Arjumand Younus,
M. Atif Qureshi,
Nasir Touheed

SUBJECTIVITY DETECTION THROUGH SOCIO-LINGUISTIC FEATURES

Abstract

Social media platforms have opened new dimensions within the information retrieval domain leading to a novel concept known as Social Information Retrieval. We argue that the concept of Social Information Retrieval can be extended by augmenting the huge amount of content on the traditional Web with the ever-growing rich Social Web content to increase the information richness of today's search engines. This paper proposes a subjectivity detection framework which can lead towards a proposed emotion-aware search engine interface. Our proposed method differs from previous subjectivity analysis approaches in that it is the first method that takes into account social features of social media platforms for the subjectivity classification task. Through experimental evaluations, we observe the accuracy of the proposed method to be 86.21% which demonstrates a promising outcome for large-scale application of our proposed subjectivity analysis technique.

Keywords: Social Search, Subjectivity, Emotion-Aware Search, Socio-linguistic.

1. INTRODUCTION

Social media platforms have given birth to novel dimensions within the field of Information Retrieval giving birth to the concept of Social Search [4, 6]. The traditional view describes Social Search as the activity of satisfying an information need through posing a question to one's social network. We argue that the concept of Social Search can be extended to exploit the vast amount of user generated content in the ever-growing content-rich Social Web. Augmenting the digital content on the Web with the user-generated content on the Social Web can lead to significant advances in Information Retrieval and can thereby help in the design of rich search engine interfaces such as emotion-aware search engines, real-time search engines, entity-focused search engines etc. In this paper we pursue such a study along the dimension of emotion. We propose a framework for detecting the level of emotional charge on a large-scale microblog in the face of a major

happening. Lastly, we propose to augment current search interfaces with data extracted from the Social Web in an attempt to provide information richness to the users of search engines.

Search engines today attempt to solve people's rich information needs directly on the search results page [2] some notable examples being answering user questions for Time Zone, Weather, Currency etc. Through the use of Social Web data an even greater degree of information richness can be provided. One example of such information richness can be the level of emotional charge associated with a real-world event or the overall opinion about a product released by a particular company. This can lead to emotion-aware search with search engines offering not only simple text information for user queries but also associated indicators displaying widespread public sentiment about the query term – this will be particularly relevant for queries of a real-time nature as users issuing such queries are interested in the level of sentiment associated with their query term. However, emotion and subjectivity detection is in itself a challenging research problem and most of solutions in the literature propose the

use of natural language processing techniques [1, 7, 8, 9, 10]. We differ from all previous approaches in that we extract and utilize socio-linguistic features for the subjectivity detection task. A significant contribution of this paper is a method for large-scale subjectivity detection [9] in the occurrence of a major real-world event. The proposed method differs significantly from the previous methods in that those methods do not take into account the nature of the microblogging platform and its social features. Our method is fundamentally built on the use of the social graph and social features for subjectivity analysis which is particularly meaningful when performing subjectivity analysis for major real-world events. The motivation behind the proposed method comes from the results of a user survey which we performed to collect thoughts of users tweeting heavily about the 2010/2011 uprisings in the Arab world. We also demonstrate the accuracy of our proposed method through experimental evaluations on tweets relating to the recent Tunisian revolution. For the purposes of evaluating our proposed model, we use the content collected from Twitter during the Tunisian uprising. This data was chosen in consideration of the major role played by Twitter during these political uprisings.

The rest of the paper is organized as follows. Section 2 discusses related work in this dimension. Section 3 presents the results of a user survey of those users who tweeted heavily during the recent political uprisings; we gather information about tweeting habits of political activists in order to identify useful socio-linguistic features that can be used for subjectivity detection on Twitter. Section 4 describes our subjectivity analysis methodology in detail along with an explanation of the underlying techniques. Section 5 presents the results of experimental evaluations of our proposed method. Finally section 6 concludes the paper.

2. USER STUDY

We first investigate the tweeting habits of users in the face of a major event in the world through a user survey which was exclusively conducted for twitterers tweeting heavily about the uprisings in Arab world. The details of this survey are presented in Section 3.1. Motivated by the results of the survey, we derive a method for performing subjectivity analysis corresponding to a major real-world event. Our method differs significantly from previous ones in that it does not use any of the previous mechanics of natural language features or dictionary-based ones, and instead takes into account Twitter's social networking features for the subjectivity analysis task.

