International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Predicting Sentiment of Tweets

Sandip D. Mali[1], Dr. S.N. Deshmukh[2], Ashish A Bhalerao[3]

[1] *Research student, Dept. of CS & IT, Dr. BAMU, Aurangabad, Maharashtra, India.*
[2] *Professor, Dept. of CS & IT, Dr. BAMU, Aurangabad, Maharashtra, India.*
[3] *Research student, Dept. of CS & IT, Dr. BAMU, Aurangabad, Maharashtra, India.*

Abstract - *Today microblogging website are very popular like twitter on which user post their views, opinion etc. The information is generated either through computer or mobile by one user and many can view them. In this paper we focus on using twitter for sentiment analysis. Sentiment analysis is challenging task for this we can use various machine learning algorithm for it, like Naive Bayes, SVM, maximum entropy etc. Sentiment analysis refers to predicting or telling the document or sentence text holds positive, negative or neutral opinion on some target. The aim of this paper is to revise the previous work and compare various techniques used in the sentiment analysis and make a broad view on sentiment analysis work. We have enclose the work from beginning to recent.*

Key Words: Twitter, Sentiment Analysis, Opinion Mining, Sentiment Classification.

1. INTRODUCTION

Launched on July 13, 2006, Twitter[1] is an extremely popular online microblogging service. It has a very large user base, consist several millions of users 23M unique users[2]. Twitter is most popular microblog website because it allows user to post its views, opinion or anything and one important thing is that message length should be less than 140 characters. Thus tweets are short text message and greatly used in sentiment analysis from text. Tweets has some common features that are explained below,

1. Username- Username is often twitter username included in tweet to direct their message. A de facto standard is to include @ before the username for e.g. @IPL2015.
2. Hash Tag- twitter allows their user to tag tweets using hash tag which has the form "#<hash-tag>" Users can use this to express what their tweet is primarily about by using keywords that best represent the content of the tweet.
3. RT: If a tweet is interesting enough, users might republish that tweet, commonly known as retweeting, and twitter employs "RT" to represent retweeting.

For product marketing the social media is widely used as a tool and its advantage is taken by peoples, governments, corporations and schools. Tweets that contain opinions are important because whenever people need to make a decision, they want to know other's opinion. The same is also true for organizations. In sentiment analysis, the first question coined is, what sentiment is? Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It exemplifies a large problem space. There are also many names and slightly different tasks, e.g. sentiment analysis, opinion mining, opinion extraction, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. [4].

Sentiment analysis has three different levels, first at document level is to whole document express a positive, negative or neutral sentiment, but this would consider document from the single user and may be related to more than one product. Second is sentence level, a single sentence is used for sentiment analysis, it gives better result than document level. Third is, aspect level, it performs fine-grained analysis. Instead of looking language constructs, it directly looks at the opinion itself [4]. Sentiment classification classifies sentence or document into subjective or objective. Subjective class means usually gives personal views or opinions and objective class expresses some factual information and facts. For e.g. "This cell phone works very well" is subjective sentence and "camera is used to capture the scene in our life" is objective sentence. For sentiment analysis we need only subjective sentences because objective sentence does not hold any subjective information and thus reduce accuracy.

1. http://www.twitter.com
2. http://blog.compete.com/2010/02/24/compete-ranks-top-sites-for-january 2010/.

2. Related work

Most of the text on microblogging websites are textual information, identifying their sentiments has become an important issue. The research in the field started with sentiment classification, which treated the problem as a text classification problem. Text classification using machine learning is a well-studied field [1], and there is ample research of the effects of using various machine learning techniques (Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) [2]. After building and testing models using Naive Bayes, Maximum entropy and Support Vector Machines (SVM), they reported that SVM showed the best performance [3].

In [3] they propose a new system architecture that automatically analyze the sentiment of microblogs or twitter, they combine this system with manually annotated data. They extracts tweets that contain opinion and filter out non opinion tweets or messages and determine their sentiment direction. For this they use Naive Bayes classifier. For short text classification they used Mutual Information and X^2 test. The final step is to determine sentiment orientation of the tweets i.e. positive or negative and they got accuracy about 67.8% for unigram and 70.39% for opinion miner.

In [5] they build models for two classification tasks: a binary task of classifying sentiment into two class positive and negative classes and a 3-way task of classifying sentiment into three class positive, negative and neutral classes. There experiments show that a unigram model is truly a hard baseline achieving over 20% over the chance baseline for both classification tasks. There feature based model that uses only 100 features achieves similar accuracy as the unigram model that uses over 10,000 features. There tree kernel based model outperforms both these models by a significant margin. They also experiment with a combination of models: combining unigrams with our features and combining our features with the tree kernel. Both these combinations outperform the unigram baseline by over 4% for both classification tasks. They use manually annotated Twitter data for their experiments.

