Proceedings of the Second International Conference on
"Emerging Research in Computing, Information, Communication and Applications"
ERCICA 2014
ISBN: 9789351072621

ELSEVIER

Feature Selection Framework for Data Analytics in Microblogs

P. Geetha[1], R. M. Chandresh[2,*] and R. Srinivasan[2]

[1]*Assistant Professor,* [2]*UG Student*
Department of IST, College of Engineering, Guindy, Anna University, Chennai 600 025, India.
e-mail: chandresh.rm@gmail.com, geethap@annauniv.edu

Abstract. The usage of social networks has been on the rise owing to the ubiquitous availability of the Internet. People use social networks to express their views, to provide their comments and to share their experiences with everyone. This data can be used by various people, from manufacturers of products, to politicians predicting election results thus making it an actively pursued research problem in social networks research. Specifically, the problem of obtaining the opinion of people about a certain entity is of high importance. A myriad of traditional learning based text classification schemes have been proposed over the years to this end. However, in microblogs such as Twitter, a lexicon based classification scheme is more suitable owing to the short length of the text. Our aim in this project is to build a novel opinion search engine that makes use of a sentimental word dictionary for text classification to deduce the intents.

Keywords: Twitter, Opinion mining, Sentiment analysis, Lexicon, Social networks.

1. Introduction

Social networks are becoming an increasingly popular phenomenon all round the world. The global availability of the Internet to the people at their fingertips has contributed to this astonishing growth of social networks over the years. It has further been fueled by the people's desire to share information with other people across the globe (the small world phenomenon). People use social networks to express their views, to provide their comments and to share their experiences, thus making them a valuable source of data for several interested parties such as product manufacturers who make use of the valuable feedbacks about their products to plan the next product, politicians who predict results of elections to name a few.

The advent of microblogs, further popularized the usage of social networks. People now had the ability to post their views and opinions in very short texts, usually about a couple of dozen words thus making it a fast and easy way to share information. This ease of usage was responsible for the exponential growth of sites such as Twitter. The relatively smaller length of texts in microblogs also makes the analysis of the data easier, thus making microblogged data, one of the most analyzed datasets for mining opinions.

Over the past decade, a number of schemes have been proposed to obtain this useful information from social network data. Most of these schemes make use of machine learning based classifiers such as the Naive Bayesian classifier, the Support Vector Machine classifier, Artificial Neural Networks etc. These methods have been compared and analyzed in [6]. However, the unique characteristic of text classification is that another alternative exists for classification which is based on a dictionary of words, with each word tagged with its associated polarity. These dictionaries can be constructed either manually, by hand-tagging each word, or automatically, by making use of a set of seed words and then extending them to form larger and larger dictionaries such as the one discussed in [3].

1.1 *Scope of the project*

Our project can be used as a tool to obtain the opinion about a given query from real-time data acquired from Twitter. Possible applications include but not limited to the measurement of the success of a product after it is released,

*Corresponding author

long-term popularity of the product, popularity of an event or a person at a given point of time, study of the global impact of an event. Our project can also be a useful tool for governments to make informed decisions by taking into consideration what the public opinion is, when passing a bill or making any such decision.

1.2 *Methodology*

The basic methodology of this project is the lexicon based intent mining of Twitter data. This is achieved with the help of a dictionary of sentiment-tagged words. The dictionary is constructed by starting with a set of seed words and then making larger and larger dictionaries by obtaining the synsets of these seed words from Wordnet [10,11].

In addition to determining straightforward sentiments using the dictionary, we also include negation detection using the NegEx algorithm [8]. This has resulted in a better efficiency. Unlike traditional methods such as the one described in [1] where only adjectives are considered to be contributing to the polarity, we incorporate the polarities of adverbs, nouns and verbs, in addition to adjectives, to determine the intent.

To give an idea of the features that contributed most to the opinion of the search term to the user, the most frequently talked about features associated with the given input query are then returned to the user. This will help the user in prioritizing the features in order to make informed decisions concerning the search term.

1.3 *Contributions of the project*

The main contributions of our project are as follows.

1. Construction of a sentiment lexicon using seed words and synsets from Wordnet.
2. Text classification using the obtained dictionary.
3. Important feature extraction and determination of overall sentiment.
4. Negation detection and incorporation of adverbs, verbs and nouns in the determination of polarity.
5. Real-time analysis of data from Twitter, thus ensuring that stale data is not being used.

