is poor. The best performance is achieved when all features are combined. A precision of 36.74% means that more than a third of all tweets that our approach classifies as relevant are indeed relevant, while the recall level (47.36%) implies that our approach discovers nearly half of all relevant tweets. Since microblog messages are very short, a significant number of tweets can be read quickly by a user when presented in response to her search request. In such a setting, we believe such a classification accuracy to be sufficient. Overall, the semantic features seem to play an important role as they lead to a performance improvement with respect to the F-Measure from 0.3965 to 0.4138. We will now analyze the impact of the different features in detail.

One of the advantages of the logistic regression model is, that it is easy to determine the most important features of the model by considering the absolute weights assigned to them. For this reason, we have listed the relevant-tweet prediction model coefficients for all employed features in Table 4. The features influencing the model the most are:

- *hasURL*: Since the feature coefficient is positive, the presence of a URL in a tweet is more indicative of relevance than non-relevance. That means, that hypothesis H2 (Section 3.3) holds.
- *isSemanticallyRelated*: The overlap between the identified DBpedia concepts in the topics and the identified DBpedia concepts in the tweets is the second most important feature in this model. This is an interesting observation, especially in comparison to the keyword-based relevance score, which is only the ninth important feature among the evaluated ones. It implies that a standard keyword-based retrieval approach, which performs well for longer documents, is less suitable for microposts.
- *isReply*: This feature, which is *true* (= 1) if a tweet is written in reply to a previously published tweet has a negative coefficient which means that tweets which are replies are less likely to be in the relevant class than tweets which are not replies, confirming hypothesis H3 (Section 3.3).
- *sentiment*: The coefficient of the sentiment feature is similarly negative, which suggests that a negative sentiment is more predictive of relevance than a positive sentiment, in line with our hypothesis H8 (Section 3.4).

We note that the keyword-based similarity, while being positively aligned with relevance, does not belong to the most important features in this model. It is superseded by syntactic as well as semantic-based features. When we consider the non-topical features only, we observe that interestingness (independent of a topic) is related to the potential amount of additional information (i.e. the presence of a URL), the clarity of the tweet overall (a tweet in reply may be only understandable in the context of the contextual tweets) and the different aspects covered in the tweet (as evident in the diversity feature). It should also be pointed out that the negative coefficients assigned to most topic-insensitive entity count features (*#entities(X)*) is in line with the results in Table 2.

5.3 Influence of Topic Characteristics on Relevance Prediction

In all reported experiments so far, we have considered the entire set of topics available to us. In this section, we investigate to what extent certain topic characteristics play a role for relevance prediction and to what extent those differences lead to a change in the logistic regression models.

Consider the following two topics: *Taco Bell filling lawsuit* (MB020[12]) and *Egyptian protesters attack museum* (MB010). While the former has a business theme and is likely to be mostly of interest to American users, the latter topic belongs into the politics category and can be considered as being of global interest, as the entire world was watching the events in Egypt unfold. Due to these differences we defined a number of topic splits. A manual annotator then decided for each split dimension into which category the topic should fall. We investigated four topic splits, three splits with two

[12] The identifiers of the topics correspond to the ones used in the official TREC dataset.

