International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319 – 6505, Impact Factor: SJIF: 5.438
Available Online at www.journalijcar.org
Volume 6; Issue 2; February 2017; Page No. 2071-2074

Research Article

SENTIMENT ANALYSIS ON TWITTER FEEDS

Aakash Patel., Ashutosh and Priyansh Gupta

Department of Information Technology IMS Engineering College Ghaziabad UP

ARTICLE INFO

Article History:

Received 29th November, 2016
Received in revised form 30thDecember, 2016
Accepted 4th January, 2017
Published online 28th February, 2017

Key words:

Polarity, Frequency, Bayes Classifier

ABSTRACT

In this paper we examine sentiment analysis on Twitter data. We are using polarity feature to examine the polarity between [-5,5] and second is frequency feature to examine how many times the word is repeating. This analysis utilises the naive Bayes Classifier to classify Tweets into positive, negative or neutral sets. Further classification is in to extremely negative and extremely positive sets. We present experimental evaluation of our dataset and classification results. A case study is presented to illustrate the use and effectiveness of the proposed system.

INTRODUCTION

Sentiment analysis is a relatively different area, which deals with extracting user opinion automatically. An example of a positive sentiment is, “natural language processing is fun” similarly, a negative sentiment is “it’s a terrible day, we will stay here at home”. There are many sentences which do not show any expression, such as news headlines, for example “company shelves wind sector plans”.

There are different ways in which micro blogging data can be increased to provide a better understanding of user opinion such problems are at the core part of natural language processing (NLP) and data mining research.

It is done to find the polarity of the words. Apart from this it classifies the emotions of the state in various categories such as sad, angry, and happy. Sentiment analysis determines the attitude of the speaker and their emotional state. Analysing tweets means to understand the feelings of the writer and what they think about the related topic. It includes the opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life.

Sentiment analysis grouped in to three main categories:-

1. Knowledge based technique
2. Statistical approaches
3. Hybrid approaches

Knowledge based technique works in the IF THEN rules. It searches particular types of words based on that it classifies the text. Words like happy, sad, afraid, bored etc are the identifier of the particular emotions.

Statistical method take advantage from machine learning such as latent semantic analysis, support vector machines, bag of words and semantic orientation that is Point wise mutual information. In more precise way to detect the holder of a sentiment i.e. the person who maintains that affective state and the target i.e. the entity about which the effect is felt.

Hybrid approaches takes advantage from both machine learning and elements of knowledge representation like ontologies and semantic network to detect semantics that are expressed in tenuous manner.

keywords - natural language processing, data mining, emotions, latent semantic, support vector machines

Related Work

Applying sentiment analysis on Twitter is the upcoming trend with researchers recognizing the scientific trials and its potential applications. The challenges unique to this problem area are largely attributed to the dominantly informal tone of the micro blogging. Pak and Paroubek rationale the use micro blogging and more particularly Twitter as a corpus for sentiment analysis.

Micro blog data like Twitter, on which users post real time reactions to and opinions about “everything”, poses newer and different challenges. Some of the early and recent results on sentiment analysis of Twitter data are by Go *et al*. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go *et al*. (2009) use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like “:)” “:-)” as positive and negative emoticons like “:(” “:-(” as negative. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM), and they report SVM outperforms other classifiers. In terms of feature space, they try a Unigram, Bigram model in conjunction with parts-of-speech (POS) features. They note that the unigram model outperforms all other models.

Specifically, bigrams and POS features do not help. Pak and Paroubek (2010) collect data following a similar distant learning paradigm. They perform a different classification task though: subjective versus objective. For subjective data they collect the tweets ending with emoticons in the same manner as Go *et al*. (2009).

keywords - MaxEnt, unigram model, bigram model, parts of speech,

Data Description

Twitter is a micro blogging websites that let the user to post 140 words long messages called tweets. Since there is a restriction on the usage of words per tweet i.e 140 words, the users frequently uses acronyms or intentionally omits some alphabets to make words shorts. Facial expressions pictorially represented using words and punctuations known as Emoticons are also used widely in tweets. Targets are to refer other twitter users using '@' symbol. Twitter users uses Hash tags to mark different topics using '#' symbol.

