Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 5, May-2016

A Survey on Emotion Based Sentiment Analysis

Krunal Pavar[1], Lokesh Sahu[2]

[1]Computer Engineering Department, PIET Vadodara, kru.pawar@gmail.com
[2]Computer Engineering Department, PIET Vadodara, Lokesh.Sahu@paruluniversity.ac.in

Abstract *— Sentiment Analysis (SA) is a new and emerging field of research which deals with information extraction and knowledge discovery from text using Natural Language Processing and Data Mining technique, which help to track the mood of public about specific products and social or political event. Emotion Detection is one of the most emerging issues in human machine interaction. Detecting emotional state of a person from textual data is an active research field along with recognizing emotions from Twitter. We are mainly focusing on recognizing emotions from twitter's tweets. Proposed emotion recognize system takes input as a tweets and produces one of the six emotion classes (i.e. happy, sad, joy, anger, surprise, disgust) as the output. The aim is to get predictive analysis about user's behavior as well as public stance and attitude towards various events around the globe.*

Keywords- Sentiment Analysis Emotion, Emoticons, Feature Extraction, SVM, Twitter data.

I. INTRODUCTION

Today, huge amounts of informal subjective text statements are accessible online with the growing availability of social networking websites, blogging and micro blogging sites. These statements are represented in several formats such as news articles, comments and review.

Sentiment Analysis (SA) has recently become the focus of many researchers due to its application and various fields [9]. As it analyzes thought and ideas, feelings, attitude, and sentiment of individuals, analysis of this type of online text is helpful and demanded for marketing research, public opinion tracking, product auditing, business research, political surveys, client correspondence surveys, improving of web shopping bases, and so on.

Sentiment Analysis is the procedure, used for automatic extracting the polarity of public's subjective opinions from plain natural language text. Sentiment Analysis is likewise known as Opinion Mining (OM). Based upon opinion of others, one can make a good decision before acquiring any products or items. Sentiment Analysis has an extensive variety of use in e commerce, which serves to figure out answer of several questions like, What do users think about our product, Which of our clients are unsatisfied, What features of our product are the worst, Who and how impacts our image, What is the public response to some event or some individual.

Opinion can be collected from any individual in the world about anything through review sites, blogs, web forums and discussion groups etc [14]. Organizations and product owners who expect to improve their products/services may strongly benefit from the rich feedback of users or customers. The most generally utilized sources for finding opinion are Blogs, review sites, raw dataset, and Micro-blogging web sites [8].

Online messages that are posted by individual in World Wide Web are mostly informal. Analysis or handling of this kind of text is often more difficult if compared with formal texts. The main difference between formal and informal text is in data preprocessing is formal text often require less preprocessing whereas informal text often contains emoticons, sarcasm, utilization of weak grammar, and non lexicon-standard words[9].Therefore, extraction of informal content is regularly more troublesome.

People frequently ask their friends, relatives, and field specialists for suggestion during the decision-making procedure, and their opinions and perspectives are based on experiences and observations. One's point of view around subject can either be positive or negative, which is known as the polarity detection of the sentiment. During Sentiment Analysis process, it requires very fast and concise information so individual can make quick and accurate decisions.

II. BACKGROUND THEORY

2.1 Type of Sentiment Analysis

In Sentiment Analysis, the sentiment mainly classified into two types as described below [2]:

2.1.1 Regular and Comparative Sentiment

A regular sentiment presents a sentiment only on a specific entity or an aspect of the entity, e.g., "Mango tastes great" which communicates a positive sentiment on the aspect taste of Mango.
It is referred to regularly as a sentiment in the literature and it has two fundamental sub-types:

A) Direct Sentiment: A direct sentiment refers to a sentiment expressed specifically on an entity or an entity aspect, e.g., "The battery life is good."
B) Indirect Sentiment: An indirect sentiment refers to a sentiment expressed indirectly on an entity or aspect of an entity based on its effects on some other entities. This sub-type frequently happens in the medicinal area, e.g., "After infusion of the medication, my joints felt more regrettable" describes an undesirable effect of the medication on "my joints", which in a roundabout way gives a negative feeling or opinion to the medication. Much of the current research focuses on direct opinions. They are less difficult to handle. Indirect opinions are regularly harder to manage.