Performance	**Measure**	**popular**	**unpopular**	**global**	**local**	**persistent**	**occasional**
	#topics	24	25	18	31	28	21
	#samples	19803	21052	16209	25646	22604	18251
	precision	0.3596	0.3579	0.3442	0.3726	0.3439	0.4072
	recall	0.4308	0.5344	0.4510	0.4884	0.4311	0.5330
	F-measure	0.3920	0.4287	0.3904	0.4227	0.3826	0.4617
Feature Category	**Feature**	**popular**	**unpopular**	**global**	**local**	**persistent**	**occasional**
keyword-based	keyword-based	0.1018	0.2475	0.1873	0.1624	0.1531	0.1958
semantic-based	semantic-based	0.1061	0.1312	0.1026	0.1028	0.0820	0.1560
	isSemanticallyRelated	1.1026	0.5546	0.9563	0.8617	0.8685	1.0908
syntactical	hasHashtag	0.1111	0.0917	0.1166	0.0843	0.0801	0.1274
	hasURL	1.3509	1.1706	1.2355	1.2676	1.3503	1.0556
	isReply	-0.5603	-0.5958	-0.6466	-0.5162	-0.4443	-0.7643
	length	0.0013	-0.0007	0.0003	0.0004	0.0016	-0.0020
semantics	#entities	0.0572	0.0117	0.0620	0.0208	0.0478	-0.0115
	#entities(person)	-0.2613	0.0552	-0.5400	0.0454	0.1088	-0.3932
	#entities(organization)	-0.0952	-0.1767	-0.2257	-0.0409	-0.1636	-0.0297
	#entities(location)	-0.1446	0.0136	-0.1368	-0.1056	-0.0583	-0.1305
	#entities(artifact)	-0.3442	-0.3725	-0.4834	-0.3086	-0.2260	-0.4835
	#entities(species)	-0.2567	-0.9599	-0.8893	-0.4792	-0.1634	-18.8129
	diversity	0.1940	0.2695	0.2776	0.1943	0.1071	0.3867
	sentiment	-0.7968	-0.1761	-0.6297	-0.4727	-0.3227	-0.7411
contextual	social context	-0.002	-0.0068	-0.0020	-0.0057	-0.0034	-0.0055

Table 5: Influence comparison of different features among different topic partitions. There are three splits shown here: popular vs. unpopular topics, global vs. local topics and persistent vs. occasional topics. While the performance measures are based on 5-fold cross-validation, the derived feature weights for the logistic regression model were determined across all topics of a split. The total number of topics is 49. For each topic split, the three features with the highest absolute coefficient are underlined. The extreme negative coefficient for *#entities(species)* and the occasional topic split is an artifact of the small training size: in none of the relevant tweets did this concept type occur.

partitions each and one split with five partitions:

- Popular/unpopular: The topics were split into popular (interesting to many users) and unpopular (interesting to few users) topics. An example of a popular topic is *2022 FIFA soccer* (MB002) - in total we found 24. In contrast, topic *NIST computer security* (MB005) was classified as unpopular (as one of 25 topics).
- Global/local: In this split, we considered the interest for the topic across the globe. The already mentioned topic MB002 is of global interest, since soccer is a highly popular sport in many countries, whereas topic *Cuomo budget cuts* (MB019) is mostly of local interest to users living or working in New York where Andrew Cuomo is the current governor. We found 18 topics to be of global and 31 topics to be of local interest.
- Persistent/occasional: This split is concerned with the interestingness of the topic over time. Some topics persist for a long time, such as MB002 (the FIFA world cup will be played in 2022), whereas other topics are only of short-term interest, e.g. *Keith Olbermann new job* (MB030). We assigned 28 topics to the persistent and 21 topics to the occasional topic partition.
- Topic themes: The topics were classified as belonging to one of five themes, either business, entertainment, sports, politics or technology. While MB002 is a sports topic, MB019 for instance is considered to be a political topic.

Our discussion of the results focuses on two aspects: (i) the difference between the models derived for each of the two partitions, and, (ii) the difference between these models (denoted $M_{splitName}$) and the model derived over all topics ($M_{allTopics}$) in Table 4. The results for the three binary topic splits are shown in Table 5.

Popularity: A comparison of the most important features of $M_{popular}$ and $M_{unpopular}$ shows few differences with the exception of a single feature: sentiment. While sentiment, and in particular a negative sentiment, is the third most important feature in $M_{popular}$, it is ranked eighth in $M_{unpopular}$. We hypothesize that unpopular topics are also partially unpopular because they do not evoke strong emotions in the users. A similar reasoning can be applied when considering the amount of relevant tweets discovered for both topic splits: while on average 67.3 tweets were found to be relevant for popular topics, only 49.9 tweets were found to be relevant for unpopular topics (the average number of relevant tweets across the entire topic set is 58.44).

Global vs. local: This split did not result in models that are significantly different from each other or from $M_{allTopics}$, indicating that—at least for our currently investigated features—a distinction between global and local topics is not useful.

