ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Twitter Sentiment Analysis Using Adaboost Classification

Sachin Madhukar Ramteke[1], Sachin N. Deshmukh[2]

M.Tech Student, Dept. of Computer Science and Information Technology, Dr B. A. M. University, Aurangabad

Maharashtra, India

Professor, Dept. of Computer Science and Information Technology, Dr B. A. M. University, Aurangabad

Maharashtra, India

ABSTRACT: In this paper, we study into the benefits of expressive features for recognizing the sentiment of Twitter messages i.e. Tweets. We analyse the effectiveness of existing lexical resources and additionally features that take information about the casual and innovative language used in Twitter. In this paper take a supervised classification approach to the problem, but authority obtaining hashtags into Twitter data for establishment training data.

KEYWORDS: Twitter, Hash Tag, Sentiment Analysis, Features Selection, Adaboost Classification

I. Introduction

Micro blogging websites are just only Social media site to which user compose small and repeated posts. There are nearly 111 micro blogging websites. Twitter is one of the popular micro blogging services. User can be read and write post messages. 140 letter in length these messages are called as Tweets [1]. Twitter is started in 2006, average number of tweets per day 58 million [2].

Twitter is also a huge platform in that different idea, thought, opinion are presented and exchanged. Not important where people came from, what religious opinions they hold, rich or poor, educated or uneducated, they comment, compliment, discuss, argue, insist and oppose over discussion. They are focused in, sharing their own emotions openly [3]. These structure include user introduce, Twitter userID, hash tags, URLs, and media Users. '@' followed by a user identifier report the user, RT stands for retweet, and '#' followed by a word characterizes a hash tag, 'Emoticon' followed by emotion represent by Special symbol [4]. Using the # symbol, users can tag their tweets, indicating the purpose towards a certain topic, e.g. "# MakeInIndia". Those tags can be used by the users to find out other tweets about the same topic. Twitter gives an analysis of so-called "trending topics", which are presently discussed by large amount of users [5].

Sentiment analysis is one of the natural language processing in which we track the mood of the public about a particular entity. It is also called as Opinion Mining which is used for creating a system to collect and examine opinions about the particular entities made in tweets. We evaluated the overall structure of these micro blog postings, the types of statement, and the grouping in positive or negative sentiment [6].

Consider an example of product feedback by the customers. As more and more users post about products and favour they used or express their legislative and religious views, micro blogging websites become beneficial sources of people's opinions and sentiments. Such data can be profitable used for marketing or social studies. [7] By utilizing the sentiment analysis the customer can know the sentiment about the products or services previously making a purchase. The company can also use sentiment analysis to know the opinion of customers about their products, so that they can inspect customer happiness and according to that they can upgrade their products [1].

Consider another example of political party's discussion. Political parties may be curious to know if people support their agenda or not. Social organizations may ask people's opinion on current discussions. All this information can be obtained from micro blogging services, as their users post everyday what they like or dislike and their opinions on many facts of their life [8].

A standard method to implement supervised sentiment analysis is the lexicon based method. It is challenging for standard lexicon-based unsupervised methods to analyse the sentiment due to the fact that expressions in social media are unstructured, informal, and fast-developing. [9] [10] Features such as part-of-speech tags such as sentiment lexicons. That have been proved useful for sentiment analysis other domains they also prove useful for sentiment analysis in Twitter? We begin to analyse this question. We use a dataset formed of collected Emoticon and Hash tag from Twitter [7].

Another problem of micro blogging is the incredible breadth of discussion that is covered. It is not an overstatement to comment that people tweet throughout anything and everything. Therefore, to be able to build systems to mine Twitter sentiments analysis about any given topic, we need a method for rapidly identifying data that can be used for training. In this paper, we explore one method for construct such data using Twitter hash tags (e.g., #MakeInIndia) to decide positive, negative, neutral tweets to use for training three-way sentiment classifiers [11].

II. RELATED WORK

The Author [12]who were the first to do sentiment analysis especially on Twitter data. They treat the problem as one of binary classification. They employ unigrams, bigrams, a combination of both and part-of-speech tags. They compare different classifiers like the Naive Bayes, Maximum Entropy and Support Vector Machine (SVM) for classification. They have provided result having 82.9% accuracy using SVM with only unigrams as features [12].

The Author [13] have used part-of-speech tags to compute the posterior probability in Naive Bayes models. They find SVM and CRF report a best result rather unpopular measure. They use the two measures entropy and salience to identify n-grams and find salience for superior measure. They have Confirm by making the observations that classification performance increases with more training data [13].