2. Literature Survey

2.1 *Existing work*

An opinion search engine exclusively for product review sites has been developed by Eirinaki *et al.* in [1]. This paper presents an algorithm which not only analyzes the overall sentiment of a document/review, but also identifies the semantic orientation of specific components of the review that lead to a particular sentiment. In this work, in addition to identifying the overall opinion, the semantic orientation of parts of the review are also identified using the High Adjective Count (HAC) algorithm.

A hybrid classification scheme for text opinion mining has been described by F. H. Khan *et al.* in [2]. They make use of the Enhanced Emoticon Classifier, the Enhanced Polarity Classifier and the SentiWordNet Classifier in combination to classify the data. They have also described an exhaustive preprocessing phase prior to the classification phase and have shown that their preprocessing techniques contribute to a significant increase in the accuracy of their classifier. An evaluation of their work using precision, recall, accuracy and F-measure has also been performed in [2].

The Semantic Orientation Calculator (SO-CAL) has been developed by Taboada *et al.* in [4] with a lexicon based classification scheme. The SO-CAL classifier has been shown to perform robustly across domains. It has also been shown to perform well on previously unencountered data. The various approaches for the construction of semantic dictionaries, both manual and automated have also been discussed. This also includes the one described by Yu *et al.* in [3]. The authors have made use of the Mechanical Turk service from Amazon, the internet crowdsourcing service, to manually tag the words in their semantic dictionary.

The sentiment classifier proposed by Lei Zhang *et al.* in [5] makes use of an entity-level lexicon-based classification. However, due to the low recall that results from this, additional tweets are then classified with a learning-based technique. The novelty of this work is that the classifier is trained to assign polarities based on training data that is obtained from the lexicon-based scheme. The authors have demonstrated a significant increase in accuracy.

2.2 *Limitations of existing work*

The opinion search engine proposed by Eirinaki *et al.* in [1] is limited in that it is applicable exclusively only to product review sites. This makes it difficult to use in similar applications in other domains. For example, it cannot

be used in applications such as predicting poll results using social network data, although the core methodology is essentially the same in both applications. Prior knowledge about the domain of interest is required in order to design the opinion search engine thus making it less robust than desirable. In addition, only adjectives are being considered for intent mining which may lead to potential loss of information. This is because the sentiment information, especially in microblogs, can be present even in other parts of speech such as adverbs and even nouns and verbs. For example, in the sentence '*The loss of their captain made their team the underdogs*', the noun *underdog* represents a negative intent. Similarly, the noun *loss* represents a negative intent. These will not be captured if only adjectives are being considered.

In the system proposed by F. H. Khan *et al.* [2], a combination of lexicon based and learning based techniques have been used. Lei Zhang *et al.* in [5] have shown that for shorter texts such as tweets, lexicon based approaches perform better than learning based approaches. Although this system has the advantage that it can be used across domains since the hybrid classifier contains a lexicon based component which performs well across domains, the use of learning based techniques instills unnecessary overheads which can be avoided. Additionally, microblogging data generally consists of one or multiple negations. This system does not incorporate detection of negations which could lead to different results altogether, thus compromising the accuracy of the system. This is true especially in the domain of politics, where usually multiple negations are being used generally.

In the sentiment orientation calculator (SO-CAL) proposed by Taboada *et al.* [4], the drawbacks of the aforementioned works have been addressed. Their work incorporates a lexicon-based approach to classify text and they make use of negation detection and also incorporate different parts of speech in the classification process. However, the NegEx [8] technique that we use performs better than their negation detection algorithm. Also, the small size of the sentiment dictionary compromises the accuracy of the classifier.

In the sentiment classifier proposed by Lei Zhang *et al.* in [5], a lexicon-based as well as a learning-based scheme is used. The major drawback of this system is that the use of a learning-based classification makes the classification tedious when the classification is performed over previously unencountered datasets (i.e.) the performance suffers when prior knowledge is not available.

2.3 *Objective of the project*

In this project, we build an opinion search engine which takes a search query as an input and outputs the general opinion about the search query based on information collected dynamically from Twitter. Given an input query by the user, the intent corresponding to that query is determined by performing a lexicon-based classification of the collected Twitter data with the help of a dictionary containing a list of words and their associated sentimental polarity. Additional features include the ability of the user to choose the parts of speech that are to be considered in the process and also the ability of the user to specify whether negations are to be detected and corrected accordingly.

