SURVEY ON EMOTION DETECTION IN SOCIAL MEDIA

Prof. SunandaBiradar[1], Sonali Porwal[2]

[1]Prof., [2]PG Scholar, Dept. of Computer Science Engineering,

BLDEA's College of Engineering and Technology, Karnataka (India)

ABSTRACT

Social Media such as Facebook and Twitter are used widely by all aged people across the World. A large repository of live real time information is thus provided by these media. Various organizations, government agencies, institutions and businesses are therefore looking for ways to monitor and analyze public response to various events, products, and services. A lot of work has been done for detecting emotions in the message and tweets. This survey paper contains survey on various technologies used for detecting emotions in the tweets or messages. These systems automatically collect messages for detecting emotions and give a summary of the messages or tweets.

Keywords: Classification, Microblogging, Naïve Bayes, Svm, Tweets

I. INTRODUCTION

With the growing use of social media such as Twitter and Facebook for information sharing, there is a need for automated tools and techniques to analyze and monitor messages or tweets. Emotion Detection requires measuring emotion and their expression within a wide range of informal messages. Detecting emotion from such a corpus is a difficult task because the messages are unstructured. Misspellings, slangs, contradictions and abbreviations are used commonly in tweets because the maximum length of tweets allowed is 140 characters. Companies and media organizations are seeking ways to find out what people feel and think about their product. Earlier emotions were detected and categorized only as positive negative or neutral. But with advancements in technology, more emotions such as joy, fear, anger, etc are also detected by the machine. Many techniques have been developed for this purpose and this survey paper contains the various technologies used for detecting emotions in Social Media.

II. LITERATURE SURVEY

Classification of text into positive, negative and neural i.e. sentiment analysis proved to be the first encouragement for emotional detection in Social media.

In their work "Opinion mining and sentiment analysis" authors Bo Pang and Lillian Lee (2008) have discussed major challenges in sentiment analysis and ways to overcome them. They have covered different techniques to perform sentiment classification and extraction on structure each dealing with different features and aspects of sentiment analysis [1].

In paper "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", authors Alexander Pak and Patrick Paroubek (2010) have collected the corpus using Twitter API which consists of text posts and formed a dataset of three classes: positive sentiments, negative sentiments, and a set of objective texts (no sentiments). They performed linguistic analysis of the collected corpus to understand the various factors that are useful for classification of the data. Using the corpus, they built a sentiment classifier, which is able to determine positive, negative and neutral sentiments for a document [2].

In the domain of six Ekman's emotion detection from tweets, Saif M. Mohammad (2012), in his work ,"#Emotional Tweets", has proposed a unique method for creating a training corpus from tweets using emotion word hash tags. He also extracted a word–emotion association lexicon from this Twitter corpus, and show that it leads to better results [3].

Uma Nagarsekaret. al in their paper, "Emotion Detection from "The SMS of the Internet" went beyond the basic sentiment classification (positive, negative and neutral) and target deeper emotion classification of Twitter data. They focused on emotion identification into Ekman's six basic emotions i.e. JOY, SURPRISE, ANGER, DISGUST, FEAR and SADNESS. They employed two diverse machine learning algorithms with three varied datasets and analyzed their outcomes. They also proved that equal distribution of emotions in training tweets results in better learning accuracies and hence better performance in the classification task[4].

Martin D. Sykoraet. al in their paper,"Emotive Ontology: Extracting fine-grained emotions from terse, informal messages", employed an ontology engineering approach to the problem of fine grained emotion detection in informal messages. It detects a range of eight high level emotions. It can also detect multi-word phrases such as 'blew-blank-blank-away' such as 'blew my mind away' and substrings such as prefixes or suffixes. The ontology into memory was developed in Python and eight basic emotions are represented as a set of Hashtable and a trie data structure. Currently the ontology covers eight emotions which are anger, confusion, disgust, fear, happiness, sadness, shame and surprise[5].

Alexander Paket. al in their paper, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", presented a method to collect a corpus with positive and negative sentiments, and a corpus of objective texts. Their method allows to collect negative and positive sentiments such that no human effort is needed for classifying the documents. Objective texts are also collected automatically. The size of the collected corpora can be arbitrarily large. They also performed statistical linguistic analysis of the collected corpus. They used the collected corpora to build a sentiment classification system for microblogging. They conducted experimental evaluations on a set of real microblogging posts to prove that their presented technique is efficient and performs better than previously proposed methods[2].