In [6] uses simple unsupervised machine learning algorithm for classifying reviews. The classification is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations (e.g., "subtle nuances") and a negative semantic orientation when it has bad associations (e.g., "very cavalier"). The first step is to use a part-of-speech tagger to identify phrases in the input text that contain adjectives or adverbs. The second step is to estimate the semantic orientation of each extracted phrase. A phrase has a positive semantic orientation when it has good associations (e.g., "romantic ambience") and a negative semantic orientation when it has bad associations (e.g., "horrific events"). The third step is to assign the given review to a class, recommended or not recommended, based on the average semantic orientation of the phrases extracted from the review. If the average is positive, the prediction is that the review recommends the item it discusses. Otherwise, the prediction is that the item is not recommended. The PMI-IR algorithm is employed to estimate the semantic orientation of a phrase. PMI-IR uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words or phrases. The semantic orientation of a given phrase is calculated by comparing its similarity to a positive reference word ("excellent") with its similarity to a negative reference word ("poor"). In experiments with 410 reviews from opinions, the algorithm attains an average accuracy of 74%.

In [7] they build a web based system called SES. The system is to predict sentiment on document level and sentence level. A document often contains more than one sentences and split it into sentences which are the input of the system. They conduct four experiments on Facebook comments and twitter tweets using four different machine learning models: decision tree, neural network, logistic regression, and random forest. There experiment results show that random forest model reaches highest accuracy.

Sentiment analysis can be classified into 2 branches. On one hand, they take state-of-the-art sentiment identification algorithms to solve problems in real applications such as summarizing customer reviews [8], ranking products [9], finding product features that imply opinions [10].

In [11] analyzes tweet sentiments about movies and attempts to predict box office revenue. The authors define different metrics to measure the popularity/sentiment of a movie and then use a linear regression model to predict box-office. On the other hand, researchers put their focus on discovering new sentiment algorithms. Bag-of-Words approach produces domain-specific lexicons. There is a vast body of research which attempts to incorporate these interactions as features in a machine learning model [12]. Rule-based approaches has been studied by many researchers [13].

3. KEY APPLICATIONS

Opinions are so important that whenever one needs to make a decision, one wants to hear other's opinions. This is true for both individuals and organizations. The technology of opinion mining thus has a tremendous scope for practical applications.

1. Individual consumers: If an individual wants to purchase a product, it is useful to see a summary of opinions of existing users so that he/she can make an informed decision. This is better than reading a large number of reviews to form a mental picture of the strengths and weaknesses of the product. He/she can also compare the summaries of opinions of competing products, which is even more useful.
2. Organizations and businesses: Opinion mining is equally, if not even more, important to businesses and organizations. For example, it is critical for a product manufacturer to know how consumers perceive its products and those of its competitors. This information is not only useful for marketing and product benchmarking but also useful for product design and product developments.

4. EXISTING TECHNIQUES-

If we look broadly on the related work, we can say that for classification naive bays, support vector machine (SVM), maximum entropy, random forest etc. machine learning algorithms are primarily used and features extracted are ngrams, bigrams, unigrams, mutual information etc. Another [5] significant effort for sentiment classification on Twitter data is by Barbosa and Feng (2010). They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words.

5. CONCLUSION

From this paper we conclude that Naive Bayes and Random Forest algorithm gives good result and accuracy may be reduced in multiple domains. The main task to increase the accuracy is to remove a text that does not hold any opinion and subsequently this reduce search space and reduce execution time.

6. REFERENCES

1. Fabrizio Sebastiani, "Machine learning in automated text categorisation". Technical Report IEI-B4-31-1999, Istituto di Elaborazione.

2. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques.", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

3. Po-Wie Liang and Bi-Ru Dai, "Opinion Mining on Social Media Data", 2013 IEEE 14th International Conference on Mobile Data Management, 978-0-7695-4973-6/13 $26.00 © 2013 IEEE ,DOI 10.1109/MDM.2013.73

4. Bing Liu," Sentiment Analysis and Opinion Mining", April 22, 2012.

5. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau, "Sentiment analysis of twitter data", In Proceedings of the Workshop on Languages in Social Media, pages 30–38. Association for Computational Linguistics 2011.

6. P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", ACL'02, 2002.

7. Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo, Ankit Agrawal, Diana Palsetia , Kathy Lee , Wei-keng Liao , Alok Choudhary, "SES: Sentiment Elicitation System for Social Media Data", Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, p.129-136, December 11-11, 2011

8. M. Hu, and B. Liu, "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177, 2004.

9. K. Zhang, and R. Narayanan, and W. Liao, and A. Choudhary, Voice of the Customers: Mining Online Customer Reviews for Product Feature-

based Ranking. 3rd Workshop on Online Social Networks, 2010.

10. Popescu, and O. Etzioni, Extracting product features and opinions from reviews. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages339–346,2005.

11. S. Asur, and B. A. Huberman, "Predicting the future with social media", Arxivpreprintar Xiv:1003.5699, 2010.

12. X. Ding, and B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining", Proceedings of the international conference on Web search and web data mining, pages 231-240, 2008.

13. Y. Choi, and C. Cardie, "Learning with compositional semantics as structural inference for subsentential sentiment analysis", Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages793–801,2008.