Temporal persistence: The same conclusion can be drawn about the temporal persistence topic split; for both models the same features are of importance which in turn are similar to $M_{allTopics}$. However, it is interesting to see that the performance (regarding all metrics) is clearly higher for the occasional (short-term) topics in comparison to the persistent (long-term) topics. For topics that have a short lifespan recall and precision are notably higher than for the other types of topics.

Topic Themes: The results of the topic split according to the theme of the topic are shown in Table 6. Three topics did not fit in one of the five categories. Since the topic set is split into five partitions, the size of some partitions is extremely small, making it difficult to reach conclusive results. We can, though, detect trends, such as the fact that relevant tweets for business topics are less likely to contain hashtags (negative coefficient), while the opposite holds for entertainment topics (positive coefficient). The

Performance	**Measure**	**business**	**entertainment**	**sports**	**politics**	**technology**
	#topics	6	12	5	21	2
	#samples	4503	9724	4669	17162	1811
	precision	0.4659	0.3691	0.1918	0.3433	0.5109
	recall	0.7904	0.5791	0.1045	0.4456	0.4653
	F-measure	0.5862	0.4508	0.1353	0.3878	0.4870
Feature Category	**Feature**	**business**	**entertainment**	**sports**	**politics**	**technology**
keyword-based	keyword-based	0.2143	0.2069	0.1021	0.1728	0.2075
semantic-based	semantic-based	0.2287	0.2246	0.0858	0.0456	0.0180
	isSemanticallyRelated	1.3821	0.4088	1.0253	1.0689	2.1150
syntactical	hasHashtag	-0.8488	0.5234	0.3752	-0.0403	-0.1503
	hasURL	2.0960	1.1429	1.2785	1.2085	0.4452
	isReply	-0.2738	-0.4784	-0.6747	-0.9130	-0.3912
	length	0.0044	0.0011	0.0050	-0.0009	0.0013
semantics	#entities	-0.2473	-0.1470	0.0853	0.0537	0.1011
	#entities(person)	-1.2929	-0.1161	-0.4852	0.0177	0.1307
	#entities(organization)	-0.0976	0.0865	-0.4259	-0.0673	-0.7318
	#entities(location)	-1.3932	-0.9327	0.3655	-0.1169	0.0875
	#entities(artifact)	-0.4003	-0.1235	-1.0891	-0.2663	-0.3943
	#entities(species)	0.0241	-19.1819	-31.0063	-0.5570	-0.6187
	diversity	0.5277	0.4540	0.3209	0.2037	0.1431
	sentiment	-1.0070	-0.3477	-1.0766	-0.5663	-0.2180
contextual	social context	-0.0067	-0.0086	-0.0047	-0.0041	-0.0155

Table 6: In line with Table 5, this table shows the influence comparison of different features when partitioning the topic set according to five broad topic themes.

semantic similarity has a large impact on all themes but entertainment. Another interesting observation is that sentiment, and in particular negative sentiment, is a prominent feature in $M_{business}$ and in $M_{politics}$ but less so in the other models.

Finally we note that there are also some features which have no impact at all, independent of the topic split employed: the length of the tweet and the social context of the user posting the message. The observation that certain topic splits lead to models that emphasize certain features also offers a natural way forward: if we are able to determine for each topic in advance to which theme or topic characteristic it belongs to, we can select the model that fits the topic best.

6. CONCLUSIONS

In this paper, we have analyzed features that can be used as indicators of a tweet's relevance and interestingness to a given topic. To achieve this, we investigated features along two dimensions: topic-dependent features and topic-independent features. We evaluated the utility of these features with a machine learning approach that allowed us to gain insights into the importance of the different features for the relevance classification.

Our main discoveries about the factors that lead to relevant tweets are the following: (i) The learned models which take advantage of semantics and topic-sensitive features outperform those which do not take the semantics and topic-sensitive features into account. (ii) The length of tweets and the social context of the user posting the message have little impact on the prediction. (iii) The importance of a feature differs depending on the characteristics of the topics. For example, the sentiment-based feature is more important for popular than for unpopular topics and the semantic similarity does not have a significant impact on entertaining topics.

The work presented here is beneficial for search & retrieval of microblogging data and contributes to the foundations of engineering search engines for microposts. In the future, we plan to investigate the social and the contextual features in depth. Moreover, we would like to investigate to what extent personal interests of the users (possibly aggregated from different Social Web platforms) can be utilized as features for personalized retrieval of microposts.