A comparative sentiment communicates a connection of contrasts between two or more entities and/or an inclination of the opinion holder based on some shared aspects of the entities. For example, the sentences, "Mango tastes great than Grapes" and "Mango tastes the best" express two comparative opinions [5].

2.2.2 Explicit and Implicit Sentiment

An explicit sentiment is a subjective statement that gives a regular or comparative sentiment, e.g., "Coke tastes great," and "Coke tastes better than Pepsi."
An implicit sentiment is an objective statement (usually expresses a desirable or undesirable fact) that implies a regular or comparative opinion, e.g., "India is at mars on his first attempt"

2.2 Task of Sentiment Analysis

The principle components of Sentiment Analysis issue are to identify the sentiment source, sentiment target, and the evaluative expressions or comments made by the opinion holder[3].
For the most part, an opinion is expressed by an individual person (opinion holder) who expresses a perspective (positive, negative, or neutral) about an entity (target object, e.g., person, item, association, occasion, service, etc.). A broad overview of the Sentiment Analysis issue [4] is presented in Figure 2.1.

2.2.1 Subjectivity and polarity classification

The task of analyzing if the sentence is objective sentence or subjective sentence is known as subjectivity classification. In which the sentences known as objective sentences, that express factual information about the world and sentence is known as subjective sentences that express some personal views, beliefs, feelings and opinions.
Sentiment polarity classification techniques are used to classify opinionated terms into positive and negative statements. A few works utilize weighting procedures to recognize the quality of subjectivity, i.e., weak positive and strong positive or weak negative and strong negative.
Sentiment lexical resources play a key part in recognizing and assessing statements of opinion. Opinion lexical resources comprise of a set of two types of words, i.e., positive polar words and negative polar words.

Positive= (great, pleasant, superb, positive, fortunate, right, superior)
Negative= (terrible, dreadful, poor, negative, unfortunate, wrong, inferior)

2.2.2 Sentiment target identification

The sentiment target identification refers to the target of the sentence, e.g., individual, item, association, occasion, service, etc. about which the sentiment is expressed. Sentiment targets at the sentence sentiment level or document sentiment level, the system should be able to have the capacity to recognize evaluative statements.

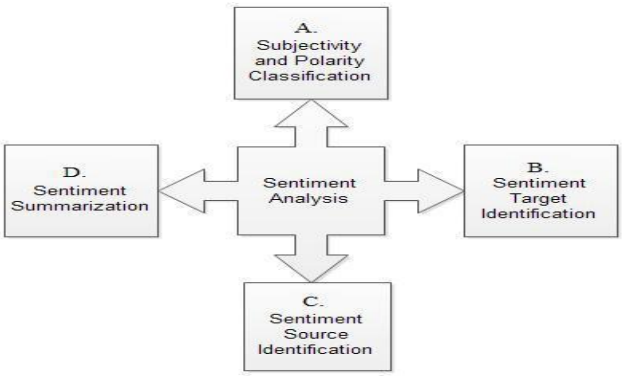

Figure 2.1: Task of Sentiment Analysis[3]

2.2.3 Sentiment source identification

An opinion holder or the source of an opinion is the individual person who expresses the opinion. For example, a statement has a great strength if it submitted by the expert rather the ordinary person. For instance, a doctor's opinion to health and medical treatment while general public opinion to a political party. This process of identifying the holder of the opinion is problem of a natural language processing.

2.2.4 Sentiment summarization

Finally at the last task, the sentiment summarization derive the extraction and categorization of entity, aspect, opinion holder, time and aspect sentiment.

