Volume 5, Issue 11, November 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Twitter Data Analysis Techniques to Extract Public Opinion

Judith Sherin Tilsha S[*], Shobha M S

Dept. of Information Science & Engineering, New Horizon College of Engineering,
Visvesvaraya Technological University, Bangalore, India

Abstract— This paper aims to provide a survey on various techniques used to mine emotions or opinions of people. Sentimental analysis has gained a considerable amount of attention from researchers since it plays a vital role in people life. The information shared in twitter has been affecting markets, political events etc. and the number of tweets per day is extremely large. So extracting public opinion from tweets is required. Most the twitter data analysis techniques are dictionary based where the words are extracted from tweets and compared against a dictionary. But these techniques lack in domain or context based semantics at the same time the strength of the techniques depends on the strength of dictionary. Later the approaches using machine learning algorithms were proposed in which a feature vector is constructed with the emotion describing words from tweets and are fed to the classifier that classifies the sentiment or opinion. This article discusses various twitter data analysis techniques that are based on dictionary and that are using the machine learning approaches.

Keywords— Word Sense Disambiguation (WSD), Hashtag, Ensemble Classifiers, Feature Vector, Sentimental Analysis

I. INTRODUCTION

The rise of social media in couple of years has changed the general perspective of networking, socialization, and personalization. Use of data from social networks for different purposes, such as election prediction, sentimental analysis, marketing, communication, business, and education, is increasing day by day. Precise extraction of valuable information from short text messages posted on social media (Twitter) is a collaborative task.

Among different social networks, Twitter is one of the most popular micro blogging services. Over 465 million twitter accounts in 2012 have generated 175 million tweets per day. People are increasingly using Twitter to share advice, opinions, news, moods, concerns, facts, and rumours.

Sentiment analysis influences users to classify whether the information about the product is satisfactory or not before they acquire it. Marketers and firms use this analysis to understand about their products or services in such a way that it can be offered as per the user's needs. The next section elaborately discusses a variety of sentimental analysis techniques.

II. LITERATURE SURVEY

A. Modeling Public Mood And Emotion: Twitter Sentiment And Socioeconomic Phenomena.[9]

In 2010, Johan Bollen, et al., found that there is a good correlation between public opinion and social, cultural, political and economics. They tried to analyze the emotions and public opinion using Profile of Mood status an established psychometric instrument and experimented with the tweets registered on US President Election [8]. Their method starts with listing out all events including fluctuations in market, death of a film or music icon, plane crash, political events etc. in timeline and collected tweets of a particular event over specific time period. Then they filtered out the tweets which doesn't express people emotions, such as those starts with 'http:' or 'www'. A normalized tweet will result in an ordered list of terms filtered for stop-words and non-alphanumeric characters, converted to lower-case, and Porter-stemmed. Which is then POMS scored.

The POMS is not intended for large-scale textual analysis. Rather, it is a psychometric questionnaire composed of 65 base terms. the POMS-scoring function P(t) maps each tweet to a six-dimensional mood vector where each dimensions of mood are, Tension, Depression, Anger, Vigour, Fatigue, and Confusion. The mood vector for a particular tweet is then converted into an unit mood vector (m). The mood vectors of a set of tweets submitted on a day will be averaged to produce an aggregate mood vector (m_d).

But the probability that the terms extracted from the tweets submitted on any given day match the given number of POMS adjectives N_p thus varies day to day. So all mood values for a given day *i* are converted into to z-scores so that they would be normalized with respect to a local mean and standard deviation observed within the period. So the normalized Z-score on a 6-dimensional time series that fluctuates around a mean of zero on a scale of 1 standard deviation is used to highlight short-term fluctuations of public mood as a result of particular short-term event.

They also adopted a 6-dimensional time series whose variance has been normalized to a scale of 1 standard deviation. This is used to assess changing mood levels over time in relation to long-term changes in socio-economic indicators. Thus they found that social, political, cultural and economic events are correlated temporary fluctuations of public mood levels and also they stress the importance of measuring mood and emotion using well-established instruments that undergone decades of empirical psychometric research.

B. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys [7]

In this paper a supervised sentiment classification framework was proposed that provides a way to utilize tagged Twitter data and smileys for classification of a wide variety of sentiment types from text. By utilizing 50 Twitter tags and 15 smileys as sentiment labels, this framework avoids the need for labor intensive manual annotation, allowing identification and classification of diverse sentiment types of short texts. The contribution of different feature types for sentiment classification was evaluated and was shown that the framework successfully identified sentiment types of untagged tweets. The quality of the sentiment identification was also confirmed by human judges. It also explores dependencies and overlap between different sentiment types represented by smileys and Twitter hashtags.

Four basic feature types for sentiment classification were used: single word features, n-gram features, pattern features and punctuation features. All feature types are combined into a single feature vector.

Hashtag-based sentiment labels five different categories: 1 – strong sentiment, 2 – most likely sentiment, 3 – context dependent sentiment, 4 – focused sentiment, and 5 – no sentiment.

The Amazon Mechanical Turk (AMT) service was used to obtain a list of the most commonly used and unambiguous ASCII smileys. From the obtained list of smileys we selected a subset of 15 smileys which were (1) provided by at least three human subjects, (2) described by at least two human subjects using the same single-word description, and (3) appear at least 1000 times in our Twitter dataset.

After constructing a feature vector the authors used k-nearest neighbors (kNN) strategy for classification. For each feature vector v in the test set, the Euclidean distance to each of the matching vectors in the training set is computed. Then the mean of all distances calculated for the matching vectors are calculated and the outliers that has distance twice than the mean are eliminated. The label assigned to v is the label of the majority of the remaining vectors.

A diverse set of feature types used for sentiment extraction including punctuation, patterns, words and n-grams contributes to the sentiment classification framework. Two different methods which allow an automatic identification of sentiment type overlap and inter-dependencies. While hashtag labels are specific to twitter data, the obtained feature vectors are not heavily Twitter-specific.

C. Opinion Mining and Sentiment Analysis on a Twitter Data Stream [3].

In this paper the performance of different classifiers in extracting sentiments from tweets are gauged and compared. Before classification starts the extracted raw tweets are pre-processed to reduce noisy data and to select tweets which are rich in emotions. Then the pre-processed data is fed into a two stage classification. The first stage classifies the incoming data into anyone of the three categories: neutral, polar and irrelevant. Then the next stage is fed with only the data classified under polar and it divides the data into positive and negative. Here the authors tried achieve more accuracy by processing only polar data and by filtering neutral and irrelevant data at the early stage.

To identify the better performing classifier the used the machine learning tool Weka which has a number of built-in classifiers? The Weka tool is provided with the pre-processed data and the authors measured the percentage of accuracy obtained for each classifier for both levels of classification. They observed that the SMO, SVM and Random Forest classifiers performs better than Naïve bayes classifier.

D. End-to-End Sentiment Analysis of Twitter Data [2]

In this paper, an end-to-end pipeline for sentiment analysis of a popular micro-blogging website called Twitter was presented. A hierarchal cascaded pipeline of three models to label a tweet as one of Objective, Neutral, Positive, Negative classes is built. The performance of this hierarchal pipeline is compared with that of a 4-way classification scheme. The trade-off between making a prediction on lesser number of tweets versus F1-measure is done.

A 4-way classifier is used to build a cascaded design, stacking 3 classifiers on top of each other: Objective versus Subjective, Polar versus Non-polar and Positive versus Negative. If the confidence of the classifier falls below a threshold, then reserve prediction on that example. This will boost the F1-measure. The relation is presented graphically and shows that one of the cascaded designs is significantly better than the other designs.