7. REFERENCES

[1] F. Abel, I. Celik, and P. Siehndel. Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter. In *ISWC '11*, Springer, 2011.

[2] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *UIST '10*, ACM, 2010.

[3] J. Chen, R. Nairn, and E. H. Chi. Speak Little and Well: Recommending Conversations in Online Social Streams. In *CHI '11*, ACM, 2011.

[4] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *WWW '10*, ACM, 2010.

[5] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *COLING '10*, Association for Computational Linguistics, 2010.

[6] A. Jadhav, H. Purohit, P. Kapanipathi, P. Ananthram, A. Ranabahu, V. Nguyen, P. N. Mendes, A. G. Smith, M. Cooney, , and A. Sheth. Twitris 2.0 : Semantically Empowered System for Understanding Perceptions From Social Data. In *Semantic Web Challenge*, 2010.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW '10*, ACM, 2010.

[8] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *SIGMOD '10*, ACM, 2010.

[9] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci '11*, 2011.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10*, ACM, 2010. ACM.

[11] J. Teevan, D. Ramage, and M. R. Morris. #TwitterSearch: a comparison of microblog search and web search. In *WSDM '11*, ACM, 2011.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM '10*, ACM, 2010.

[13] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, ACM, 2001. ACM.

Understanding co-evolution of social and content networks on Twitter

Philipp Singer
Knowledge Management Institute
Graz University of Technology
Graz, Austria
philipp.singer@tugraz.at

Claudia Wagner
DIGITAL Intelligent Information Systems
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Markus Strohmaier
Knowledge Management Institute and Know-Center
Graz University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Social media has become an integral part of today's web and allows users to share content and socialize. Understanding the factors that influence how users evolve over time - for example how their social network and their contents co-evolve - is an issue of both theoretical and practical relevance. This paper sets out to study the temporal co-evolution of content and social networks on Twitter and bi-directional influences between them by using multilevel time series regression models. Our findings suggest that on Twitter social networks have a strong influence on content networks over time, and that social network properties, such as users' number of followers, strongly influence how active and informative users are. While our investigations are limited to one small dataset obtained from Twitter, our analysis opens up a path towards more systematic studies of network co-evolution on platforms such as Twitter or Facebook. Our results are relevant for researchers and social media hosts interested in understanding how content-related and social activities of social media users evolve over time and which factors impact their co-evolution.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks;
J.4 [**Computer Applications**]: Social and behavioral sciences—*Sociology*

General Terms

Experimentation, Human Factors, Measurement

Keywords

Microblog, Twitter, Influence Patterns, Semantic Analysis, Time Series

1. INTRODUCTION

Social media applications such as blogs, message boards or microblogs allow users to share content and socialize. Hosting such social media applications can however be costly, and social media hosts need to ensure that their users remain active and their platform remains popular. Monitoring and analyzing behavior of social media users and their social and content co-evolution over time can provide valuable information on the factors which impact the activity and popularity of such social media applications. Activity and popularity are often measured by the growth of content produced by users and/or the growth of its social network. In recent work we have analyzed how the tagging behavior of users influences the emergence of global tag semantics [3]. However, as a research community we know little about the factors that impact the activity and popularity of social media applications and we know even less about how users' content-related activities (e.g., their tweeting, retweeting or hashtagging behavior) influence their social activities (i.e., their following behavior) and vice versa.

This paper sets out to explore factors that impact the co-evolution of users' content-related and social activities based on a dataset consisting of randomly chosen users taken from Twitter's public timeline by using a multilevel time series regression model. Unlike previous research, we focus on measuring dynamic bi-directional influence between these networks in order to identify which content-related factors impact the evolution of social networks and vice versa. This analysis enables us to tackle questions such as "Does growth of a user's followers increase the number of links or hashtags they use per tweet?" or "Does an increase in users' popularity imply that their tweets will be retweeted more often on average?".

Our results reveal interesting insights into influence patterns in content networks, social networks and between them. Our observations and implications are relevant for researchers interested in social network analysis, text mining and behavioral user studies, as well as for social media hosts who need to understand the factors that influence the evolution of users' content-related and social activities on their platforms.