2.1. TWEETING HABITS

We conducted an online survey as a preliminary user study on how Twitter is used during a major world event. 321 active Twitter users participated in the survey and this helped us in laying the groundwork for the proposed method for subjectivity analysis. The questionnaire was kept very brief and to-the-point so as to gather a large number of responses in order to be useful for our novel subjectivity analysis method. The main questions included in the questionnaire were as follows:

• Do you agree with the statement that Twitter is an effective tool for political activism?

• Have you used Twitter for getting or sharing news about current uprisings in Arab world?

• Even when you shared news item related to the current uprising, did you have a sentiment attached with that sharing?

Other significant questions along with the answer choices are as follows:

• How often did you come across charged sentiments when you read any tweet related to current uprising in Arab world?

One answer could be selected for this question with following answer choices: Almost every other tweet reflected charged sentiments, quite often, rarely.

• In what way did you share news related to these political uprisings through Twitter?

Multiple answers could be selected for this question with following answer choices: My tweets were dominated by news of these uprisings, I used my tweets to express solidarity with the people in these regions and involved in the uprisings, I also discussed on Twitter about these uprisings, I used Twitter as a medium to express my emotion in the face of these uprisings.

299 respondents (93%) agreed on Twitter being an effective tool for political activism. 282 (88%) of the respondents agreed that they used Twitter for getting or sharing news about uprisings in Arab world. To the very significant question of whether or not there was an associated sentiment with sharing of news related to the political uprising, 273 (85%) of the respondents responded with a yes. Only 16 (5%) of the respondents answered that they rarely encountered a tweet with highly charged sentiment with respect to the political uprisings. 260 (81%) of the respondents who tweeted about the political uprisings responded that they also conversed about these political happenings on Twitter and used their tweets to express solidarity with people in those regions.

The frequent encounter of tweets with highly charged sentiments in face of major events shows that the traditional sentiment analysis methods might not scale well given the large-scale sentiment found under such contexts. A twitterer associating a sentiment even with a shared news item reveals an interesting pattern, and this, combined with the engagement in conversations about political events is the fundamental feature upon which we build our method as described in the next Section.

3. SOCIAL FEATURES FOR SUBJECTIVITY DETECTION

In this paper, we consider the specific problem of subjectivity analysis. Subjectivity analysis refers to the task of classifying a piece of text as containing or not containing a sentiment. The pieces of text containing sentiment are referred to as subjective, while those that do not contain any sentiment as referred to as objective. The problem of identifying sentiment can be considered as a binary classification task comprising two major steps: 1) selection of features for description of tweets and 2) applying a classifier using the selected features. As opposed to traditional subjectivity analysis methods applied on Twitter data [lrec paper], we used Twitter's social features for the task. This ensures a high level of accuracy along with a promising solution that can scale well for future large-scale social media applications. To the best of our knowledge, this is the first work that proposes taking into account social network features for the sentiment classification task.

3.1. FEATURES

3.1.1. SOCIAL FEATURES

Overall five features were taken into consideration. Four of these features are based on Twitter's social features and motivated through the user survey results presented in Section 2.1. Table 1 lists these three features along with a description of each.

The first of these features *graph_conv_ratio* firstly takes into account the following and follower network of a user on Twitter. It is a general observation that the growing popularity of Twitter as a news media has caused many news and media outlets to migrate to Twitter and they maintain their channels' Twitter account for news dissemination and propagation. One obvious point to note in the social network of these news and media outlets is that they have a large number of followers but a negligible number of followings, and tweets shared by news media are generally of an objective nature. In light of these observations it makes sense to utilize the following to follower ratio for subjectivity analysis of tweets. However, the following is to follower ratio alone may not serve as a good feature for the subjectivity analysis task since some popular political figures from the developing world also have a large number of fol-

Table 1. Twitter's Social Network Features for Sentiment Classification Task

Feature name	Description
graph_conv_ratio	The following is to follower ratio of Twitterer multiplied by his number of conversations
num_convs	Number of conversations in last 30 tweets of Twitterer
num_lists	Number of lists

lowers while they follow less people, but their tweets on the other hand contain highly charged sentiments. To overcome the limitation produced by the following is to follower ratio, we use a special social feature of Twitter which is the ability to converse with other members of the Twittersphere using the @ symbol. The popular figures who express charged sentiments on Twitter also engage actively in conversations related to political events which is why we combine number of conversations in a user's last 30 tweets with the following is to follower ratio to use it as the feature *graph_conv_ratio*.