The Author [14]used two-step classifier. The first step, tweets are classified as subjective or objective. The subjective tweets are classified as positive or negative. Divide their features into two categories: meta-features and tweet syntax. The first group holds features such as part-of-speech tags from a part-of-speech dictionary and the subjectivity and polarity of words in the MPQA lexicon negation word and weighted by the occurrence of positive and negative words in the training data. The second group holds Twitter-specific features such as retweets, hash tags, emoticons etc. They get the best results using a SVM classifier for both steps and provide 81.9% accuracy for the subjectivity detection step, 81.3% accuracy for polarity detection, and report a unigram baseline of 72.4% and 79.1%, respectively. They find that the meta-features are very important for the polarity detection and the tweet syntax features are more important for subjectivity detection [14].

The Author [15] calculated the impact of the shortness of tweets on sentiment analysis. They collect tweets with five categories “entertainment, products and services, sport, current affairs and companies”. For their machine learning classifiers they have used unigrams, bigrams and trigrams as well as part-of- speech n-grams. They find the Naïve Bayes classifier to outperform SVM. They report their best result for binary positive/negative classification having 74.85% accuracy and 61.3% for the ternary case, both using Naive Bayes and unigrams [15].

The Author [16] have provided the challenges of the large size of Twitter data streams. They propose a new kappa-based sliding window measure for finding classification performance in data streams. They experiment with the Stanford Twitter Sentiment dataset and the Edinburgh Twitter Corpus of [19**]**, using emoticons. They use only unigrams as features. The author report 82.45% accuracy on the test set of the first corpus using Naive Bayes and 86.26% accuracy on the second corpus using stochastic gradient descent (SGD) as best results for the binary classification task [16].

The Author [17] used hash tags and emoticons as noisy labels to label the data set of[18]. They used words, n-grams (2-5), tweet length, punctuation and numbers of exclamation marks, question marks, and quotes and capitalized words in the sentence as features. Additionally, they identify special patterns of high-frequency words and content words and use those as features as well. As best result for their k- nearest neighbor like classification strategy they reported an average harmonic F-score of 86.0% for binary classification. The author find words, patterns and more other features to be useful like punctuation features , while n-grams increase classification performance only marginally, despite of their strategy to use only tokens that exceed a 0.5% frequency threshold in the n-grams in the training data [17].

III. Data Collections

Twitter, with nearly 320 million active users in January 2016and over 350,000 million messages per minute. It has very quickly turned into a very profitable for organizations to monitor their prestige, credit and brands by retrieving and analysing the sentiment of the Tweets messages by the user about their remarks, markets, and other challengers [25]. We use Twitter messages in our experiment for development and training, we use the hashtagged data set (Dataset) from Twitter API. The Twitter API has a parameter that specifies in which language you want to extract tweets and we set this parameter to English. We acquire 5000 tweets of Hash Tag. The number of Twitter messages and the distribution across classes.

Table 1: Dataset Details

No. of Tweets	No. of Positive	No. of Negative	No. of Neutral
5000	788	1464	2748

To create the hashtagged data set, we first filter out duplicate tweets, non-English tweets, and tweets that do not contain hashtags. We investigate the distribution of hashtags and identify what we hope will be sets of frequent hashtags that are indicative of positive, negative, and neutral messages. These hashtags are used to select the tweets that will be used for development and training. The 15 most-used hashtags in the Twitter corpus. We identify all hashtags that appear at least 100 times in the Twitter corpus. From these, we selected the top hashtags. We detect most useful for recognizing positive, negative and neutral tweets as given in Table1.

Opinion Mining and Sentiment Analysis are the classification of Text Mining which refer to the process of retrieving related information and nontrivial patterns from unstructured script topics. Sentiment Analysis and Opinion Mining may appear to be identical as a traditional text mining, but it varies because of following facts. Sentiment Classification is the binary polarity classification which deals with a relatively small number of classes [4]. Sentiment classification is simple task compared to text auto categorization. While Opinion mining displays numerous extra tasks other than sentiment polarity detection.