The pipeline for end-to-end classification of tweets into one of four categories is simple: 1) crawl the tweets from the web, 2) pre-process and normalize the tweets, 3) extract features and finally 4) build classifiers that classify the tweets into one of four categories: Objective, Neutral, Positive, Negative.

Following is a list of possible classifier designs:

1). Build a 4-way classifier. Note, in a 4-way classification scheme, a multi-class one-versus-all SVM builds 4 models, one for identifying each class. Each model is built by treating one class as positive and the remaining three classes as negative.

2). Build a hierarchy of 3 cascaded models: Objective versus Subjective, Polar versus Non-Polar and Positive versus Negative. But there is one design decision to be taken here: while building the Polar versus Non-polar model, do we want to treat both Neutral and Objective examples as Non-polar and only Neutral examples as Non-polar? This decision affects the way we create the Polar versus Non-polar model.

E. Precise Tweet Classification and Sentiment Analysis [12]

The paper by Rabia Batool et al proposed to enhance preciseness of sentiment or information extracted from tweets related to health care domain [5]. The system proposed by the authors is shown in Fig 1.

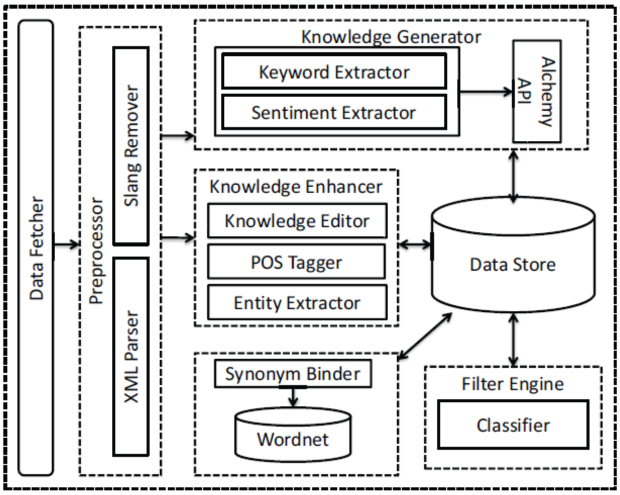

Fig 1: System architecture proposed for tweet classification by Rabia Batool *et al.*[12]

To extract tweets, the Archivist, a service that uses Twitter Search API to find and archive tweets was used. Tweets returned by Archivist are in XML format that need to be processed by DOM parser and are stored in a repository. For extraction of keywords, entities and sentiments Alchemy API[1] was used that utilizes natural language processing technology and machine learning algorithms to analyse content and extract key phrases, named entity, and topic level sentiments. Table I shows keywords and associated sentiments, extracted by knowledge generator from tweets

Table I: Knowledge Extracted by Knowledge generator [12]

Tweets	Keywords	Sentiments
Exercise is very good for diabetic patient	Exercise Diabetic patient	Positive
I am Scott Malkinson and I have got diabetes please help	Scott Malkinson	Neutral

Twitter data also contains slangs and repeated character like *plz* and *gooood* instead of *please* and *good* and spelling mistake which could affect knowledge extraction process. So he information extracted by knowledge generator is not accurate. In order to enhance the accuracy of retrieved knowledge the knowledge enhancer module incorporates the addition of subjects, verbs, objects, and entities in knowledge as shown in Table II.

Table II: Information Extracted knowledge Generator and Knowledge Enhancer [12]

Tweet	Knowledge extracted by knowledge generator	Knowledge extracted by knowledge enhancer
I am Scott Malkinson and I have got diabetes plz help	Scott Malkinson	Scott Malkinson, diabetes
RT @qytaralore: physiology Viagra online cialis commercial ads http://t.co/cPMJqQ6w impotence in young men diabetes	Physiology Viagra online commercial ads, young men, impotence	Viagra, diabetes

Then the synonym binder connects synonyms with words and stores them into data store. It also covers many word structure problems associated with words e.g., it extracts synonym of calories as calorie and exercises as exercise. WordNet dictionary is used to bind synonyms with entity and keywords.

