IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS ON TWITTER DATA USING MACHINE LEARNING

Koppada Hema

M. Tech

Department Of Computer Science And Systems EngineeringAndhra University College Of Engineering Visakhapatnam-530003.

ABSTRACT

This project tackles the issue of tweet sentiment analysis, which involves categorizing tweets into those that indicate good, negative, or neutral mood. Twitter is a social networking and microblogging website that enables users to post 140-character maximum status updates With over200 million registered users, of which 100 million are active users and half of them log in at least daily, it is a service that is rapidly growing. Each day, it generates approximately 250 million tweets.We intend to reflect the public opinion by assessing the feelings stated in the tweets in light of thissignificant usage. Numerous applications require the analysis of public mood, including businessesattempting to gauge the market response to their products, the prediction of political outcomes, andthe analysis of socioeconomic phenomena like stock exchange. The goal of this project is to createa practical classifier that can accurately and automatically identify the sentiment of an unidentifiedtweet stream.

1. INTRODUCTION

As opposed to more traditional internet articles and web blogs, we believe that twitter providesa more accurate representation of public opinion. The rationale is that, when compared to conventional blogging platforms, twitter has a considerably higher volume of pertinent material. Because there are many more users who tweet than those who regularly update web blogs, the response on twitter is also quicker and more inclusive. In macro-scale socioeconomic phenomena like anticipating the stock market rate of a certain company, public sentiment analysis is crucial. This might be accomplished by examining the general public opinion of the company over time and utilising economics methods to determine the relationship between the public opinion and the firm'sstock market valuation. Since Twitter allows us to download streams of geo-tagged tweets for specific locations, businesses may also evaluate how well their product is responding in the marketand which areas of the market are it having a favourable response and in which a bad response. If businesses can gather this data, they may analyse the causes of geographically diverse responses and sell their products more effectively by coming up with suitable solutions like forming appropriate market groups. Another developing use for sentiment analysis is making predictions about the outcomes of popular political elections and surveys. In one such study, which was carriedout in Germany for the purpose of forecasting the results of federal elections, Tumasjan et al. cameto the conclusion that Twitter is a good indicator of offline mood.

Domain Introduction

Twitter sentiment analysis falls under the categories of "Pattern Classification" and "Data Mining" in this project. Both of these concepts are intimately related and intertwined, and they mayboth be properly defined as the automatic (unsupervised) or semi-automatic (supervised) process of finding "useful" patterns in vast sets of data. The project would heavily rely on "Natural Language Processing" techniques to extract important patterns and features from the massive dataset of tweets and on "Machine Learning" techniques to accurately categorize individualunlabeled data samples (tweets) according to whichever pattern model best describes them. Formal language-based features and informal blogging-based characteristics can be separated into two primary groupings that can be utilized for modeling patterns and classification. Language-based features are those that deal with formal linguistics and include the parts of speech that each sentenceis tagged with as well as the preceding sentiment polarity of certain words and phrases. Prior sentiment polarity describes the inherent innate inclination of some words and phrases to express particular and specific sentiments in general. For instance, the term "excellent" carries a strong connotation of positivity, whereas the word "evil" carries a strong connotation of negativity. Therefore, whenever a word with a positive connotation is employed in a sentence, there is a good possibility that the sentence as a whole will be positive. The two categories of classification methodsare supervised vs. unsupervised, and non-adaptive vs. adaptive/reinforcement methods. When we have pre-labeled data samples available, we may train our classifier using a supervised approach. In order to classify an unlabeled data sample according to the pattern that best represents it, the classifier must first be trained by using the pre-labeled data to extract features that best model the patterns and differences between each of the distinct classes.

2. LITERATURE SURVEY AND RELATED WORKLimitations of Prior Art

Since sentiment analysis in the context of microblogging is still a relatively unexplored area of study, there is still much to be discovered. A substantial amount of similar past work has been doneon phrase level sentiment analysis as well as sentiment analysis of user reviews, documents, web blogs, and publications. These are distinct from Twitter mostly due to the 140-character character limit per tweet, which pushes users to express opinions in extremely condensed prose. The best sentiment classification results are obtained using supervised learning methods like Naive Bayes and Support Vector Machines, however the manual labeling needed for this method is quite expensive. Unsupervised and semi-supervised methods have received some attention. Many researchers experimenting with novel characteristics and classification methods simply compares their findings to baseline performance. In order to choose the best features and most effective classification techniques for particular applications, proper and formal comparisons between these findings obtained by various features and classification techniques are required.