Aditya Joshi et. al in their paper, "C-Feel-It: A Sentiment Analyzer for Micro-blogs", described a web based system, C-Feel-It, categorizes tweets as positive, negative or objective and gives an aggregate sentiment score that represents a sentiment snapshot for a search string. It predicts sentiment in micro-blogs on Twitter. The weighted majority voting principle is used to predict sentiment of a tweet[6].

EfthymiosKouloumpiset. al in their paper, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", proposed a system that evaluates the usefulness of existing lexical resources and captures information about the informal and creative language used in microblogging.It uses three different copora of Twitter messages in experiment. Hashtagged data sets are used for development and training. Emoticon data set from http://twittersentiment.com . Manually annotated data set produced by iSieveCorporation(iSIEVE) are used for evaluation[7].

III. METHODOLOGY

Uma Nagarsekar et.[4] Al in their system, proposed a methodology that is divided into different stages as shown in Figure 1.

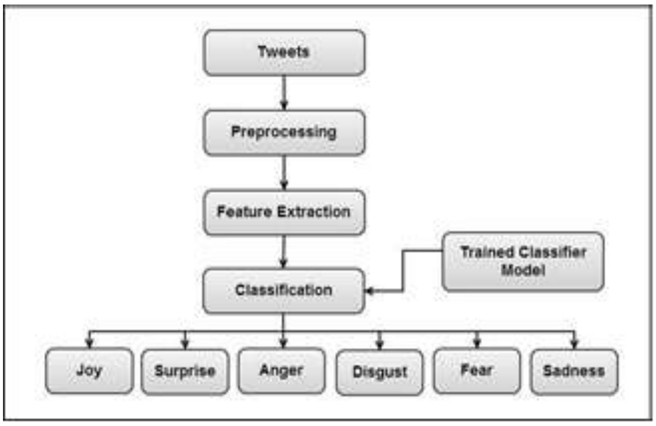

Fig 1. Emotion Analysis of Proposed Methodology [4]

The five stages are as follows:

1. Collection of tweets: The input to the emotion analyser is a user entered keyword based on which recent tweets over the last seven days are fetched from Twitter using its Search API.
2. Pre-processing: The procedure for pre-processing consists of the following five steps:
   a) Converting all the tweets collected to the lower case
   b) Removing the URL's present in the tweets and replacing it with URL
   c) Replacing any usernames present in the tweets to AT_USER
   d) Converting the hash tags to normal words i.e. removing the #
   e) Removing any unnecessary characters, extra spaces etc.
3. Feature Extraction : The input is the pre-processed tweet which is first filtered using the four steps mentioned below:
   a) Removing all stop words like a, the, is, etc. which don't indicate any emotion
   b) Replacing two or more repeating letters in a tweet by two letters of the same
   c) Removing punctuation like commas, single/double quotes, question marks, etc.

d) Discarding all words which don't start with an alphabet

4. Classification: A classifier is a learning model with associated learning algorithms that analyse data and recognize patterns which can be used for classification. They used two supervised classifiers: Naïve Bayes[8] and Support Vector Machines[9] that model the probability of an input being in a particular class by predicting the categorical emotion labels (joy, surprise, anger, disgust, sadness, fear). The authors have used SVM with a linear kernel [10].
5. Visualization of labeled tweets: A web interface is used for clear and concise visualization of the labelled tweets. It acts as a means for the user to interact with the system. The results of the classification are displayed in the form of a stream of the original tweets and the associated class label.[4]

Martin D. Sykoraet. Al[5] in their paper proposed the following system:

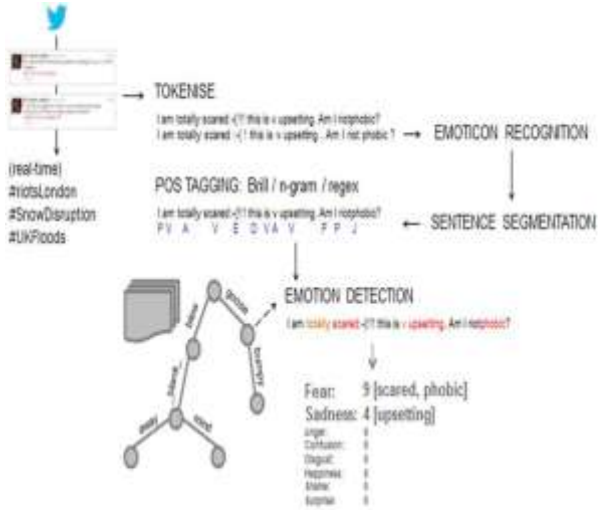

Fig. 2 Emotion Detection Proposed Methodology [2]