2.3 Level of Sentiment Analysis

In Sentiment Analysis, the information or data collected from the reviews has been investigated mainly at three Sentiment Analysis levels [2]

2.3.1 Document Sentiment Level

The task at this level is to identify whether an entire sentiment document expresses a positive or negative sentiment. For example, given a product review, the system finds out whether the review expresses an overall positive or negative sentiment about any item or product. This task is usually known as document-level sentiment classification.

2.3.2 Sentence Sentiment Level

The task at this level goes to the sentences and figures out if each sentence expressed a positive, negative, or neutral sentiment. Neutral usually defines no opinion. This level of analysis is closely related to subjectivity classification, which recognizes sentences as objective sentences, that express factual information about the world and subjective sentences that express some personal views, beliefs and feelings. This task of classifying whether a sentence is subjective or objective is known as subjectivity classification.

2.3.3 Entity and Aspect/Feature Sentiment Level

Above described both the document sentiment level and the sentence sentiment level do not analyze what exactly people liked and did not like. Aspect level helps to derive polarity (positive or negative) and a target of sentiment. A sentiment without its target being recognized is of restricted use. Finding out the target of sentiment helps to understand the Sentiment Analysis problem better.

III. RELATED WORK

In [4], they contributes to the sentiment analysis for customers' review classification which is helpful to analyze the information in the form of the number of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of these two. For this they first pre-processed the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms namely: Naïve Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature. Finally they measured the performance of classifier in terms of recall, precision and accuracy.

The work in [5], they introduce a novel solution to SA of short informal texts with a main focus on Twitter posts known as "tweets". They compare state-of-the-art SA methods against a novel hybrid method. The hybrid method utilizes a Sentiment Lexicon to generate a new set of features to train a linear Support Vector Machine (SVM) classifier. They further illustrate that their hybrid method outperforms the state of- the-art unigram baseline.

The work in [6], To proposes a Tweets Sentiment Analysis Model (TSAM) that can spot the societal interest and general people's opinions in regard to a social event. In this paper, Australian federal election 2010 event was taken as an example for sentiment analysis experiments. They are primarily interested in the sentiment of the specific political candidates, i.e., two primary minister candidates - Julia Gillard and Tony Abbot. In this paper, work has demonstrated that building a lexicon-based sentiment analysis intelligent system is doable and can be very beneficial. Their experimental results demonstrate the effectiveness of the system.

In [7], they predict all emotion labels of each tweet. We use graphic emoticons, punctuation expressions together with a tiny but accurate lexicon to label data and provide a Multi-label Emotion Classification algorithm (MEC) for tweets in Weibo (so called Chinese Twitter). Our method has superior performance to the state-of-the-art method under both single-label and multi label evaluation measures. They also carried out a case study on Weibo dataset of Malaysia Missing Flight.

In [8], they propose a simple and completely automatic methodology for analyzing sentiment of users in Twitter. Firstly, they built a Twitter corpus by grouping tweets expressing positive and negative polarity through a completely automatic procedure by using only emoticons in tweets. Then, they have built a simple sentiment classifier where an actual stream of tweets from Twitter is processed and its content classified as positive, negative or neutral. The classification is made without the use of any pre-defined polarity lexicon.

The work in [9] they go beyond basic sentiment classification (positive, negative and neutral) and target deeper emotion classification of Twitter data. They have focused on emotion identification into Ekman's six basic emotions i.e. Joy, Surprise, Anger, Disgust, Fear and Sadness. They have employed two diverse machine learning algorithms with three varied datasets and analyzed their outcomes. They show how equal distribution of emotions in training tweets results in better learning accuracies and hence better performance in the classification task.

