International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 09, September-2015

Sentiment Analysis for Frontier Security issues using Phrase Patterns

Mrs. Suneetha Eluri
Asst. Prof
CSE
JNTU KAKINADA
East Godavari, India

Ramakrishna Reddy Madireddy
CSE
JNTU KAKINADA
East Godavari, India

***Abstract*— Sentiment analysis is an area of data mining where we study the opinions and appraisals of people using natural language processing and computational linguistics. But the new challenge task for sentiment analysis is on frontiers related data such as migration and firing at LOC etc. Retrieve tweets from twitter and pre-process them. Integrate LinearSvc an algorithm for classification with Phrase extraction where Phrases are important in extracting sensitive and useful information which is important for sentiment classification. Phrases can express sentiment information more expeditiously than individual words. Here, using Part of-speech (POS) based rules and dependency relations in the tweets sentiment rich phrases are extracted to convey contextual and syntactic information. Then the tweets are classified using LinearSvc and phrase patterns into positive, negative and neutral tweets. The accuracy of sentiment classification is improved by using these methods. So that one can gather information related to a topic efficiently if the sentiment is known i.e. either positive or negative.**

Keywords*— Sentiment analysis, Preprocessing, LinearSvc, Phrase Patterns, Pos based rules, Dependency Relations.

I. INTRODUCTION

Sentiment Analysis or Opinion Mining is a Natural Language Processing and Information Extraction Task that identifies the user's views or opinions expressed in the form of positive, negative or neutral comments and quotes underlying the text. Sentiment Analysis on social networks are used for different purposes, such as election prediction, marketing, communication, business, medical, education and movie reviews[6]. In modern years, public opinion shows the diversity and multiple characteristics. And as the fast development of Internet, Social media as a new form of information transfer is becoming a significant channel for expressing public sentiments, feelings, pooling public wisdom in people's social life. At present, an important and large amount of information is being carried through web-based social media.

To disseminate ideas on hot topics and emergencies of education, many people using internet are willing to go through social media. Main platforms for expressing views are twitter, Facebook, BlogSpot and news channels. The views provided are unobstructed. The opinion provided in internet is very fast. A hot incident with comments will light the fuse of public opinion. Variety of ideas and views are setting up their significance and influencing through wide variety of ways. At present china is evolving in social transformation, IEPS influence is becoming high. At this time government should recognize the negative public opinion properly, otherwise it will form a larger community, public safety threat.

Especially, Twitter came forth to be an important social media providing most recent information and opinions on current issues and topics of any kind. This led security-related organizations to monitor twitter to gather information of the people. The twitter streams are very useful so there is a continuous interest on them.

In predicting elections if a party got more number of positive tweets it will be anticipated that it may win and the party with more number of negative tweets will lose. Marketing has wide variety of uses by using sentiment analysis for example product can be found whether it is successful or not based on the views of the people and those are rated accordingly. We can find degree of concern of a disease spreading contagiously [12]. Online education whether it is useful or not can be known by using this analysis.

A. Research Challenges

Sentiment analysis of Financial news using feature sets and classifiers for predicting the stock market for example Norwegian financial news articles are classified into four categories by using machine learning techniques with 71% accuracy. Here it has to find the stock market ups and downs by using news articles.

Border security related issues are also major topics to be taken care of [1]. What is happening in the borders and why they are crossing the border to gather such type of information. We should analyze the data. Illegal Bangladeshi immigrants are entering into India, to find the sentiment of the Indians towards them classify the tweets into positive, negative and neutral.

II. RELATED WORK

Twitter is a relatively ancient source of information, research to use Twitter in the context of awareness situations, crisis management and security-related intelligence gathering has received much attention. Movie reviews and product reviews are easily classified using sentiment analysis. Because the data in review sentences is grammatically correct than the tweets. On the other hand tweets are informal, short, colloquial and can contain abbreviations and slangs. Thus many nlp tools fail to work with tweets such as tokenizer, dependency parser etc. Current sentiment analysis approach follow a similar machine learning approach like text classification.

