Twitter Blogs Mining using Supervised Algorithm

Geetanjali S. Potdar
Pune University, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India.

Phursule R.N.
Pune University, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India.

ABSTRACT
Twitter has become one of the most popular micro blogging platforms recently. Near about 800 Millions of users can uses twitter micro-blogging platform to share their thoughts and opinions about different aspects? Therefore, Twitter is considered as a rich source of huge amount of information for decision making, data mining and Sentiment analysis. Sentiment analysis refers to a classification problem where the main focus is to predict the polarity of words and then classify them into positive, negative and neutral feelings with the aim of identifying attitude and opinions that are expressed in any form or language. Sentiment analysis over Twitter offers organizations a fast and effective way to monitor the public's feelings towards their products, brand, business, directors, etc. A wide range of features and methods for training sentiment classifiers for Twitter datasets have been researched in recent years with varying results. The primary issues in previous techniques are data sacristy, classification accuracy, and sarcasm, as they incorrectly classify most of the tweets with a very high percentage of tweets incorrectly classified as neutral. This work focuses on these problems and presents a supervised learning algorithm for twitter feeds classification based on a hybrid approach. The proposed method includes various pre-processing steps before feeding the text to the classifier. Experimental results show that the proposed technique overcomes the previous limitations and achieves higher accuracy, precision and higher recall when compared to similar techniques.

Keywords
Opinion Mining, Sentiment Analysis, hybrid supervised learning Methods, Social Media.

1. INTRODUCTION
Due to the growth of internet is 100% per year many user access and share data on Internet every day. Twitter, with nearly 800 million users and over 350 million messages per day, has quickly become a gold mine for organizations to monitor their reputation and brands by extracting and analyzing the sentiment of the Tweets posted by the public about them, their markets, and competitors. Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Social media network is a graph consisting of nodes and links used to represent social relations on social network sites. In Fig. 1 node represents entities and link represents the link between entities.

Sentiment analysis over Twitter data and other similar micro-blogs faces several new challenges due to the typical short length and irregular structure of such content.

Following are some challenges faced in sentiment analysis of Twitter feeds

- Named Entity Recognition (NER):-This is a method of extracting entities such as people, organization and locations from twitter corpus.
- Anaphora Resolution:- the process of resolving the problem of what a pronoun or noun phrase refers to. We both had a dinner and went for a walk, it was awful. What does It refers to?
- Parsing:- the process of identifying the subject and object of the sentence. The verb and adjective are referring to what?
- Sarcasm:- Sarcasm means what does a verb actually stand for? Does bad mean bad or good?

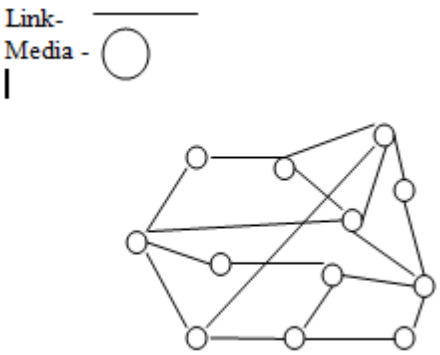


Fig1.1 : Social Media Network Showing Node and Links

1.1 Research Areas in Opinion Mining:

Current research is focusing on

1. Customer Reviews on Individual Product Feature Based Ranking.
2. Overall positive and negative polarity at paragraph level.
3. Ranking of best paragraph or sentence based on best feature and their polarity involved.
4. Continuous Improvement of the algorithms for opinion detection.
5. Decrease the human effort needed to analyze contents.
6. Semantic analysis through lexicon/corpus of words with known sentiment for sentiment classification.
7. Identification of highly rated experts.

1.2 Key challenges and Research Questions of Opinion Mining

1. Product reviews, comments and feedback could be in different languages (English, Urdu, Arabic etc), therefore to tackle each language according to its orientation is a challenging task.

2. As noun words are considered as feature words but Verbs and adjectives can also be used as feature words which are difficult to identify.

3. If a customer-A comments on mobile phone, the voice quality is excellent and customer-B comments, sound quality of phone is very good. Both are talking about same feature but with different wording. To group the synonym words is also a challenging task.

4. Orientation of opinion words could be different according to situation. For example camera size of mobile phone is small. Here adjective small used in positive sense but if customer parallel said that the battery time is also small.