Apart from utilization of number of conversations in combination with the following is to follower ratio, we also use it as a separate feature for our subjectivity classification task. The last social feature taken into account is the number of lists a user has been added to. The Twitter list feature reveals additional information about a user [5] and is a useful feature for the subjectivity classification task. However, a potential drawback of directly using the number of lists of a user is that normally the Twitter accounts of news channels are added to a large number of lists. To overcome this drawback, we consider the number of lists of a user only if the number of conversations in his last 30 tweets exceeds a threshold of five.

3.1.2. Linguistic Feature

The linguistic feature we use takes into account textual similarity between last 30 tweets of a user. The main hypothesis behind this feature is that users tweeting with high level of sentiment tend to have lexically similarity in their tweets, implying that they tend to tweet about same topic. We calculate this text similarity using the well-known cosine measure defined on the document vectors in the tf-idf term space. Here, each tweet from the last 30 tweets of a user is treated as a document. The larger the value of cosine similarity, the larger is the chance of user's tweets being subjective.

3.1.3. Socio-linguistic Feature

The presence of a hashtag in a tweet is used as a socio-linguistic feature. It is derived from the classical linguistic concept of topic-focus articulation [3] where a written piece of text has a theme also known as «topic». A significant characteristic of social media platforms such as Twitter is the short length of the text that can be posted – and on account of this feature users tend to tweet about a topic in successive tweets denoting that they are talking about same topic with the special Twitter feature known as hashtag. Hashtag implies beginning a word in a tweet with the symbol «#» in order to denote the topic or keyword of their tweet. We use this feature to extract the topic-focus articulation of a tweet with the presence of a hashtag denoting the higher chance of charged sentiments.

4. EXPERIMENTAL EVALUATIONS

This Section presents details of experimental evaluations of the proposed subjectivity analysis techniques. We tested our proposed method using Naïve Bayes classifier, a simple statistical classifier which works well with small training data. Further, the Naïve Bayes classifier works on strong independence assumptions and this property is well-suited to our task.

Using the dataset collected of hourly tweets using Twitter Search API corresponding to query term «Tunisia» we performed an experiment to study the accuracy of the proposed method utilizing all the features mentioned in Section 4.2. For the dataset gathered during the period of 17th January to 29th January, we randomly select 500 tweets for each day labeling all of them with the help of a manual annotator. The labels assigned are «subjective» or «objective» based on the sentiment expressed in a particular tweet. Of these 500 labeled tweets, 150 of them are used as the training data and the remaining 350 were used as the test data. For each collected tweet, we also collect the username of the Twitter user who tweeted it. Through the username, we gather the features described in Section 3.1 for each user. The 150 tweets serving as the training data train our Naïve Bayes classifier, while for the remaining 350 tweets we extract the class using the Naïve Bayes classifier. The true labels assigned by manual annotator are then compared against those obtained with the classifier, we refer to those that do not match as mismatched labels. Accuracy is calculated as follows

$$Accuracy = \frac{Number\ of\ Mismatched\ Labels}{350} *100.$$

The accuracy is calculated for all the days and the experimental results are shown in fig. 1. As shown in fig. 1, the accuracy does not show much variation for the time period of our study. The average accuracy we achieve is 86.21% which is reasonably good.

5. CONCLUSIONS

We have described a subjectivity classification framework for Twitter based on tweeting habits of users. The proposed method was largely motivated by a user survey of Twitter-ers tweeting heavily during a major world event. Our techniques are particularly well-suited within the context of major political events in the world given the large-scale political activism on Twitter. Our experimental evaluations demonstrated that our approach yielded a high level of accuracy and thus, use of the social networking features for sentiment analysis tasks seems to be a promising solution for future research in this dimension. So far research within the sentiment analysis field has looked to natural language processing. We argue that social network analysis presents a feasible solution to the sentiment analysis problem and this is particularly true for large-scale opinions expressed on social media platforms.
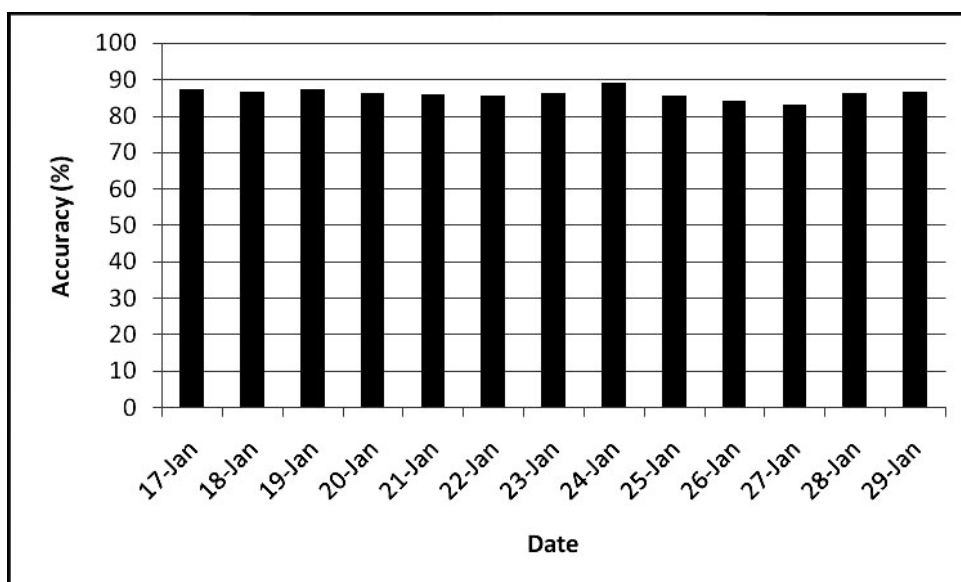