Related Work

Due to its simplicity and effective performance, the bag-of-words model is one of the most frequently used feature models for practically all text classification problems. The approach considers the text to be categorized as a bag or collection of distinct words without any connectionbetween or dependent on one another; thus, it entirely ignores the grammar and word order withinthe text. This design is also quite well-liked in Several researchers have employed sentiment analysis. Using unigrams as features is the simplest method to include this model in our classifier.In our text, an n-gram is typically defined as a contiguous series of "n" words that stand alone from all other words and grammatical structures. An online tool called the Multi-Perspective-Question- Answering (MPQA) has a subjectivity lexicon that categorizes 4,850 terms as either "positive" or "negative" and having "strong" or "weak" subjectivity. Another tool that indicates the likelihood that a word falls into the positive, negative, or neutral categories is SentiWordNet3.0.

3. IMPLEMENTATION STUDY

Unigrams are just a grouping of distinct words in the text that need to be categorized, and we make the assumption that the presence or absence of other words in the text will not have an impacton the likelihood of recurrence of any given word. Although it is a relatively simplistic assumption,it has been demonstrated to offer fairly acceptable performance. Assigning unigrams with a certainprior polarity and averaging the overall polarity of the text are two straightforward methods for using unigrams as features. The overall polarity of the text can be computed by adding the prior polarities of individual unigrams. Prior polarity of words can be used as a feature in three different ways. Using publicly accessible internet lexicons or dictionaries that map a word to its preceding polarity is the easier unsupervised method.

4. PROPOSED METHODOLOGY

The method involves building a unique prior polarity dictionary from our training data based onhow often each word appears in each specific class. For instance, if a specific word appears more frequently in the positive labelled sentences in our training dataset (relative to other classes), then we can determine the likelihood of that term being in the positive class is greater than the likelihoodthat it will be in any other class. It has been demonstrated that this strategy performs better since the prior polarity of the words is more matched and fitted to a certain sort of text and is not as general as in the previous approach. The latter, however, requires supervision because the trainingdata must first be classified into the proper classes in order to determine the relative frequency of aword in each class Kouloumpis and others.

5. METHODOLOGIESMODULES

We will first discuss our findings for the classifications of objective vs. subjective and positivevs. negative. These findings serve as the foundation of our classification strategy. For both of theseoutcomes, we only use the traits that made the short list. This indicates that we have 5 features forthe objective/subjective classification and 3 features for the positive/negative classification.

We utilize the Naive Bayes classification algorithm for both of these outcomes since that is the algorithm we really use in our initial step of categorization. Furthermore, 10-fold cross validation was used to get all of the provided statistics. Each of the 10 values we receive from the cross validation is averaged.

We specify that only subjectively categorized tweets are utilized to determine the results of polarityclassification, which distinguishes between positive and negative classifications. Nevertheless, in the case of the final classification technique, any such requirement is eliminated, and essentially, classifications for objectivity and polarity are applied to all tweets regardless of whether they are marked as objective or subjective. The accuracy of neutral class decreases from 82.1% to 73% if we use our classification instead of Wilson et al.'s (results are shown in Tables 2 and 3 of this study), but this is still an improvement. However, we report noticeably better results for all other courses. Wilson et al.'s results, while not based on Twitter data, are from phrase level sentiment analysis, which is conceptually quite similar to Twitter sentiment analysis.

Koulompis et al. report an average F- measure of 68% versus these results. Their average F- measure, however, falls to 65% when they factor in another subset of their data (which they referto as the HASH data). In comparison, we attain an average F-measure of over 70%, demonstratingsuperior performance to both of these outcomes. Additionally, we only employ 8 features and 9,000 tagged tweets, whereas their technique uses roughly 15 features overall and more than 220,000 tweets for their training set. Our unigram word models are also less complex than theirs because their word models include negation.