2. METHODOLOGY

Since we aim to gain insights into the temporal evolution of content networks and social networks, we apply *time se-*

ries modeling [2] based on the work by Wang and Groth [6] who provide a framework to measure the bi-directional influence between social and content network properties. In this work we apply an *autoregressive model* in order to model our time series data. An autoregressive model is a model that goes back p time units in the regression and has the ability to make predictions. This model can be defined as $AR(p)$, where the parameter p determines the order of the model. An autoregressive model aims to estimate an observation as a weighted sum of previous observations, which is the number of the parameter p. In this work we apply a simple model, which calculates each variable independently and further only includes values from the last time unit. The calculated coefficients of the model can determine the influences between variables over time.

In regression analysis variables often stem from different levels. So called *multilevel regression models* are an appropriate way to model such data. Hence, the measurement occasion is the basic unit which is nested under an individual, the cluster unit. In our dataset we have such a hierarchical nested structure. For each day different properties are measured repeatedly, but all of these values belong to different individuals in our study. If we would apply a simple autoregressive model to our data we would ignore the difference between each user and would just calculate the so-called *fixed effects*, because we can not assume that all cluster-specific influences are included as covariates in the analysis [4]. The advantage of such multilevel regression models is now that they add *random effects* to the fixed effects to also consider variations among our individuals. Since we measure different properties repeatedly for different days and different individuals in our study, our dataset has a hierarchical nested structure. Therefore, we utilize a *multilevel autoregressive regression model* which is defined as follows:

$$x_{i,p}^{(t)} = a_i^T x_p^{(t-1)} + \epsilon_i^{(t)} + b_{i,p}^T x_p^{(t-1)} + \epsilon_{i,p}^{(t)} \qquad (1)$$

In this equation $x_p^{(t)} = (x_{i,p}^{(t)}, ..., x_{m,p}^{(t)})^T$ represents a vector, which contains the variables for an individual p at time t. Furthermore, $a_i = (a_{i,1}, ..., a_{im})^T$ represents the fixed effect coefficients and $b_i = (b_{i,1}, ..., b_{im})^T$ represents the random effect coefficients. It is assumed that $\epsilon_i^{(t)}$ and $\epsilon_{i,p}^{(t)}$ is the noise with Gaussian distribution for the fixed and random effects respectively. It has zero mean and variance σ_ϵ^2. To compare the fixed effects to each other, the variables in the random effect regression equations need to be linearly transformed to represent standardized values. How this is done and how the model is finally applied to our data is described in section 4.

3. DATASET

We chose Twitter as a platform for studying the co-evolution of communication content and social networks, since it is a popular micro-blogging service. We explore one *random dataset* in this work, which was crawled within a time period of 30 days. This random dataset consists of random users from the public timeline who do not have anything special in common.

To generate the random dataset, we randomly chose 1500 users from the public Twitter timeline who we used as *seed users*. We used the public timeline method from the Twitter API to sample users rather than using random user IDs since the timeline method is biased towards active Twitter users. To ensure that our random sample of seed users consists of active, English-speaking Twitter users, we further only kept users who mainly tweet in English, have at least 80 followers, 40 followees and 200 tweets. We also had to remove users from our dataset who deleted or protected their account during the 30 days of crawling. Hence, we ended up having 1.188 seed users for whom we were able to crawl their social network (i.e., their followers and followees) and their tweets and retweets. To identify retweets we used the flag provided by the official Twitter API and to extract URLs we used a regular expression. During a 30 day time period (from 15.03.2011 to 14.04.2011) we polled the data daily at about the same time.

4. EXPERIMENTAL SETUP

The goal of our experiments is to study the co-evolution of social and content networks of Twitter users and influence patterns between them. In order to achieve this we firstly created a social and content network for each specific time point.

Social network: The social network is a one-mode directed network, where each vertex represents a user and the edges between these vertices represent the directed follow-relations between two users at a certain point in time. The constructed social network of seed users only reflects a sub-part of a greater network. Therefore it makes no sense to calculate and analyze specific network properties such as betweenness centrality or clustering coefficient, because these properties depend on the whole network and we only have data available for a certain sub-network.

