A Proposed Solution for Sentiment Analysis on Tweets to Extract Emotions from Ambiguous Statements

Dyuti Shukla
Dwarkadas J. Sanghvi College of Engineering
Mumbai, Maharshtra, India

Mihika Shah
Dwarkadas J. Sanghvi College of Engineering
Mumbai, Maharshtra, India

Prerna Parmeshwaran
Dwarkadas J. Sanghvi College of Engineering
Mumbai, Maharshtra, India

Prof. Kiran Bhowmick
Dwarkadas J. Sanghvi College of Engineering
Mumbai, Maharshtra, India

***Abstract*-- As consumers move towards social media platforms like Twitter and Facebook to air their views about a variety of products, performing sentiment analysis on their responses becomes a desirable activity that can return a wealth of information about public perception. However, information posted in such networks is designed for human consumption instead of computers, and the nuances a person can catch from them are difficult for a machine to interpret. Thus most work in this field has always concentrated on polarity detection of the opinion into three broad fields of positive, negative or neutral. In this paper we aim to look at other techniques and emotion models that would aid us in helping computers understand the emotions attached to such ambiguous statements. We compare various techniques used for sentiment analysis to that end, and propose a solution for the same.**

Keywords: Sentiment analysis, emotion models

I. INTRODUCTION

Social media is a goldmine of opinions - people from all walks of life take to such platforms to express what affects them most. The ease of access they provide, along with their global reach make them very effective carriers of all kinds of expressions, be it a product review, a political statement or a reaction. Of all available platforms, Twitter is the preferred destination for such reviews, and is often used by companies as an evaluation measure for their products and services. Tweets and retweets are very helpful indicators of public sentiment, and utilizing the wealth of information they provide is imperative to any company aiming for higher sales, better growth or hoping to correct a failing product.

However, Twitter data is humongous and cannot be perused by a human, which is where data mining and machine learning techniques become useful. They can process a large volume of tweets in smaller instances, drawing effective conclusions and providing relevant data. However, information on Twitter is written for the understanding of fellow humans and the ambiguity in the statements, which a human reader can easily understand proves to be a difficult task for a computer. Several other factors also contribute to ambiguity in statements, including contrasting polarities, slang, use of non standard spelling for exaggeration or emphasis and context dependent words. Without the inference and cognitive ability of a human, a computer struggles with examining such statements.

Finally, in order to provide a more accurate picture of the user's opinions, it is simply not enough to classify them by polarity. More in-depth knowledge of their expression is needed, thus in this project, we will aim to extract human emotions conveyed in the tweet with the help of existing emotion models from the domain of psychology.

II. LITERATURE SURVEY

A. Sentiment Analysis Techniques

The sentiment analysis methods discussed below are used to find the polarity of the text. There are many approaches to this problem. The two most popular ones are the lexicon based and the learning based approaches. There is a third, hybrid approach [1] that combines the ones listed above. Finally, concept level sentiment analysis methods are also used for this task.

Lexicon Based Methods

As the name suggests, lexicon based methods make use of a dictionary, or a lexicon, that contains opinion words and their associated sentiments. The tweets are retrieved and pre-processed. Then, the tweets are parsed and tokenized. Each token is compared with the entries in the lexicon. If the token exists in the lexicon, and the associated sentiment is positive, the overall score is incremented. If the sentiment is negative, the overall score is reduced. Finally, after the parsing is complete, the score is compared to a threshold value. If the score is greater than the threshold, the polarity is said to be positive, else the polarity is said to be negative.

The lexicon based approach to sentiment analysis seems very intuitive and is easy to implement. However, this method will fail for ambiguous phrases, because the words in such sentences do not necessarily match the primary meaning of the word. Also, the lexicon based approach does not usually take the context of the opinion into consideration. Finally, the construction of an exhaustive lexicon is a tedious process.

Learning Based Methods

Most learning based approaches use classification techniques to determine the polarity of the text. Usually, the learning used is supervised learning, and therefore, a training dataset is required. The most popular techniques used to perform learning based analysis are Naive Bayes, Support Vector Machines and Maximum Entropy [5]. Feature selection is an important step in machine learning, because these features determine the results of the classifiers. Some of the features that are used for performing sentiment analysis are: term presence and frequency, Part of Speech (POS) tagging, negations encountered in the sentence and the list of opinion words/ phrases amongst other features.