For example, Earthquake shakes Twitter Users: Real-time Event Detection by Social Sensors presents an SVM-based approach to classify tweets on earthquakes and typhoons and a probabilistic spatio-temporal model to find the centre and the trajectory of the earthquake events, whereas Natural Language Processing to the Rescue? Extracting" Situational Awareness reports on an approach of using linguistic features to detect tweets with content relevant to situational awareness during mass emergencies [3]. Sentiment analysis in red river floods studies the content of tweets posted during Red River floods and Oklahoma Grassfires in order to identify event features which can be automatically extracted. Utilization of Twitter for increasing situational awareness in the context of earthquakes is also reported etc are there. Most of the reported work in the context of using Twitter.Maintaining the Integrity of the Specifications.

III. PROPOSED SYSTEM

In this research we try to classify tweets on frontier security related issue for example illegal immigration from Bangladesh to India by using sentiment rich phrase patterns and Linear support vector machine learning classification.

A. Tweet extraction

By using twitter API and R language tweets on particular topic are extracted. Here in Twitter an application is created and the tweets are gathered on specified search term i.e. keyword search after authentication. For example keyword illegal migration is used then tweets containing that word are retrieved. Tweets are gathered from particular location using geo tagged search. Mobile devices which are having geo tagging feature active can share its location. Roughly 20% of the tweets are geo located using the location search. Location query requires longitude and latitude values of the place and the radius of the region. For example "immigration and geocode: 72, 10, 500km" will return tweets containing immigration keyword and in the range of 500km from the centre 72,10 coordinates.

B. Tweet Preprocessing

The collected tweets are preprocessed to get the text part only i.e. words only to do that we remove punctuations, urls, special symbols, lowercasing, emotion replacement and word normalization etc.

Remove punctuations: The symbols like comma, full stop, at the rate of, colon and semicolon etc. are eliminated.
Remove urls: The tweets are obtained with the username and urls those are removed from the tweets.
Emotion replacement: Tweets are compared with a emotion list like senti strength and replace emotions with their polarity.
Word normalization: The tokens from tweets are compared to words in Roget's Thesaurus. If not matches with any word, repeated letters are reduced to two or one until a match is found in the dictionary (e.g. "excellllllent" becomes "excellllent", "excelllent", and subsequently "excellent"). Words in this form are treated as "stressed" words.
User and topic labelling: The topics mentioned refers to with # in the tweet are replaced by topic name, and @ symbol by person.
Affect word matching: The words in tweet (tokens) are matched against three dissimilar sentiment lexicons: MicroWNOp, Linguistic Inquiry and Word Count (LIWC) [8], and General Inquirer (GI) [7] which were organised into four categories high positive, positive, high negative and negative. Matched words are denoted with their label of sentiment. The version of data without replacements is also maintained to compare with results.
Bag of words and Rule based Algorithm: Tweets contain special characteristics in terms of style those are not in traditional way they contain emotions :). They will go in wrong direction if we do not carefully analyse them.

C. Feature extraction

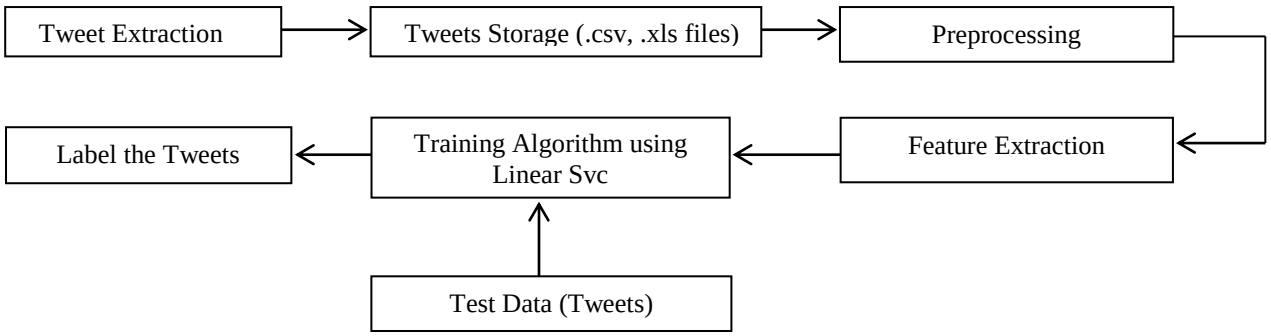
A classifier whether it is successful or not can be determined by its feature vector. A good feature vector will classify accurately. So in implementing a classifier, feature vector plays a key role. The training to classify is learnt from the feature vector and builds a model or algorithm that classifies the unknown data [13].