Content network: The content network at each point in time is a two-mode network, which connects users and tweets via authoring-relations. From these user-tweet networks one can extract specific tweet features, such as hashtags, links or retweet information, and build, for example, a user-hashtag network. It would also be possible to create further types of content networks, such as hashtag co-occurrence networks (see [5] for further types), but we leave the investigation of such network types open for future research.

Overall, the social networks capture the social following relations between users, whereas our content networks account the tweets users publish. Finally, we can connect both networks via their user vertexes, since we know which user in the social network corresponds to which user in the content network and vice versa.

A further step towards our final results is the normalization of our available data. This is done by subtracting the time-overall mean and dividing the result by the time-overall standard deviation. The fixed effects can now be analyzed as the effect of one standard deviation of change in the independent variable on the number of standard deviations change in the dependent variable [6].

Based on the prepared data, the final model described in section 2 can be applied to identify potential influences between social and content network properties over time. Ta-

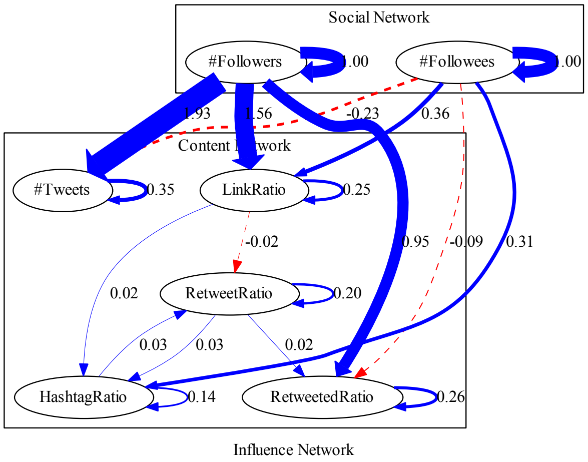


Figure 1: Influence network between the content and social network of a randomly chosen set of Twitter users. An arrow between two properties indicates that the value of one property at time t has a positive or negative effect on the value of the other property at time $t+1$. Red dashed arrows represent negative effects and blue solid arrows represent positive effects. The thickness of the lines indicates the weight of the influence relations. Only statistically significant influences are illustrated.

ble 1 describes each social and content network property used throughout our experiment. The properties are calculated for a corresponding social or content network at each time point t of the random Twitter dataset. The dependent variable of the model is always a property at time t and the independent variable are all properties at time $t-1$ including the dependent variable at that time. Including the dependent variable in that step allows us to detect if a variable's previous value influences it's future value. Finally, the resulting statistical significant coefficients show a relationship between an independent variable at time $t-1$ and a dependent variable at time t. Positive coefficients indicate that a high value of a property leads to an increase of another property, while negative coefficients indicate that a high value of a property leads to an decrease of another property. To reveal positive and negative influence relations between properties within and across different networks, we visualize them as graphical influence network.

Table 1: Social and content network properties

Network type	Property	Description
Social	#Followers	The number of followers a user v has on a specific time point t.
Social	#Followees	The number of followees a user v has on a specific time point t.
Content	#Tweets	The number of tweets a user v has authored on a specific time point t.
Content	Hashtag ratio	The number of hashtags used by a user v on a specific time point t, normalized by the number of daily tweets authored by him/her.
Content	Retweet ratio	The number of retweets (originally authored by other users) by a user v on a specific time point t, normalized by the number of tweets he/she published that day.
Content	Retweeted ratio	The number of tweets produced by a user v on a specific time point t that were retweeted by other users, normalized by the number of tweets user v published that day.

5. RESULTS

Our results reveal interesting influence patterns between social networks and content networks. The influence network in figure 1 shows the correlations detected in the multilevel regression analysis via arrows that point out influences between a property at time t and another property at time $t+1$.

The influence network reveals significant influences of social properties on content network properties. The strongest positive effects can be observed between the number of followers of a user and the content network properties - i.e., users' number of followers positively influences their link ratio, their retweeted ratio and their number of tweets. This indicates that users start providing more tweets and also more links in their tweets if their number of followers increases. Not surprisingly, users' tweets are also more likely to get retweeted if their number of followers increases, because more users are potentially reading their tweets.