In this pipeline, messages were initially segmented into sentences based on punctuation. They used a tweaked version of Potts regex based tokeniser, which covers a wide range of emoticons. It also breaks up certain tokens on their suffices or prefixes for instance lipophobic would become lipo+phobic. This enables for efficient and fast matching of strings using Hashtables. A trie facilitates an efficient search for the longest possible phrase. Once an emotion was found, intensifiers, negators and conjunctions that occur before the emotion token would be matched too, if they satisfy some constraint.

The next step is Parts-Of-Speech tagging. Finally, as emotions, intensifiers, negators, conjunctions are matched, a running total of the emotional charge score for each found emotion is kept. Then, overall emotionality strength score will be computed by taking sum, average and maximum for each tweet.In future, the authors further intend to extend the ontology[5]

Using Twitter API the authors Alexander Paket. Al [2] collected a corpus of text posts and formed a dataset of three classes: positive sentiments, negative sentiments, and a set of objective texts (no sentiments). To collect negative and positive sentiments, they followed the same procedure as in (Read, 2005; Go et al., 2009). They queried Twitter for two types of emoticons:

• Happy emoticons: ":-)", ":)", "=)", ":D" etc.

• Sad emoticons: ":-(", ":(", "=(", ";(" etc.

The two types of collected corpora will be used to train a classifier to recognize positive and negative sentiments. They used the presence of an n-gram as a binary feature. The process of obtaining ngrams from a Twitter post is as follows:

1. Filtering – Remove URL links (e.g.://example.com), Twitter user names (e.g.@alex – with symbol @ indicating a user name),Twitter special words (such as "RT"6), and emoticons.
2. Tokenization – Segment text by splitting it by spaces and punctuation marks, and form a bag of words. However, it should be made sure that short forms such as "don't", "I'll", "she'd" remains as one word.
3. Removing stopwords – Remove articles ("a", "an","the") from the bag of words.
4. Constructing n-grams – Make a set of n-grams out of consecutive words. A negation (such as "no" and "not") is attached to a word which precedes it or follows it. For example, a sentence "I do not like fish" will form two bigrams: "I do+not", "do+not like", "not+like fish". Such a procedure allows to improve the

accuracy of the classification since the negation plays a special role in an opinion and sentiment expression(Wilson et al., 2005).[3]

Aditya Joshi et. al[6] proposed the following methodology:

Input: Search string

Output: It is a two level- tweet wise prediction and overall prediction. For tweet-wise prediction, sentiment prediction for each of the resources is returned. For overall prediction, it combines sentiment from all tweets to return the percentage of positive, negative and objective content retrieved for the search string.

It is divided into three parts:

1. Tweet fetcher: It obtains tweets related to a search string entered by the user. It uses live feeds from Twitter which retrieves atleast 50 tweets about the keyword in English. The result is in XML format.
2. Tweet Sentiment Predictor: It predicts sentiment for a single tweet. It has three fundamental blocks- Preprocessor, Emoticon-based Sentiment Predictor, Lexicon based sentiment predictor. Preprocessor deals with getting clean tweets. It handles extensions and contradictions. Extensions: Eg-'besssst' such extensions are normalized to their dictionary equivalent. It indicates a strong sentiment, so the extended word is replaced by two occurences of the contracted word. In Chat lingo normalization the words used in tweets are not present in lexical resources. A dictionary downloaded from http://chat.reichards.net is used to replace chat word by its dictionary equivalent. Emoticon based Sentiment Predictoris used for the visual representations of emotions frequently used on Internet. Emoticon mapping from http://chat.reichards.net/smiley.shtml is used.
3. Tweet Sentiment Collaborator**:** It gives overall prediction with respect to a keyword in the form of percentage of positive, negative and objective content. For each resource, the following scores are determine

posscore[r] =Xmi=1piwpi

negscore[r] =Xmi=1niwni

objscore[r] =Xmi=1oiwoi

where

posscore[r] = Positive score

negscore[r] = Negative score

objscore[r] = Objective score

m = Number of resources used for prediction

pi, ni, oi = Positive,negative& objective count of tweetpredicted respectively using resource i

wpi,wni, ooi = Weights for respective classes derivedfor each resource i[4]

EfthymiosKouloumpiset. al proposed the following methodology:

Hashtagged Data set- The Edinburgh Corpus contains 97 million tweets collected over a period of two months. The hashtagged data set is created by filtering duplicate tweets, non-English tweets, and tweets that do not contain hashtags. The final set of messages to be included in HASH dataset is selected by identifying tags that appear atleast 1000times in corpus. Then, the top hastags are marked as positive, negative and neutral tweets.