IV. PROPOSED METHODOLOGY

3.1 Overview of Proposed Work

In our system, we have proposed a methodology that is divided into different stages as shown in Figure 3.1. The five stages are as follows:

1) Collection of tweets
2) Pre-processing
3) Feature Extraction
4) Classification

3.2 Flow of the Proposed System

As demonstrated in the figure, methodology to extract the sentiment contains the several steps that are described below:

1. Collection of Tweets

The input to the emotion analyzer is a user entered keyword based on which recent tweets which fetched from Twitter using its Search API. The Twitter Search API is a dedicated API for running searches against the real-time index of recent tweets. Each request will return up to more than 100 tweets, for a single query. By running the search script we can keep up with most search topics without missing any tweets. For our system, We gathered our dataset by consulting the Twitter API and making use of word spotting based on occurrence of the word we are querying the recent tweets.

2. Pre-Processing

Twitter data is unstructured data. It needs to be processed before it can be used. Hence the tweets obtained are cleaned to remove unwanted discrepancies and retain only information that will help in determining the underlying emotion. This makes data easier to process in the later stages.

The procedure for pre-processing consists of the following steps:

1) Removing all non-English Tweets.
2) Converting all the tweets collected to the lower case.
3) Removing the URLs – erased all string that describes links or hyperlinks present in the tweets.
4) Replacing any usernames present in the tweets to @username – removed the username and because these are not considers for sentiments.
5) Converting the hash tags to normal words because hash tags can provide some helpful information, so it is useful to replace them with the literally same word without the hash. E.g. #Happy replaced with Happy.
6) Removing any unnecessary characters, extra spaces etc.
7) Remove all the number from tweets and also remove all words which don't start with an alphabet, for example 9th, 9:15am.
8) Removing punctuation like commas, single/double quotes, question marks, etc. at the beginning and end of each word in a tweet. E.g. Happy!!!!!! Replaced with Happy.
9) Replacing two or more repeating letters in a tweet by two letters of the same in tweets, sometimes users repeat letters to stress the emotion or feelings. E.g. Happpyy, Happyyyyyyyy for 'Happy'. Looked for 2 or more repetitive letters in words and replaced with 2 of the same.

3. Feature Extraction

Extraction of features is a very important concept that is responsible for the accuracy of the system. To decide what features are relevant to the classifier, we need a feature extractor. The one we have used returns a dictionary indicating what words are contained in the input passed. Here, the input is the pre-processed tweet which is first filtered using the steps mentioned below:

- Polarity Score of the Tweet
- Removing all stop words like a, the, is, etc. which don't indicate any emotion.
- Replace the emoticons with similar mining of word i.e. ☺ with happy.
- **Use of Negation Method** : The appearance of negative words may change the opinion orientation like not happy is equivalent to sad.

- **Use of Unigram Model** : The feature extraction method, extracts the aspect (adjective) from the dataset. Later this adjective is used to show the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using unigram model. Unigram model extracts the adjective and segregates it. It discards the preceding and successive word occurring with the adjective in the sentences. For above example, i.e. "Driving Happy" through unigram model, only Happy is extracted from the sentence.

Once the tweets are filtered, the output of the feature extractor is a list of the feature words present in the tweet.

4. Classification

A classifier is a learning model with associated learning algorithms that analyze data and recognize patterns which can be used for classification. We have used two supervised classifiers: Support Vector Machines and Naive Bayes that model the probability of an input being in a particular class by predicting the categorical emotion labels (Happy, Sad, Anger, Fear, Surprise, Disgust). Any one of the two classification methods may be used by the user to perform emotion analysis. Both of them have been trained with a pre-classified dataset and tested for accuracy. These trained classifier models, for every tweet, take the input as the feature word list extracted from the tweets and return the class of the tweet as one of the six universal Ekman emotion labels.

❖ Machine learning approach

Machine learning approach relies on the famous ML algorithms to solve the SA as a regular text classification problem that makes use of syntactic and/or linguistic features.

Text Classification Definition: We have a set of training records D = {X1, X2, . . ., Xn} where each record is labeled to a class. The classification model is related to the features in the underlying record to one of the class labels. Then for a given instance of unknown class, the model is used to predict a class label for it.

a) Support Vector Machines Classifiers (SVM)

The main principle of SVM is to determine linear separators in the search space which can best separate the different classes. In figure 3.2 there are 2 classes x, o and there are 3 Hyperplanes A, B and C. Hyperplane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation.