The novelty of our project lies in the fact that it employs a lexicon-based approach to classify text thus making it robust across domains. This implies that little to no prior knowledge is required when using this approach to mine intents. We employ the NegEx algorithm as used in [8] to identify negated words and hence invert the sentiment. Our analysis shows that this performs better than the one used by Taboada *et al.* in [4].

3. System Architecture

The following sequence provides an overall view of the events that occur when a search query is provided to the system.

- The search query given by the user is used to obtain relevant tweets, dynamically from Twitter. This is achieved by making use of the Python API for Twitter: Twython. We make use of the REST API to obtain relevant tweets which have been recently posted.
- The collected tweets are then preprocessed to remove any potential noise from the tweets. We make use of a comprehensive preprocessing scheme which removes URLs, usernames, hashtags.
- For the Naive Bayesian classification which acts as a baseline for our comparison, a bag of words approach is used for selecting features. The classified data is then stored separately for future use.
- For the lexicon based classification, the sentiment dictionary is first constructed using a set of tagged seed words. The dictionary is then expanded using synsets obtained from Wordnet. This construction process occurs only once and the dictionary is then stored for future use.

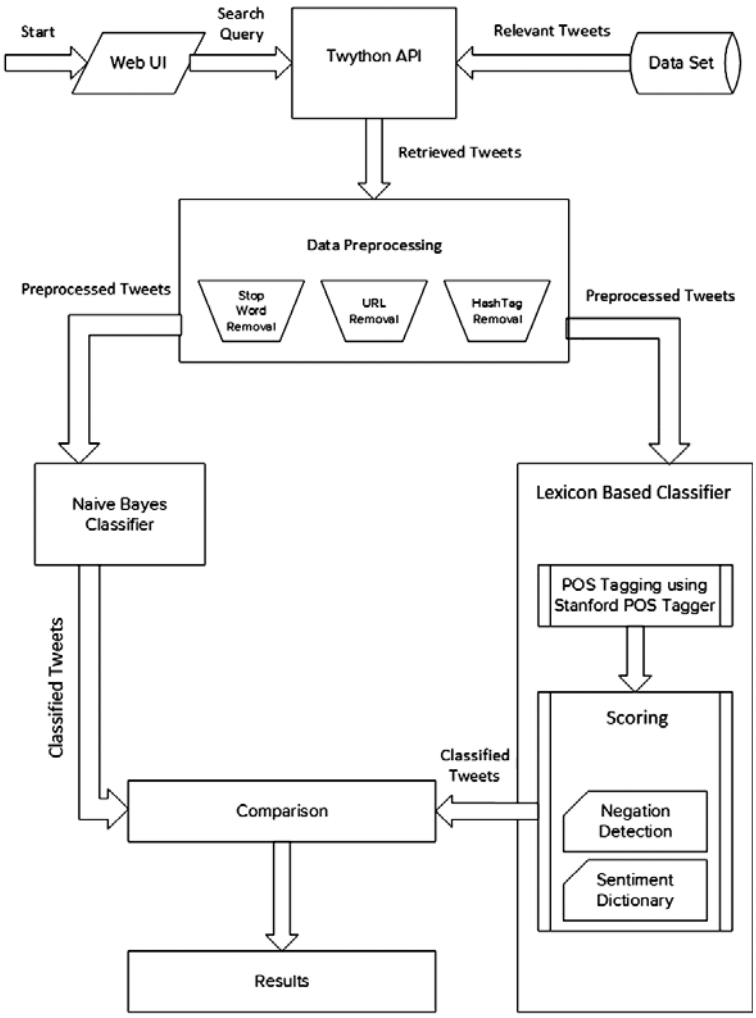


Figure 1. Overall architecture of the system.

- A parts of speech tagging is performed on the tweets using the Stanford POS tagger [9]. The tagged tweets are then fed into the document scorer which then returns the overall intent of the tweet. This is again stored for future use.
- At the same time, the nouns which act as features are collected in the form of a hash map with its associated sentiment in the tweet.
- When all the obtained tweets have been processed, the results are again read from the files and displayed in a Web UI to the user.

Figure 1 shows the architecture of our opinion search engine.