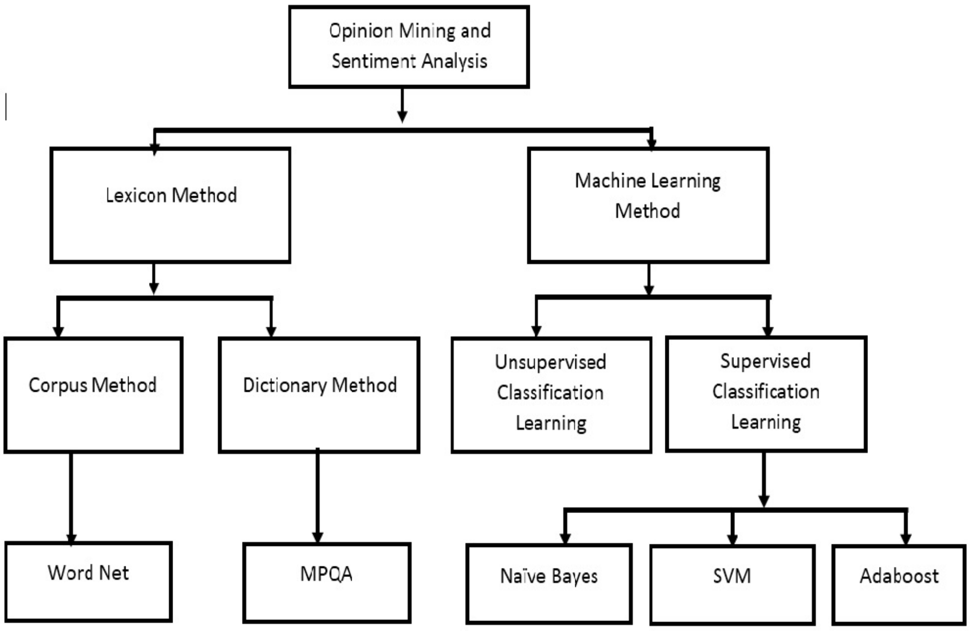
Opinion Mining and Sentiment Analysis
Lexicon Method
Machine Learning Method
Corpus Method
Dictionary Method
Unsupervised Classification Learning
Supervised Classification Learning
Word Net
MPQA
Naïve Bayes
SVM
Adaboost

Figure 1: Sentiment Analysis Flow

IV. PREPROCESSING

A. Tokenization*:*

It is the process of separating a stream of text up into words, symbols and other meaningful elements known as “Tokens”. Tokens can be separated by using whitespace characters and/or punctuation characters. It is done so that we can look at tokens as individual components that make up a tweet.

B. Normalization:

The normalization process is a refinement process after identifying data i.e. the presence of abbreviations within the tweet is identified and then abbreviations are replaced by their actual meaning. e.g., “OMG" by "Oh My God” [11]. Words which have same letters more than two times and does not appear in the lexicon are reduced to the word by eliminating repeated letter and putting it once. For example, the enlarged word "YESSSSS" is diminish as "YES". Tweet may be normalized by converting it to lowercase which makes it’s comparison with an English dictionary easier. Finally, the occurrence of any special Twitter tokens is identified (e.g., #hashtags, URLs) and placeholders indicating the token type are substituted.

C. POS TAGGING (Part of Speech tagging):

POS Tagging is the process of assigning a tag to each word in the sentence as to which grammatical part of speech that word belongs to, i.e. noun, verb, adjective, adverb, coordinating conjunction etc. [1].

V. FEATURES SELECTION

A. N gram Features Selection:

Creating word into N-gram to identify. Removed the stop word into N-gram list [13] is shown in Table 2. Finally, all N-grams are identified in the training data and ranked according to their information gain, measured using Chi-squared. We use the N-grams in a bag of-words fashion.

B. Lexicon Features Selection:

Words listed the MPQA subjectivity lexicon [20] are tagged with their prior polarity: positive, negative, or neutral. We create three features based on the presence of any words from the lexicon.

C. POS Tagging:

For each tweet, we have features for counts of the number of verbs, adverbs, adjectives, nouns.

VI. EXPERIMENT AND RESULT

Our goal was to performing and calculating Chi-squared result using N-gram feature selection. For these experiments is two-fold. First, we need to analyse whether our training data with labels obtain from hash tags (Dataset). It is useful for training data sentiment classification. We need to analyse the usefulness of the features from section for sentiment analysis in Twitter data.

Table 2: N-Gram Features Selection Training Data

Pre Processing Action	Training data for 4000 tweets
Tokenization	61456 Tokens
Stop Word	13969 Stopwords
Word List	11297 Unique word

For our first set of experiments we use the Dataset started by randomly sampling 10% of the Dataset to use as a Testing set. Training Dataset is used for N-gram features selection. To train a classifier, we sample tweets from the training data and use this data to train AdaBoost.MH [21] models. We repeat this process ten times and average the performance of the models.

Table 3: Features of Classifier Performance (AdaBoost)

Features	Recall	Precision	F score
All	0.98	0.98	0.976
N-gram Feature	0.7685	0.770	0.675
Lexicon Features	0.54	0.81	0.52

The average F-measure for the All Tweets using manually baseline and all the features on the Dataset. Table 3 shows the average F-measure for the baseline and two of features: N-grams and lexicon features, Table 3 shows the accuracy for these same experiments. Interestingly, the best performance on the evaluation data comes from using the N-grams together with the lexicon features and features are included, the improvements drop or disappear. The best results on the evaluation data comes from the N-grams, lexical and Twitter features trained on the hashtagged data. The Table 4 showing accuracy of hash tag (Dataset) in two fold cross validation.