A feature can be represented by its absence or presence of words in the tweet. The data to be trained will consists of positive, neutral and negative tweets. Split each tweet into words then add each word to feature vector. This type of adding single words to feature vector is called 'unigrams' approach. We will filter out the words which don't have any sentiment like words is, was , then that etc.

D. Classification

This Sentiment detection module retrieves the tweets from social media using keywords search, then it will store the tweets in the form of .csv and .xls files. This module is trained by collecting the features that is words it may be unigrams or bigrams or n-grams. The polarity of words taken into account to calculate the polarity of tweet. This module classifies each tweet into one of the three labels that are positive, negative and neutral. This is domain independent. Language of the text is analysed which reflects the mood and attitude of public using twitter rather than the negativeness or positiveness of the fact. The whole process of classification is shown in architecture.
This sentiment analysis system is based on a mixed approach, which incorporates supervised learning with a Support Vector Machines linear kernel, on bigram and unigram features; Even so we generalize some of the n-gram features of training data. By using abstract labels related to the word polarity (e.g. "NEGATIVE", "POSITIVE"), as found in dictionaries of sentiment, lists of emotion, slang lists and other media-specific features. We do not use any particular language analysis software. Same approach is used to apply for any language by using alike language-specific dictionaries

Architecture:



There are many variations of support vector machines. LIBSVM is a variation of svm where it will implement large-scale regularized linear classification and regression. LiblineaR algorithm features include cross validation for model selection, multi-class classification (one-vs-the rest, and Crammer & Singer method), weights for unbalanced data or probability estimates (logistic regression only). The estimation of the models is as speed as compared to other libraries. Tweets are classified using supervised learning technique Support Vector Machines Sequential Minimal Optimization (SVM SMO) with a linear kernel, based on boolean features. This sentiment analysis system analyses proposed tweets, which employs supervised learning with a Support Vector Machines Sequential Minimal Optimization (SVM SMO) linear kernel, on unigram and bigram features; however some of the n-gram features are generalized in the training data using abstract labels corresponding to the word polarity (e.g.”POSITIVE”, ”NEGATIVE”), as found in sentiment dictionaries, emoticon lists, slang lists and other social media-specific features. SMO algorithm did not scale to large number of samples as compared to LinearSVC. Furthermore, LinearSVC uses one vs the rest scheme whereas svm smo multi-class mode is implemented using one vs one scheme. LinearSVC has more flexibility in the choice of penalties and loss functions and should scale better. Now, LinearSvc [4] is a machine learning algorithm which is used for classification. It is a linear kernel algorithm which takes the input tweets and trains the algorithm to classify based on polarity of the words. It classifies the tweets using a hyper plane equation

$$w.x + b = 0$$

x is the variable holding tweets i=1,2,…n

w is the normal vector drawn to hyper plane.

$x_i.w + b = +1$ for $y_i = +1$

$x_i.w + b = -1$ for $y_i = -1$

x_i, y_i are the input tweets and the class label respectively.

$$D=\left\{(x_i,\ y_i)\middle| x_i \in R^p,\ y_i \in \{-1,1\}\right\}_{i=1}^{n}$$

Normalization of w and classification is done by using Lagrange multipliers.

Phrases are extracted from tweets and semantic orientation of each and every phrase is computed [10]. Phrase is assumed to be semantically positive or negative if the phrase occurs predominantly and frequently in one class (positive or negative). If a phrase has high positive or negative polarity value that indicates that phrase has appeared more with positive sentences known as positive words. The association strength between a phrase and negative and positive sentences is calculated by using Point Wise Mutual information (PMI).

$$PMI(c, pos) = \log \frac{p(c, pos)}{p(c)p(pos)}$$

$$PMI(c, neg) = \log \frac{p(c, neg)}{p(c)p(neg)}$$

Probability of a phrase that it occurs in positive documents =P(c, pos), i.e. occurrence of a phrase in positive documents divided by total number of documents that are positive.

Probability that a phrase occurs in negative document = P(c, neg) i.e. occurrence of a phrase divided by total number of documents that are negative.