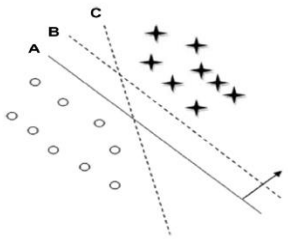

Figure 3.2 : Support vector machine on a classification[13]

Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories. SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyperplane .

b) Naive Bayes Classifier (NB)

The Naive Bayes classifier is the simplest and most commonly used classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.[13]

$$P(\text{label}\,|\,\text{features}) = \frac{P(\text{label}) * P(\text{fearures}\,|\,\text{label})}{P(\text{features})}$$

P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features | label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred. Given the Naive assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(\text{label}\,|\,\text{features}) = \frac{P(\text{label}) * P(f1\,|\,\text{label}) * \ldots\ldots * P(fn\,|\,\text{label})}{P(\text{features})}$$

V. CONCLUSION

The goal of this work is to classify the sentence according to its emotion. The proposed system involves the emotion classification. It contains emotion class happy, sad, anger, fear, surprise, disgust. The detection of the emotion features of the sentence is the most important step in emotion recognition. This process of extracting the text having emotion deals with finding the emotion feature set from the sentence. And also emotion recognize from the text entered by user on twitter with direct word or indirect emotions like emoticons or smiley faces. From Recognize of emotion we get the clear idea about the marketing research, public opinion tracking, product auditing, business research, political surveys, Events and so on.

REFERENCES

[1] Han & Kamber, "Data Mining: Concepts and Technique", 2nd edition, 2006.
[2] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012.
[3] Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan, Ashraf Ullah, "Mining opinion components from unstructured reviews: A review", Journal of King Saud University – Computer and Information Sciences, 2014.
[4] Alec Go, Richa Bhayani, Lei Huang, "Twitter Sentiment Classification using Distant Supervision", CS224N project report, Stanford, 2009, pp. 1-12.
[5] Lijun Shi, Jing Zhang, Xuegang Hu, "Subjective Relation Identification in Chinese Opinion Mining Based on Sentential Features and Ensemble Classifier", Computer Science and Information Technology, vol. 8, 2010.
[6] Blessy Selvam, S.Abirami, "A survey on Opinion Mining Framework", International Journal of Advanced Research in Computer and Communication Engineering vol. 2, 2013.
[7] Geetika Gautam, Divakar yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", IEEE, 2014
[8] Seyed-Ali Bahrainian, Andreas Dengel, "Sentiment Analysis using Sentiment Features", IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.
[9] Xujuan Zhou , Xiaohui Tao, Jianming Yong , Zhenyu Yang, "2.5.3 Sentiment Analysis on Tweets for Social Events", IEEE 17th International Conference on Computer Supported Cooperative Work in Design, 2013
[10] Jun YANG, Lan JIANG, Chongjun WANG and Junyuan XIE, "Multi-Label Emotion Classification for Tweets in Weibo : Method and Application", IEEE 26th International Conference on Tools with Artificial Intelligence,2014.
[11] Diego Terrana, Agnese Augello, Giovanni Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream", IEEE International Conference on Semantic Computing,2014.
[12] Uma Nagarsekar, Priyanka Kulkarni, Aditi Mhapsekar, Dr. Dhananjay R. Kalbande, "Emotion Detection from "The SMS of the Internet"", IEEE Recent Advances in Intelligent Computational Systems (RAICS),2013.
[13] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey, Production and hosting by Elsevier B.V. on behalf of Ain Shams University,2014.
[14] Khairullah Khan, Baharum B. Baharudin, Aurangzeb Khan, Fazal-e-Malik, "Mining Opinion from Text Documents: A Survey", 3rd IEEE International Conferences on Digital Ecosystem and Technology, 2009.
[15] Haruna Isah, Paul Trundle, Daniel Neagu, "Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis", IEEE,2014