4. Evaluation

We evaluate our system by comparing its performance with that of the classical Naive-Bayesian based classification scheme. The following are the results that are obtained for our system. The precision, recall and accuracy for the proposed work is tabulated below.

4.1 *Parameters for evaluation*

We evaluate our system by comparing the accuracy of classification under different settings of the following parameters.

Table 1. Precision for different window sizes.

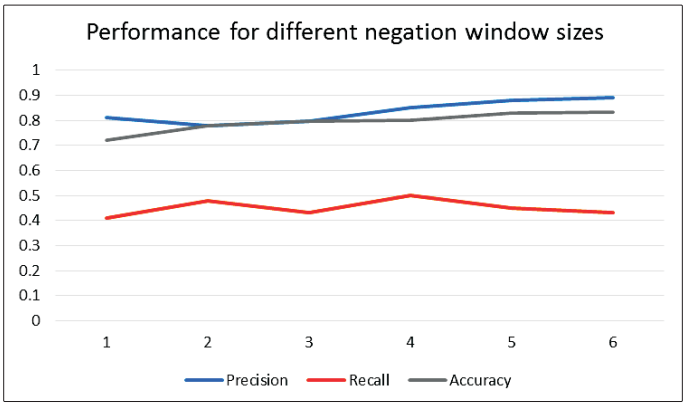
Window Size	1	2	3	4	5	6
Precision	0.81	0.78	0.797	0.85	0.88	0.89

Table 2. Recall for different window sizes.

Window Size	1	2	3	4	5	6
Recall	0.41	0.48	0.43	0.50	0.45	0.43

Table 3. Accuracy for different window sizes.

Window Size	1	2	3	4	5	6
Accuracy	0.72	0.78	0.797	0.80	0.83	0.832



1. Negation Detection and Negation Window Size.
2. Parts-of-Speech that are being used in measuring the polarity.
3. Domain to which the search term belongs to.
4. Possible distance of adjectives or adverbs from nouns.
5. Size of the sentiment dictionary.

4.2 *Results*

The usage of the NegEx algorithm [8] to detect negations increases the efficiency of the system. However, beyond a window size of 5, the efficiency reaches saturation as shown in table 3. The table contains the accuracy values for the case when only adjectives are considered. A similar rise in accuracy is observed when the other parts of speech are also considered with the highest achieved accuracy being 0.85. The precision and recall values have been tabulated in table 1 and table 2.

Traditional lexicon based sentiment analysis has been performed predominantly based on adjectives only. In this project, we provide the user with the option to choose which parts of speech are to be analyzed. Experimental results on a sample query for 100 tweets show that the combination of adjectives and adverbs significantly increases the accuracy of the system. Although the inclusion of nouns and verbs also contribute to the accuracy, they are relatively insignificant.

When the NegEx algorithm [8] is being utilized in conjunction with the parts of speech, the accuracy increases considerably, especially in the domain of politics. The increase in accuracy for various parts of speech is shown in table 6. The precision and recall values are tabulated in table 4 and table 5 respectively.

The domain under consideration plays a key role in the accuracy of our system. The relationships in this case are quite complex. For instance, tweets about politics tend to be involving a lot of negations and hence, the usage of NegEx in this domain yields better results. Whereas, tweets about products tend to consist of strong adjectives and adverbs

Table 4. Precision across various parts of speech.

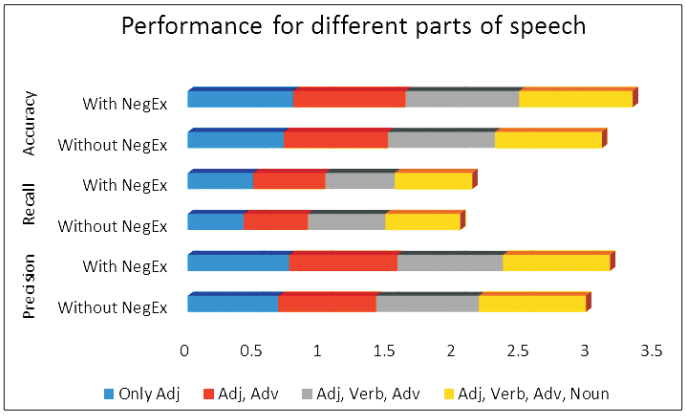
-	Only Adj	Adj, Adv	Adj, Verb, Adv	Adj, Verb, Adv, Noun
Without NegEx	0.68	0.73	0.77	0.80
With NegEx	0.76	0.81	0.79	0.80

Table 5. Recall across various parts of speech.