Here small represent negative orientation to battery of phone. To identify the polarity of same adjective words in different situation is also a challenging task.

5. As the customer comment in free format, he can use abbreviation, short words, and roman language in reviews.

For example cam for camera, pic for picture, f9 for fine, gud for good etc. To deal with such type of language need a lot of work to mine opinion.

6. Different people have different writing styles, same sentence may contain positive as well as negative opinion, so it is difficult to parse sentence as positive or negative in case of sentence level opinion mining.

7. In Bing Liu approach opinion always classified only in two categories positive and negative but Neutral opinion also expressed sometimes. Liu considers only adjective as opinion words but opinion can also expressed as adverb, adjectives and verb. For example like is a verb but also an opinion word. His approach finds the implicit features because it extracts the sentences contain at least one feature word. So the features commented by customer indirectly are ignored.

1.3 Different Techniques Used for Opinion Mining

Data mining techniques used to extract the knowledge and information are: generalization, classification, clustering, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks, genetic algorithm, decision tree, multi agent systems, CRISP-DM model, churn reduction, Case Based Reasoning and many more

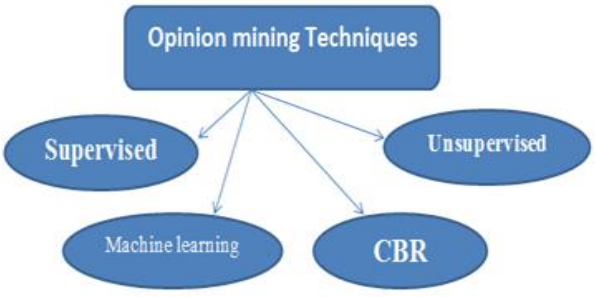


Fig1.2: Techniques of Opinion Mining

2. EXISTING SYSTEM

In this section we discussed about the related work issues. The primary issues in the current system techniques are classification accuracy, data sparsity and sarcasm, as they incorrectly classify most of the tweets with a very high percentage of tweets incorrectly classified as neutral. This project focuses on these problems and presents an algorithm for twitter feeds classification based on the supervised learning algorithm.

3. PROPOSED SYSTEM

3.1 System Overview

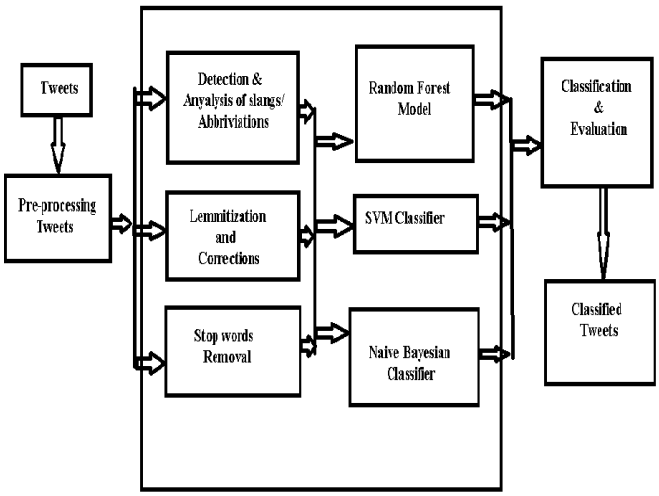


Fig 3.1: Framework of the Proposed System

3.1.1. Data Acquisition

The fundamental purpose of data acquisition module is to obtain the twitter feeds with sparse features in continuous fashion. The Twitter streaming API allows real time access to publicly available data on OSN. Twitter4J library has been used for this purpose. The library was con-figured to extract only English language tweets. The tweets serve as input to pre-processing module and then they are further classified as positive, negative or neutral.

3.1.2 Pre-processing:

The pre-processing module involves performing intensive processing steps at each tweet individually and then passes each refined tweet to the classifier. This consists of following steps:

- Look up for meaning of each word in three English dictionaries (Word Net/Spell Check /J Spell). The words that are not found illustrate that they are either slangs or abbreviations. For example, the tweet ?@xyz u and Jane are gud friends?. ?u? and ?gud? will not return any meaning.
- Abbreviations and/or shorthand notations will be replaced by expansions. Net lingo and sms dictionary are used for this purpose. Our example tweet will now be represented as,? @xyz you and Jane are good friends?
- The next step is to apply lemmatization. Lemmatization is used to stem the words and apply corrections. For example, when ?happiness? is stemmed to ?happi?.
- Apply spell checking of the tweet in order to correct the effects of the lemmatized. This step feeds the remaining words in the spell checker and substitute with the best match. We have used Jazzy Spell Checker, JSpell and Snow ball for spell checking. For instance, ?happi? is corrected to ?happy?.