Then the filter engine classifies data based on seed list which is useful to classify without explicit keywords. Thus the authors found that when data searched from Twitter using keywords, it does not return all related data which can provide useful information. Using short text classification, all information related to specific topic are extracted. With the addition of verbs and entities the authors achieved preciseness in sentiment division also.

F. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis [8]

Geetika Gautam et.al contributed to the sentiment analysis for customers' review classification from unstructured tweets using machine learning classification algorithms such as Naïve bayes classifier, maximum entropy and Support Vector Machines (SVM). They also adopted the semantic orientation of tweets using WordNet. Fig 2 shows the system proposed by Geetika et al.

The authors used twitter data set which is already labelled with negative and positive polarity. This raw data is preprocessed in order to reduce the chances of having inconsistency and redundancy. Then the unigram model is used to extract features by segregating the adjectives from pre-processed tweets. Ex: From "Painting Beautiful" only *beautiful* is extracted.

The authors use supervised learning algorithms such as naïve bayes[4], maximum entropy[6] and SVM for training and classification. In order to improve the accuracy the semantic analysis of words is done using WordNet, a database where semantically similar words are linked together.

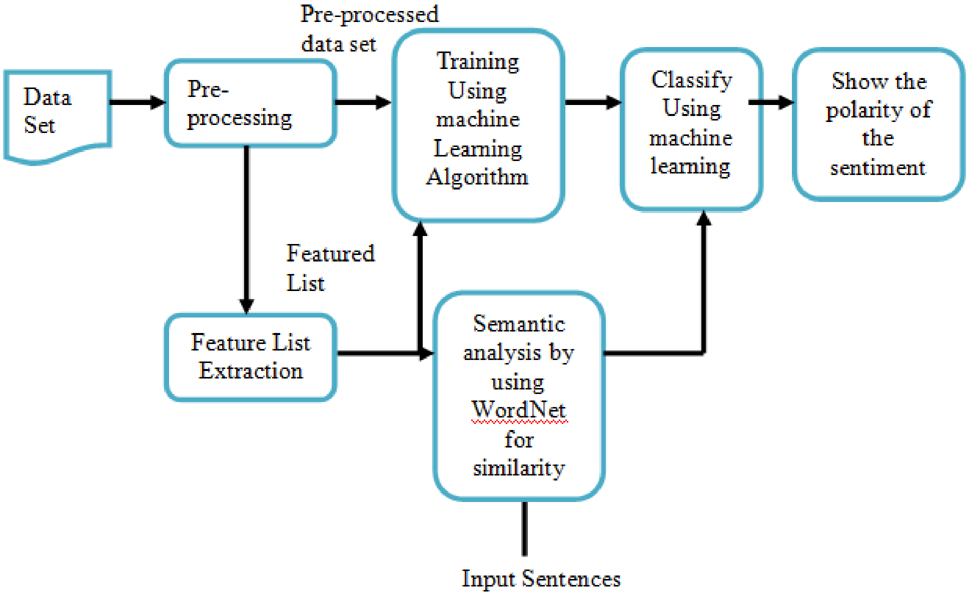
Pre-processed data set
Data Set
Pre-processing
Training Using machine Learning Algorithm
Classify Using machine learning
Show the polarity of the sentiment
Featured List
Feature List Extraction
Semantic analysis by using WordNet for similarity
Input Sentences

Fig 2: Module Diagram for Sentiment Analysis

The key task is to use the stored documents that contain terms and then check the similarity with the words that the user uses in their sentences. Thus it is helpful to show the polarity of the sentiment for the users. For example in the tweet "I am happy" the word "happy'' being an adjective gets selected and is compared with the stored feature vector for synonyms. Assume 2 words; 'glad' and 'satisfied' tend to be very similar to the word 'happy'. Now after the semantic analysis, 'glad' replaces 'happy' which gives a positive polarity.