PMI value difference is the polarity of a phrase [11].

$$SO(c) = PMI(c, pos) - PMI(c, neg)$$

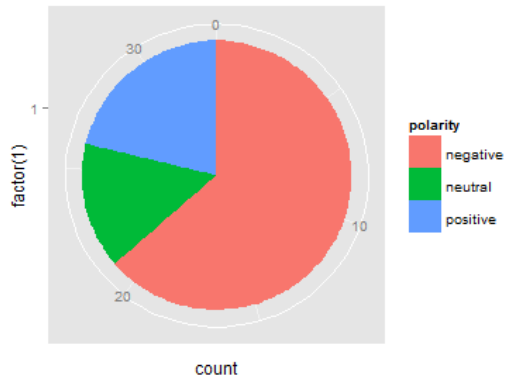
$$SO(c) = \log \frac{p(c, pos)/p(pos)}{p(c, neg)/p(neg)}$$

$$SO(c) = \log \frac{P(c, pos)}{P(c, neg)}$$

Calculate the polarity of each unigram and bigrams (phrases) by using PMI methods and using lists provided by LIWC and Sentiwordnet. Train the algorithm using Linear Svc to classify the tweets. Test the algorithm by providing the list of tweets. Compare the result with manually classified tweets.

IV. ANALYSIS

Here issue is illegal Bangladeshis migration into India, for sample we retrieved 37 tweets from twitter API. Those are preprocessed by using R language like elimination of punctuation, urls, special symbols and word normalization. By using Linear Svc and phrase extraction tweets are classified into positive, negative and neutral. Among those most of the tweets are classified as negative so there are more number of people are opposing the migration. Here we got more accuracy in classification by using linear Svc with phrase extraction than using simple svm. By using svm we got 85% accuracy but by using svc we got 90% accuracy.



(i).figure representing no. of negative, neutral and positive tweets

V. CONCLUSION & FUTURE WORK

We study the usefulness of the sentiment analysis as a technique for observing interesting data and facts on Twitter. Gathered the tweets and pre processed by filtering them and classified using different algorithms this classification is useful in determining the intentions of the people. So one can take action corresponding to the views of public. Twitter is used as a sensor for gathering the sentiment of the people by classifying the tweets. In the future work Consider long term of tweets so that we can calculate the real value of tweets otherwise there may be false tweets which leads to misconception [5] [9]. We can classify the tweets efficiently by considering the non english tweets in various languages.

REFERENCES

1. Jakub Piskorski, Hristo Tanev, Alexandra Balahur Exploiting Twitter for Border Security-Related Intelligence Gathering, 2013 European Intelligence and Security Informatics Conference.
2. J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines." Advances in Kernel Methods - Support Vector Learning, Tech. Rep., 1998.
3. M. Atkinson, J. Piskorski, R. Yangarber, and E. van der Goot, "Multilingual Real-Time Event Extraction for Border Security Intelligence Gathering." in *Open Source Intelligence and Counter-terrorism*, U. K. Wiil, Ed. Springer, LNCS, Vol. 2, 2011
4. http://scikitlearn.org/stable/modules/generated/sklearn Svm.LinearSVC.html#sklearn.svm.LinearSVC and http://scikit-lear.org/stable/modules /svm.html
5. C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility onTwitter." in *Proceedings of the 20th International Conference on WorldWide Web*. New York, USA: ACM, 2011, pp. 675–684.
6. Thet TT, Na JC., Khoo CSG., Shakthikumar S., Sentiment Analysis of Movie Reviews on Discussion Boards using a Linguistic Approach, In Proceedings of the 1st international CIKM workshop on Topic sentiment analysis for mass opinion, Pages 81-84, 2009.
7. P. Stone, D. Dunphy, M. Smith, and D. Ogilvie, "The General Inquirer: A Computer Approach to Content Analysis." 1966.
8. Y. Tausczik and J. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.
9. C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter." in *Proceedings of the 20th International Conference on World Wide Web*. New York, USA: ACM, 2011, pp. 675–684.
10. Basant Agarwal, Vijay Kumar Sharma, and Namita MittalSentiment Classification of Review Documents using Phrase Patterns, *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI).*
11. Turney PD., Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, ACL, pp.417-424, 2002.
12. S. Burton, K.Tanner, C. Giraud-Carrier, J.West, and M.Barnes, "Right Time, Right Place Health Communication on Twitter: Value and Accuracy of Location Information." *Journal of Medical Internet Research*, vol. 6, no. 14, 2012.
13. Ravikiran janardhana, how to build a twitter sentiment For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].