3.1.3 Supervised learning

Supervised learning is the machine learning task of inferring a function from supervised train-ing data. Supervised learning algorithm performs the following steps:

1. Determine the type of training examples.
For example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.

2. Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.

3. Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are de-scriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.

4. Determine the structure of the learned function For example, the engineer may choose to use support vector machines or decision trees.

Complete the design. Run the learning algorithm on the gathered training set.

5. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.

6. Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

A. RANDOM FOREST MODEL:

Compositional semantic rule helps in learning the meaning of contextual information for random forest. It has a rule to identify the meaning of the sentences. The compose function provides the polarity of the compound expression. For example, this book is not informative, the word informative specifies the positive sentiment but the previous word not alters the sentiment of the sentence. This is addressed by polarity (not (arg1)) = polarity (arg1). This work is based on an algorithm in the sentiment elicitation system proposed by Zhang et al [13]. The random forest model is trained based on the rules given in the Table1 to classify

Compose(arg1,arg2) =	if arg1 is negative: if arg2 is not neutral : return: polarity (arg2) else: return -1 else if arg1 is positive and arg2 is not neutral: return polarity(arg2) else if polarity(arg1) equals polarity (arg2): return 2 polarity(arg1) else if (arg1 is positive and arg2 is neutral) or (arg2 is positive and arg1 is neutral): return polarity(arg1) + polarity (arg2) else: return 0

Tabel 1: Compose Function

B. SVM CLASSIFIER:

In general, the work in the context of supervised sentiment analysis mainly focused on lexeme based features for sentiment classification. WordNet [14] is a large lexical database which provides different senses for a single word. Replacing the word by its sense will improve the accuracy of a sentiment classifier. The WordNet senses are better features compared to word. Every word is replaced by its corresponding synset ID. The first digit in ID refers to parts-of-speech and the remaining digits refer to its meaning. Thus, the SVM classifier [16] is trained based on senses as features.

C. NAIVE BAYESIAN CLASSIFIER:

The semantics concepts as feature for supervised sentiment classifier can provide better classification [14]. The entity extractor like Alchemy API, Zemanta can be used to extract entity and concepts. The concepts are inserted as additional features in the training data. The multinomial nave Bayesian classifier performs the classification. Nave Bayesian classifier is a simple probabilistic classifier. The semantic concepts are included into the training set by interpolation method. The language model with the Interpolation component is given by:

$$Pf(W|C) = \propto Pu(W|C) + \sum I\beta iP(W, Fi, C)i \qquad (1)$$

Where Pu(WjC) is the original unigram model calculated via maximum likelihood estimation. P(W,Fi,C) is the interpolation component which can be decomposed into

$$P(W, Fi, C) = \sum jP(W|fij).P(fij|C)j \qquad (2)$$

4. EXPERIMENTAL RESULTS

4.1 Experimental Setup

Program for Twitter Opinion Mining is written in JAVA. This program is run/executed on Net Beans v5.5.1 or later platform. To write this programs requires JAVA v1.7 or later. This JAVA program can be executed on any windows operating system from windows 7 latest OS. While executing this program Internet connection is mandatory.

4.2 Dataset

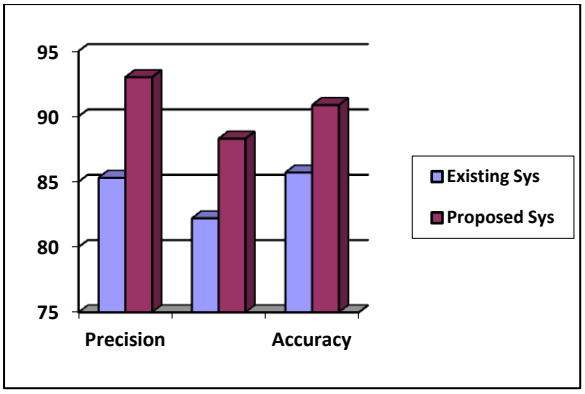
In this system we used the tweeter Blogs Dataset it contains public opinion about particular product or topic.