Twitter API was used to acquire the tweets . Twitter has let it data publicly available through Twitter API. The data is collected by streaming the real time data and archiving it with the use of Streaming API. We gathered a sample of 600 tweets based on a particular topic. The Tweets retrieved also contained languages other than eliminated languages. The Remaining tweets were marked as positive negative or neutral. The Tweets that were not understood by human annotator were also eliminated.
keywords - emoticons, targets

Resources

The main resources used area List of words with polarity, an Emoticon Dictionary and an acronym dictionary. The list of words(AFINN-111 Dictionary) contains 2490 words collected through an online source rated for valence with an integer between minus five (negative) and plus five (positive).

Table -1 AFINN-111 Dictionary

Words	Polarity
Adore	3
Aggressive	-2
Bitch	-5
Breathtaking	5
Celebrate	3

170 emoticons that were listed on the Wikipedia along with their facial expression and meaning were labelled in the Emoticon dictionary. Each emoticon was labelled from the set of labels Extremely negative, Extremely Positive, Positive, Neutral and Negative.

Table 2 Emoticons Dictionary

Emoticons	Polarity
:) :) : o) :] : 3	Positive
: D : D8DxDXD	Extremely Positive
: = : = = = = < =3	Negative
D : D8D =DXv:vDx	Extremely Negative
>:)B)B) :) :) >	Neutral

About 5000 acronyms were detected along with their translations. For Example lol is translated to laugh out loud.

Table 3 Acronym Dictionary

Acronym	Polarity
admin	administrator
asap	As soon as possible
omg	Oh my god
lol	laugh out loud
rofl	rolling over floor laughing

keywords - emoticon dictionary

System Architecture and Processing

Pre-Processing of Tweets

The Pre-Processing of tweets is done in the following steps.
Tokenization-Tokenization is one of the most basic, yet most important, steps in text analysis. Tokenisation is the process of splitting a stream of text into sub stream which is known as token. Tokens are usually words or phrases. While this is a well understood problem with several out-of-the-box solutions from popular libraries, Twitter data pose some challenges because of the nature of the language. Emoticons are replaced with their corresponding polarity using emoticon dictionary. URLs are replaced with |U| and tags are replaced with |T|.

Normalization- In normalization process the abbreviations present in the tweets are noted and are replaced by their translations using the acronym dictionary (ex- BRB-> Be Right Back). Informal Intensifiers such as all caps and character repetition i.e a character appearing more than thrice is also identified and reduced (for ex- coooooool is converted to coool.).

Features

We use a variety of features for our classification experiments. For the baseline, we use unigrams and bigrams. We also include features typically used in sentiment analysis, namely features representing information from a sentiment lexicon and POS features. Finally, we include features to capture some of the more domain-specific language of micro blogging.

n-gram features

To identify a set of useful n-grams, we first remove stop words. We then perform rudimentary negation detection by attaching the word not to a word that precedes or follows a negation term. This has proved useful in previous work (Pak and Paroubek 2010). Finally, all unigrams and bigrams are identified in the training data and ranked according to their information gain, measured using Chi-squared. For our experiments, we use the top 1,000 n-grams in a bagof-words fashion.

Lexicon features

Words listed the MPQA subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2009) are tagged with their prior polarity: positive, negative, or neutral. We create three features based on the presence of any words from the lexicon.

Part-of-speech features

For each tweet, we have features for counts of the number of verbs, adverbs, adjectives, nouns, and any other parts of speech.

Micro-blogging features

We create binary features that capture the presence of positive, negative, and neutral emoticons and abbreviations and the presence of intensifiers (e.g., all-caps and character repetitions). For the emoticons and abbreviations, we use the Internet Lingo Dictionary (Wasden 2006) and various internet slang dictionaries available online.

keywords - tokenization, normalization, intensifier

Case Study

This section will show the results of the classifiers and the analysis carried out on a case study. With the above described software, it is possible to obtain some training sets for the classifiers. In our case study, they consist of:

- 56000 instances (polarity)
- 16000 instances (subjectivity)

These instances have been obtained by exploring more than 60 channels on the social network. In the generated models, the selected features are consistent with our expectations: the typical expressions of a certain feeling (such as smileys, or some words that express appreciation or disgust) show a higher probability of belonging to the class of that feeling, rather than to the class of the opposite sentiment.