-	Only Adj	Adj, Adv	Adj, Verb, Adv	Adj, Verb, Adv, Noun
Without NegEx	0.42	0.48	0.58	0.56
With NegEx	0.49	0.54	0.52	0.58

Table 6. Accuracy various parts of speech.

-	Only Adj	Adj, Adv	Adj, Verb, Adv	Adj, Verb, Adv, Noun
Without NegEx	0.72	0.78	0.80	0.80
With NegEx	0.79	0.84	0.85	0.85



and hence yield better results when both adjectives and adverbs are being considered rather than adjectives only. This is summarized in table 7, table 8 and table 9.

The observations are consistent with the expected results, thus the proposed work performs consistently works across domain with much efficiency than the existing work.

Table 7. Precision across different domains.

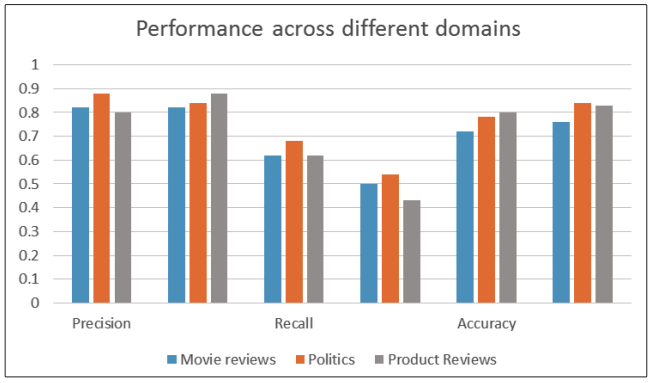
-	Movie reviews	Politics	Product Reviews
Existing Work (Without NegEx)	0.82	0.88	0.80
Proposed Work (With NegEx)	0.82	0.84	0.88

Table 8. Recall across different domains.

-	Movie reviews	Politics	Product Reviews
Existing Work (Without NegEx)	0.62	0.68	0.62
Proposed Work (With NegEx)	0.50	0.54	0.43

Table 9. Accuracy for proposed work across different domains.

-	Movie reviews	Politics	Product Reviews
Existing Work (Without NegEx)	0.72	0.78	0.80
Proposed Work (With NegEx)	0.76	0.84	0.83



5. Conclusion

The above analyses shows that our system performs consistently across domains and is more accurate due to the incorporation of the NegEx algorithm and also due to the inclusion of other parts of speech such as adverbs and verbs. This is also shown to be better for particular domains such as politics.

Our work can be extended by designing a system that automatically suggests the user whether or not to include negation detection and also suggest which parts of speech to analyze based on the given search query. This can also be achieved through the usage of machine learning based techniques to determine which parts of speech to consider when a particular domain is considered

References

[1] Magdalini Eirinaki, Shamita Pisal and Japinder Singh, "Feature Based Opinion Mining and Ranking", Journal of Computer and System Sciences, Vol. 78, pp. 1175–1184, 2012.

[2] Farhan Hassan Khan, Saba Bashir and Usman Qamar, "TOM: Twitter Opinion Mining Framework using Hybrid Classification Scheme", Decision Support Systems, 2013.

[3] Dragut, Eduard C., Clement Yu, Prasad Sistla and Weiyi Meng., "Construction of a Sentimental Word Dictionary", Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1761–1764, 2010.

[4] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu and Bing Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", White Paper, HP Laboratories.

[5] Maite Taboada, Julian Brooke, Milan Toloski, Kimberly Voll and Manfred Stede, "Lexicon-Based Methods for Sentiment Analysis", Computational Linguistics, Vol. 37.2, pp. 267–307, 2011.

[6] Alexander Pak and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Language Resources and Evaluation Conference, 2010.

[7] Pang, Bo and Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval 2.1-2, pp. 1–135, 2008.

[8] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper and Bruce G. Buchanan, "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries", Journal of Biomedical Informatics, 34, pp. 301–310, 2001.

[9] Kristina Toutanova and Christopher D. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger", Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 63–70, 2000.

[10] George A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM, Vol. 38, No. 11, pp. 39–41, 1995.

[11] Christiane Fellbaum, "WordNet: An Electronic Lexical Database", Cambridge, MA: MIT Press 1998.