Further, figure 1 shows that the number of followees of the social network has positive and negative influences on the content network in our random dataset. While the positive effects point to the link and hashtag ratio, the negative effects point to the number of tweets and the retweeted ratio. This suggests that users who start following other users also start using more hashtags and links. One possible explanation for this is that users get influenced by the links and hashtags used by the users they follow and might therefore use them more often in their own tweets. The negative effect of the number of followees on the number of tweets and the retweeted ratio suggests that users who start following many other users start behaving more like passive readers rather than active content providers.

Another observation of our experiment is that all properties influence themselves positively, which indicates that users who are active one day, tend to be even more active the next day. This indicates for example, that users who attract new followers one day tend to attract more new followers the day after.

6. CONCLUSIONS AND FUTURE WORK

The main contributions of this paper are the following: (i) We applied multilevel time series regression models to one selected Twitter dataset consisting of social and content network data and (ii) we explored influence patterns between social and content networks on Twitter. In our experiments we studied how the properties of social and content networks co-evolve over time. We showed that the adopted approach allows answering interesting questions about how users' behavior on Twitter evolves over time and the factors that impact this evolution. While our results are limited to the dataset used, our work illuminates a path towards studying complex dynamics of network evolution on systems such as Twitter. Our analyses may also facilitate social media hosts to promote certain features of the platform and steer users and their behavior. For example, one can see from our analysis that usage of content features, such as hashtags and links, is highly influenced by social network properties such as the number of followers of a user. Therefore, social media hosts could try to encourage users to use more content features by introducing new measures such as a friend recommender techniques which might impact the social network of users. However, further work is warranted to study these ideas.

Overall, our findings on one small Twitter dataset suggest that there are manifold sources of influence between social and content network properties. Our results indicate that users' behavior and the co-evolution of content and social networks on Twitter is driven by social factors rather than content factors. Previous research by Anagnostopoulos et al. [1] showed that content on Flickr is not strongly influenced by social factors. This may suggest that different social media applications may be driven by different factors. The experimental setup used in our work can be applied to different datasets to study these questions in the future. Nevertheless, further work is required to confirm or refute this observations on other, larger datasets.

Our experiments suggest that the number of followers powerfully influences properties of the content network. One interpreation for that is that the number of followers is a very important motivation for Twitter users to add more content and use more content features like hashtags, URLs or retweets. However, the number of users a user is following can also have a negative influence on content network properties as one can see from figure 1. Our results suggests that an increase of a user's followees (i.e., the number of users he/she follows) implies that the user starts tweeting less and that his/her tweets get less frequently retweeted.

Further, our findings show that all properties influence themselves positively. This does not mean that the values of all properties always increase over time, but that they tend to increase depending on how much they increased the day before. For example, a Twitter user who started posting more links at day t, is likely to post even more links at day $t + 1$ or a user who gain new followers at day t is likely to gain even more new followers at day $t + 1$.

To summarize, our work highlights the existence of interesting influence relationships between content and social networks on Twitter, and shows that multilevel time series regression analysis can be used to reveal such relationships and to study how they evolve over time. Based on the techniques developed by Wang and Groth [6], our work investigated influence patterns in a new domain, i.e. on microblogging platforms like Twitter. Our results are relevant for researchers interested in social network analysis, text mining and behavioral user studies, as well as for community hosts who need to understand the factors that influence the evolution of their users in terms of their content-related and social behavior.

Acknowledgments

This work is in part funded by the FWF Austrian Science Fund Grant I677 and the Know-Center Graz. Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science.

7. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In Y. Li, B. Liu, and S. Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 7–15. ACM, 2008.

[2] G. Kitagawa. *Introduction to Time Series Modeling (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 2010.

[3] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 521–530, New York, NY, USA, 2010. ACM.

[4] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman and Hall/CRC, 2004.

[5] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.

[6] S. Wang and P. Groth. Measuring the dynamic bi-directional influence between content and social networks. In P. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Pan, I. Horrocks, and B. Glimm, editors, *The Semantic Web Ű ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 814–829. Springer Berlin / Heidelberg, 2010.