Emoticon data set- Go, Bhayani and Huang created an emoticon data by collecting tweets with positive ':)' and negative ': ('emoticons. The set contains 381381 positive and 150,570 negative.

iSeive data set- Contains 4,000 tweets. It contains data on selected topics, and label of each tweet reflects its sentiment towards the tweet's topic. Used for evaluation.

Preprocessing- It contains 3 steps-

1. Tokenization- Emoticons and abbreviations are identified.
2. Normalization- Abbreviations replaced by their actual meaning(eg-brbis be right back). All caps word are made into lower case and instances of repeated characters are replaced by a single character.
3. Parts-Of-Speech tagging is done[7].

IV. CONCLUSION

This survey paper contains various methodologies that have been employed for detecting emotions in tweets. Emotions are complex entities just like spontaneously occurring behaviours. Their expression varies from person to person and also changes over time. This makes them difficult to manipulate and measure experimentally without disrupting its expression. Hence it becomes increasingly difficult to identify them.

Larger datasets give better results for supervised training. Currently a large dataset consisting of emotion labelled tweets is unavailable. This created an urge to create our own custom dataset to proceed with the research. But the small size of the dataset is a shortcoming. Also, it cannot be assured that all the tweets have been correctly tagged in the training corpus. There is a possibility that incorrectly tagged entries may be present and such entries which affect the accuracy.

Furthermore, it is possible that a single tweet may represent multiple emotions or the same tweet can mean different emotions to different people. Consider the example of a Tweet tweeted on 20th April 2013, "@ WhoAhMai yeah fair enough Fear Sucks!" The above tweet represents joy, fear and disgust at the same time. It becomes almost impossible to classify such Tweets into a single class of emotion, affecting the accuracy.

Currently a major shortcoming of the dataset is that it does not contain a neutral label which means that the tweet is emotionless. Such a label is essential since it is not necessary that every tweet will portray an emotion. This problem persists since existing datasets do not identify such a label.

REFERENCES

[1] Bo Pang, Lillian Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, 2008, 1–135.

[2] Alexander Pak, Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, In Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10 , 2010.

[3] Saif M. Mohammad, #Emotional Tweets, In Proceedings of the First Joint Conference on Lexical and Computational Semantics (FJCLCS), 2012.

[4] Uma Nagarsekar, PriyankaKulkarni, AditiMhapsekar, Dr. Dhananjay R. Kalbande, Emotion Detection from "The SMS of the Internet, IEEE Recent Advances in Intelligent Computational Systems (RAICS),2013

[5] Martin D. Sykora, Thomas W. Jackson, Ann O'Brien; Suzanne Elayan,Emotive Ontology: Extracting fine-grained emotions from terse, informal messages,IADIS International Journal on Computer Science and Information Systems Vol. 8, No. 2,106-118 ISSN: 1646-3692.

[6] Aditya Joshi,Balamurali A R,Pushpak Bhattacharyya,RajatMohanty,C-Feel-It: A Sentiment Analyzer for Micro-blogs, Proceedings of the ACL-HLTPortland, Oregon, USA, Association for Computational Linguistics System Demonstrations,21 June 2011, 127–132.

[7] EfthymiosKouloumpis, TheresaWilson, Johanna Moore, Twitter Sentiment Analysis: “The Good the Bad and the OMG!”, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media

[8] Christopher Potts, Sentiment Symposium Tutorial: Classifiers, Internet: http://sentiment.christopherpotts.net / classifiers.html,[Mar. 20, 2013].

[9] Thomas Hill, PawełLewicki,Statistics methods and applications a comprehensive reference for science industry and data mining. StatSoft, Inc., 2006 , 319

[10] Alec Go, RichaBhayani, Lei Huang, Twitter Sentiment Classification using Distant Supervision, Stanford Digital Library Technologies Project, 2009