Fig. 1. Subjectivity Classification Accuracy Per Day

References

1. *Bollen J., Pepe A., Mao H.* 2010. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. CoRR, abs/0911.1583, 2009.

2. *Chilton L. B., Teevan J.* Addressing people's information needs directly in a web search result page. In Proc. WWW 2011, ACM Press, 2010. P. 27–36.

3. *Eva H., Barbara H., P., and Petr, S.* Topic-Focus Articulation, Tripartite Structures, and Semantic Content. Studies in Linguistics and Philosophy 71. Dordrecht: Kluwer.

4. *Horowitz D., Kamvar S.D.* The anatomy of a large-scale social search engine // In Proc. WWW 2010, ACM Press, 2010. P. 591–600.

5. *Kim D., Jo Y., Moon I., Oh, A.* Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users. In Proc. Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems (CHI 2010), 2002.

6. *Morris M.R., Teevan J., Panovich K.* A Comparison of Information Seeking Using Search Engines and Social Networks // In Proc. ICWSM 2010.

7. O'Connor, B., Balasubramanyan, R., Routledge, B.R., and Smith, N.A. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series // In Proceedings of the International AAAI Conference on Weblogs and Social Media (Washington DC, USA, May, 2010), 2010.

8. *Pak A., Paroubek P.* Twitter as a Corpus for Sentiment Analysis and Opinion Mining // In Proceedings LREC 2010.

9. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis // In Foundations and Trends in Information Retrieval, January, 2008. Vol. 2, № 1–2. P. 1–135.

10. *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 (Philadelphia, USA, July, 2002), 2002.

11. *Ramakrishnan G., Jadhav A., Joshi A., Chakrabarti S., Bhattacharyya.* Question Answering via Bayesian Inference on Lexical Relations // In ACL Workshop on Multilingual Summarization and Question Answering, 2003.

ОПРЕДЕЛЕНИЕ СУБЪЕКТИВНОСТИ С ПОМОЩЬЮ СОЦИАЛЬНО-ЛИНГВИСТИЧЕСКИХ СВОЙСТВ

Аннотация

Социальные платформы открыли новые направления в области информационного поиска, объединенные концепцией социального информационного поиска. Мы считаем, что эта концепция может разнообразить информацию, представляемую пользователям современными поисковыми машинами, путем комбинирования огромного количества информации традиционного Интернета с быстро растущим и разнообразным социальным контентом. Данная статья предлагает систему определения субъективности, которая может быть использована в предлагаемой поисковой системе с пользовательским интерфейсом, учитывающим эмоциональную окраску результатов. Описанный метод отличается от ранее предложенных использованием специфических свойств социальных сетей для задачи определения субъективности. При экспериментальной оценке предложенного метода мы получили точность 86.21%, что является многообещающим результатом для полномасштабного применения технологии оценки субъективности в поисковых машинах.

Ключевые слова: социальный поиск, субъективность, эмоционально-окрашенный поиск, социально-лингвистический.

Arjumand Younus,
Faculty of Computer Science
Institute of Business Administration
Karachi, Pakistan,
ayounus@iba.edu.pk,

M. Atif Qureshi,
Faculty of Computer Science
Institute of Business Administration
Karachi, Pakistan,
maqureshi@iba.edu.pk,

Nasir Touheed,
Faculty of Computer Science
Institute of Business Administration
Karachi, Pakistan,
ntouheed@iba.edu.pk.