6. RESULTS AND DISCUSSION SCREEN SHOTS

Classes	True Positive	False Positive	Recall	Precision	F-measure
Objective	0.73	0.26	0.74	0.73	0.73
Subjective	0.74	0.27	0.725	0.73	0.73
Average	0.73	0.27	0.73	0.73	0.73

Fig 1: Results From Objective / Subjective Classification

Classes	True Positive	False Positive	Recall	Precision	F-measure
Positive	0.84	0.19	0.86	0.84	0.85
Negative	0.81	0.16	0.79	0.81	0.80
Average	0.83	0.18	0.83	0.83	0.83

Fig 2:- Results From Polarity Classification (Positive / Negative)

Features	Naive Bayes	Max Entropy	SVM
Unigram	81.3%	80.5%	82.2%
Bigram	81.6%	79.1%	78.8%
Unigram + Bigram	82.7%	83.0%	81.6%
Unigram + POS	79.9%	79.9%	81.9%

Fig 3: - Positive / Negative Classification Results Presented By (1-9)

Classes	True Positive	False Positive	Recall	Precision	F-measure
Objective	0.77	0.27	0.77	0.75	0.76
Positive	0.66	0.11	0.66	0.70	0.68
Negative	0.60	0.10	0.59	0.61	0.60
Average	0.70	0.19	0.703	0.703	0.703

Fig 4: Final Results Using Svm At Step 2 And Naive Bayes At Step 1

7. CONCLUSION AND FUTURE SCOPE

Particularly in the area of microblogging, sentiment analysis is still a work in progress and farfrom being finished. Therefore, we offer a few concepts that we believe are worth pursuing in thefuture and could lead to even better performance. We currently only use the most basic unigrammodels, but we may make those models better by including additional data, such as how closelya word is related to a negation word. In order to add the effect of negation into the model, we may define a window previous to the word under examination (the window could, for example,be of two or three words). The polarity should be affected more by the negation word the closer it is to the unigram whose prior polarity is to be determined. For instance, if the negative is immediately behind a word, it might just change the word's polarity. The further the negation isfrom the word, the less impact it should have the impact of bigrams and trigrams may be studied,but aside from that, we are just concentrating on unigrams at this time. According to the literature review section, bigrams combined with unigrams typically result in improved performance. We are currently investigating parts of speech outside of the unigram models, although we may attempt to include POS datawithin them in the future. So let's imagine that, rather than computing a single probability for each word like P(word | obj), we may instead have several probabilities for each word depending on which Part of Speech it belongs to. P(word | obj, verb), P(word | obj, noun), and P(word | obj, adjective), for instance, may exist. According to Pang et al. [5], who employed a somewhat similar methodology, adding POS information to every unigram does not significantly affect performance (with Naive Bayes performing slightly better and SVM having a slight decrease in performance), but if only adjective unigrams are added,accuracy decreases significantly. Not least of all, we can try to simulate human faith in our system. A tweetcan be plotted on the 2-dimensional objectivity/subjectivity and positivity/negativity planes using five human labelers, for instance, to distinguish between tweets where all five labels agree, only four agree, onlythree agree, or when no majority vote is attained. For creating optimum class borders, we might create a custom cost function that gives the maximum weight to tweets with agreement from all five labels and decreases in weight as the number of agreements increases. In this approach, sentiment analysis can illustrate the effects of human confidence.

8. REFERENCES

[1] Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. *Discovery Science, Lecture Notes in Computer Science*, 2010,
Volume 6332/2010, 1-15, DOI: 10.1007/978-3-642-16184-1_1

[2] Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. *Project Technical Report, Stanford University*, 2009.

[3] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining.*In Proceedings of international conference on Language Resources and Evaluation (LREC)*, 2010. [4]Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe. Predicting Electionswith Twitter: What 140 Characters Reveal about Political Sentiment. *In Proceedings of AAAI Conferenceon Weblogs and Social Media (ICWSM)*, 2010.