Table 4: Average Accuracy of Hash tag (Dataset)

Features	Accuracy
All	97.30%
N-gram Features	88.70%
Lexicon Features	90.12%

VII. CONCLUSION

Our experiments on twitter sentiment analysis show that N-gram features may not be useful for sentiment analysis in the microblogging domain. More research is needed to determine whether the N-gram features are of low accuracy due to the results of the N-gram tokenization features are less useful for sentiment analysis in this domain. Features from an existing sentiment lexicon were somewhat useful in conjunction with microblogging features, but the microblogging features (i.e., the presence of intensifiers and positive/negative/neutral emoticons and abbreviations) were clearly the most useful. Using hashtags to collect training data did prove useful. However, which method produces the better training data and find whether the two sources of training data are compatible or interdependent, it may depend on the type of features used. In that experiments combination of features is better accuracy show that when microblogging features are included, the benefit of Hashtag training data is lessened.

REFERENCES

1. Varsha Sahayak, Vijaya Shete, Apashabi Pathan, “Sentiment Analysis on Twitter Data”. International Journal of Innovative Research in Advanced Engineering (IJIRAE) Issue 1, Volume 2 (January 2015) page no. 178-183, January 2015.
2. Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media page no. 178-185 (2010)
3. Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang, “Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach”. Conference on Information and Knowledge Management 2011 page No. 1031 -1040
4. Kishori K. Pawar, Pukhraj P Shrishrimal, R. R. Deshmukh, “Twitter Sentiment Analysis: A Review”. International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015 page no. 957 -964
5. Tobias Günther, “Sentiment Analysis of Microblogs” page no. 8 , June 2013
6. Waghode Poonam B,Prof. Mayura Kinikar, “Twitter Sentiment Analysis with Emoticons”. International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 4 Issue 4 April 2015, Page No. 11315-11321
7. Geeta G. Dayalani, Prof. B. K. Patil, “Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets”. International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014 page no. 438-445
8. Swapna R. Kharche, Prof. Lokesh Bijole “Review on Sentiment Analysis of Twitter Data”. International Journal Of Computer Science And Applications Vol. 8, issue.2 , page No. 53-56Apr-June 2015
9. Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu, “Unsupervised Sentiment Analysis with Emotional Signals”. International World Wide Web Conference Committee (IW3C2). Page no. 978-989, 2013
10. Waghode Poonam B, Prof. Mayura Kinikar, “Twitter Sentiment Analysis with Emoticons”. International Journal Of Engineering And Computer Science Volume 4 Issue 4 Page No. 11315-11321,April 2015,
11. Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!”. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media page no. 538-541, 2011
12. Alec Go, Richa Bhayani, and Lei Huang. “Twitter sentiment classification using distant supervision”. CS224N Project Report, Stanford, 2009.
13. Alexander Pak and Patrick Paroubek. “Twitter as a corpus for sentiment analysis and opinion mining”. In Proceedings of LREC, Page no. 1320-1326 ,2010.
14. Luciano Barbosa and Junlan Feng. “Robust sentiment detection on twitter from biased and noisy data”. In Proceedings of the 23rd International Conference on Computational Linguistics pages no 36–44, 2010
15. Adam Bermingham and Alan F Smeaton. “Classifying sentiment in microblogs: is brevity an advantage?”. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836. ACM, 2010.
16. Albert Bifet and Eibe Frank. “Sentiment knowledge discovery in twitter streaming data”. In Discovery Science, pages 1–15. Springer, 2010.

17. Dmitry Davidov, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys". In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 241–249. ACL, 2010.
18. Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. "Improved part-of-speech tagging for online conversational text with word clusters". In Proceedings of NAACL 2013, 2013.
19. Sasa Petrovic, Miles Osborne, and Victor Lavrenko. "The Edinburgh twitter corpus". In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 25–26, 2010
20. Janyce Wiebe, Theresa Wilson, and Claire Cardie. "Annotating expressions of opinions and emotions in language". Language Resources and Evaluation, 39(2-3):165–210, 2005.
21. Schapire, R. E., and Singer, Y."BoosTexter: A boosting-based system for text categorization". Machine Learning 39(2/3):135–168, 2000

BIOGRAPHY

Sachin Madhukar Ramteke Received B.E in Computer Technology from K. D. K. College Of Engineering, Nagpur, R.T.M.N. University, Nagpur in 2014. Currently pursuing M.Tech in Computer Science and Engineering from Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.

Dr. Sachin N. Deshmukh completed his Ph.D. and M.E in Computer Science and Engineering. He is currently a Professor in Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India. His research area is "Text mining, Social Web mining and Intension Mining". He is a member of Adhoc Board of Studies in BioInformatics and Liberal arts of Dr. B. A. M. University Aurangabad and Adhoc Board of Computer Science at Shivaji University Kolhapur.