4.3 Results

The experimental results is perform on Pentium(R) Dual Core processor with installed memory of 2.00 GB RAM. I will train the Semantic Sentiment Miner by the twitter dataset and tested and try to obtain the average accuracy of the system. The performance is evaluated by three measures. They are precision, recall and accuracy. Following table 1 shows the results that are obtained after using the hybrid approach of the supervised learning methods.

Positive Sentimate			**Negative Sentimate**		
Precision	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**
93.61	88.3	90.87	88.4	93.64	90.94

Table 2: Result Table



5. ACKNOWLEDGMENTS

This is a great pleasure and immense satisfaction to express my deepest sense of gratitude and thanks to everyone who has directly or indirectly helped me in completing my Dissertation work successfully. I express my gratitude towards project guide Prof. R. N. Phursule and Prof. S.R.Todmal Head of Department of Computer Engineering. JSPM's Imperial college of Engineering and Research, Wagholi Pune.

6. CONCLUSION

This project has proposed a new algorithm that is Supervised Learning Algorithm for twitter sentiment analysis for increases the accuracy, precision and recall. I have also discussed challenges that are faced during sentiment analysis and proposed the algorithm that resolves these issues and increases the classification accuracy effectively.

In this Project I have used Supervised Learning Algorithm to get Accuracy, Precision & recall at maximum level (more than 90%). Under Supervised Learning Algorithm I have used below three classifiers and maximize the performance. Analysis of Twitter Blogs which is extracted from the Twitter API. Applied Data Acquisition Technique on the analyzed data.

Preprocessed Data is successfully achieved by using preprocessing steps. Sentiment analysis over Twitter offers organizations a fast and effective way to monitor the publics' feelings towards their brand, business, directors, etc.

Sentiment analysis is used to predict the polarity of words and then classify them into positive and negative feelings with the aim of identifying attitude and opinions that are expressed in any form & in English language. In this project sentiment analysis was done by using following classifiers.1. SVM CLASSIFIER 2. NAIVE BAYESIAN CLASSIFIER.

7. REFERENCES

[1] TOM: Twitter Opinion Mining Framework Using Hybrid classification scheme 2013

[2] A. Cui, M. Zhang, Y. Liu, S. Ma,Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis, Springer-Verlag, Berlin, Heidelberg, 2011, Vol. 24, no. 1,January 2012.

[3] A. Bifet, E. Frank,Sentiment Knowledge Discovery in Twitter Streaming Data Published in the Proceedings Springer- Verlag, Berlin, Heidelberg, 2010

[4] A. Bifet, G. Holmes, B. Pfahringer,MOA-TweetReader: realtime analysis in twitter streaming data.

[5] S. Ye, S.F.Wu, Measuring message propagation and social influence on Twitter.com 2013.

[6] S. Argamon, K. Bloom, A. Esuli, F. Sebastiani, Automatically determining attitude type and force for sentiment analysis Berlin Heidelberg, 2009.

[7] X. Fu, Y. Guo, W. Guo, Z. Wang, et al.,Aspect and sentiment extraction based on information-theoretic co-clustering, in: J. Wang, G.G.. ISNN 2012.

[8] A. Nagy, J. Stamberger, Crowd sentiment detection during disasters and crises Proceedings of the 9th International ISCRAM Conference 2012.

[9] A. Montejo-Raez, E. Mart?nez-Camara, M.T. Mart?n-Valdivia, L.A. Urena-Lopez, RandomWalk weighting over SentiWordNet for sentiment polarity detection on Twitter, Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 310, 2012.

[10] R. Ortega, A. Fonseca, M. Mendoza, Y. Guti'errez, SSA-UO: unsupervised Twitter sentiment analysis, in: A. Montoyo (Ed.), Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, pp. 501–507,(Atlanta,Georgia).

[11] F. Bravo-Marquez, M. Mendoza, B. Poblete, Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis, WISDOM'13, Chicago, IL, USA, 2013.

[12] J. Kim, J. Yoo, H. Lim, H. Qiu, Z. Kozareva, A. Galstyan, Sentiment Prediction using Collaborative Filtering, Association for the Advancement of Artificial Intelligence, 2013.