The authors showed that the naïve bayes classifier gives better results than other two. They also suggested that the training data set can be increased to improve the feature vector related sentence identification process and also to extend WordNet for the summarization of the reviews.

G. NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers [10].

In this paper an enhanced way of sentiment classification is proposed by adding semantics in the feature vector computation and using ensemble methods for classification. This system constructed the feature vector by considering the relationship between the words in tweets by using WordNet's Synset. It also focused on resolving the ambiguity associated with polysemic words by using Word Sense Disambiguation (WSD) [11].

The constructed feature vector is given to various traditional machine learning classifiers such as Naïve Bayes, Maximum entropy and Support Vector Machines (SVM). The authors show improvement in efficiency of sentiment classification by using ensemble approach. The ensemble methods basically aim to combine the predictions of several traditional machine learning algorithms so that it can build a model that minimizes the biasness of a single algorithm.

The following parameters were used to estimate their system performance:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall}$$

They also tested their system with different number of tweets. With different scenarios such as key words only, keywords with Synset, keywords with WSD and Synsets the authors observed the effectiveness of feature vector. Finally they claimed that use of Synset and WSD overcome the limitations of traditional bag-of words methods and use of ensemble based classification showing improved performance by 3-5 % than earlier machine learning algorithms.

III. CONCLUSION

This paper discussed the insights of techniques that analyses twitter data used for public opinion mining. Some of the sentimental analysis techniques presented here are based on dictionary approach whereas the others used machine learning approaches. This survey gives us an observation that data analysis techniques based on machine learning classifiers performs better than dictionary based approaches. In order to achieve better classification performance the researchers used the ensemble classifiers that performs better than individual machine learning approaches. Apart from classifying sentiments from twitter data, the precision of classification also need to be improved and the same achieved in [12].

REFERENCES

[1]. *Alchemy API*, (Last visited in March 2012). [Online]. Available: www.alchemyapi.com

[2] Apoor v Agarwal, Jasneet Singh Sabharwal, *End-to-End Sentiment Analysis of Twitter Data*, Workshop on Information Extraction and Entity Analytics on Social Media Data, pages 39–44, COLING 2012, Mumbai, December 2012.

[3] Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, ThiruchittampalamRagavan ,Nadarajah Prasath, Shehan Perera, *Opinion Mining and Sentiment Analysis on a Twitter Data Stream*, 2012 IEEE.

[4]. B.Ren ,L.Cheng, *Research of Classification System based on Naïve Bayes and Meta Class*, Second International Conference on Information and Computing Science, ICIC 09, Vol(3), pp. 154 – 156, 2009.

[5] C. Fellbaum, WordNet. Wiley Online Library, 1998, *Healthcare tweet chats*, (Last visited in October 2012). [Online]. Available: http://www.symplur.com/healthcare-hashtags/tweet-chats/

[6]. C.I.Tsatsoulis, M.Hofmann, *Focusing on Maximum Entropy Classification of Lyrics by Tom Waits*, IEEE International on Advance Computing Conference (IACC), pp. 664 – 667, 2014.

[7] Dmitry Davidov, Oren Tsur, Ari Rappoport, *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*, Coling 2010: Poster Volume, pages 241–249, Beijing, August 2010

[8] Geetika Gautam, Divakar yadav, *Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis*, ©2014 IEEE

[9] Johan Bollen, Alberto Pepe Huina Mao, *Modeling public mood and emotion: Twitter sentiment and socioeconomic Phenomena*, April 2630, 2010, Raleigh, North Carolina

[10] Monisha Kanakaraj, Ram Mohana Reddy Guddeti, *NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers*, 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), c 2015 IEEE.

[11] M. Stevenson and Y. Wilks, *Word-sense disambiguation*, *The Oxford Handbook of Comp. Linguistics*, pp. 249–265, 2003.

[12] Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool and Sungyoung Lee, *Precise Tweet Classification and Sentiment Analysis*, ©2013 IEEE.