The obtained results by the classifiers using cross validation (with folds = 10) on the training sets showed an accuracy of:

- 72.35% (polarity classifier)
- 73.60% (subjectivity classifier)

These results show that the model of the classifiers contains effective features for the recognition of the sentiment of a message.

The case study that was considered in this work was the recent demonetization in India.

The total of 2000 tweets were collected for the experimentation.

For evaluating the performances of our system, we conducted a simple survey through a group of persons in our department. In this way, we selected and classified 100 messages that show a clear opinion on the singer. Then, we used those messages as a test. The results of the classifiers showed an accuracy of 84% for the polarity and 88% for subjectivity.

keywords - training set, accuracy

CONCLUSION

Our experiments on twitter sentiment analysis shows that part-of-speech features may not be useful for sentiment analysis in the micro blogging domain. More research is needed to determine whether the POS features are just of poor quality due to the results of the tagger or whether POS features are just less useful for sentiment analysis in this domain. Features from an existing sentiment lexicon were somewhat useful in conjunction with micro blogging features, but the micro blogging features (i.e., the presence of intensifiers and positive/negative/neutral emoticons and abbreviations) were clearly the most useful. Using hash tags to collect training data did prove useful, as did using data collected based on positive and negative emoticons. However, which method produces the better training data and whether the two sources of training data are complementary may depend on the type of features used. Our experiments show that when micro blogging features are included, the benefit of emoticon training data is lessened.

keywords - tagger, whether, lessended

Reference

1. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau.“Sentiment analysis of Twitter data”, Procs of the Workshop on Languages in Social Media (LSM '11), pp. 30-38, 2011.
2. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Vol.11, No. 1, 2009.
3. Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.
4. L. Bing, “Web Data Mining: Exploring Hyperlinks, Contents, and Usage
5. Data,” Data-Centric Systems and Applications, 2nd ed., Springer, pp. 1- 603, 2011.
6. D. Kim, Y. Jo, I-C. Moon, and A. Oh, Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users, Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems (CHI 2010).
7. L. Bing. Sentiment Analysis and Opinion Mining, Morgan & Claypool
8. Publishers, 2012.
9. Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proc. of Coling*.
10. Jalaj S. Modha,Gayatri S. Pandi, Sandip J. Modha," Automatic Sentiment Analysis for Unstructured" ,DataVolume 3, International Journal of Advanced Research in Computer Science and Software Engineering Issue 12, December 2013.
11. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC.Vol. 10. 2010.
12. Rajni Singh, Rajdeep Kaur,"Sentiment Analysis on Social Media and Online Review" International Journal of Computer Applications (0975 8887) Volume 121 No.20, July 2015.
13. F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, 1-47, March 2002.
14. Acronym list. [Online]. Available: http://www.noslang.com/dictionary/
15. Afinn-111. [Online]. Available: http://www2.imm.dtu.dk/pubdb/views/
16. publication details.php?id=6010
17. Emoticon list. [Online]. Available: http://en.wikipedia.org/wiki/List of
18. emoticons
19. Twitter. [Online]. Available: http://twitter.com
20. Twitter nlp. [Online]. Available: https://github.com/brendano/
21. ark-tweet-nlp/tree/master/src/cmu/arktweetnlp
22. M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, How Does the

Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? Proc. of the 4th Int'l AAAI Conference on Weblogs and Social Media, George Washington University, Washington, DC, May 23-26, 2010.

23. Walaa Medhat, Ahmed Hassan, Hoda Korashy," Sentiment analysis algorithms and applications: A survey", Ain Shams University Ain Shams Engineering Journal 2014.
24. P. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. ACL
25. J. Read, Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification, Proc. of ACL-05, 43rd Meeting of the Association for Computational Linguistics, 2005.

****