The classifiers that are used are:

1. Naive Bayes Classifier:
   Naive Bayes Classifier is a probabilistic classifier which uses Bayes Theorem.The basic assumption of this classifier is that the features are assumed independent. This model can be combined with a decision rule. There are a few variations to the Naive Bayes Model, two of which are the Bernoulli Model and the Multinomial Model. The advantage with Naive Bayes is that it is extremely simple to implement, but since features are assumed to be independent of each other, POS cannot be used.

2. Support Vector Machines (SVM):
   The main idea for SVM for sentiment classification is to find a hyperplane that divides the documents as per the sentiment, and the margin between these classes should be as high as possible. The principle of SVM is Structural Risk Minimization [5]. The objective is to find a hypothesis h for which the error is the lowest. If we symbolize the hyper plane as h and the tweet as t, and represent the classes into which the tweet has to be classified as $C_j \in \{1,-1\}$ corresponding to the sentiment of the tweet, the solution can be written as:

$$\vec{h} = \sum \alpha_i C_i \vec{t}_j$$

   The texts that have α>0 are the ones which contribute in finding the hyperplane. SVMs can handle large feature spaces with a high number of dimensions. Also, SVM does not assume any feature to be irrelevant. However, the main problem with SVM is that it is difficult to identify which features are more important for classification.

3. Maximum Entropy:
   The main principle of maximum entropy is that a uniform model is preferred to satisfy the given constraints. It can be used to estimate any probability distribution. In maximum entropy, the training data is used to set constraints on the conditional distribution. Each constraint should express a characteristic of the training data that should also be present in the learned distribution.

Machine learning methods are somewhat more successful than lexicon based methods because they take the various features and their relationships into consideration while deciding the polarity of the sentence. However, the requirement of a large training dataset is a drawback.

Hybrid Methods

Hybrid methods use a combination of lexical and learning based methods. They show higher performance than the two methods used alone [1].

An example of a hybrid approach is feature based sentiment analysis [2]. The steps are as follows:

1. Feature and opinion extraction: A tokenized sentence is given as input and the output is a list of feature words and a list of opinion words in the sentence.
2. Anaphora resolution: Examples of anaphora include the usage of pronouns in a sentence so as to avoid the repetitive usage of a noun. Usage of anaphora makes the sentence ambiguous because it is difficult for the computer to map opinion words to the corresponding feature. Hence, backtracking is used to resolve the usage of anaphora, so as to map the opinions to the correct features.
3. Feasibility analysis: Extraneous words that are not related to the process of sentiment analysis are eliminated in this step.
4. Statistical features identification: A set of positive and negative seed words already exists in the form of the sentiment dictionary. In this step, the correlation of the opinion words with either of these sets is calculated.
5. Sentiment determination: Now that the nature of the opinion words has been calculated, the overall sentiment expressed in the sentence is determined using various machine learning algorithms. Usually, supervised ML techniques are used for this purpose.

Yet another hybrid approach [6] determines the polarity of the sentence taking the use of emoticons into consideration. It uses both lexical and machine learning techniques. The underlying assumption of this method is that the overall orientation of the emoticons used in a sentence is the same as the overall sentiment expressed by the words in that sentence. In simpler terms, the emotions expressed using words and emoticons in one particular sentence would be the same. A list of annotated emoticons (i.e. a list of emoticons and their related sentiments) has to be used for this task. Also, a lexicon containing words and their sentiments has to be used. The assumption about the orientations of the emoticons and the words in the sentence makes it quite easy to implement, because it is usually correct. However, in ambiguous phrases, this assumption may not hold true and this method would fail in such a case.

Concept Level Sentiment Analysis

Concept level methods [4] can be used to determine emotions that are expressed subtly, as long as they can be connected or related to concepts that are present in the sentences. Concept level techniques depend on large semantic dictionaries that act as a repository for semantic words and concepts. Thesaurus and Commonsense dictionaries are two commonly used propagation methods, but the use of Thesauruses are generally limited to lexical dictionaries only. A thesaurus is used to map syntactically similar words by means of some defined syntax rules. Commonsense dictionaries on the other hand map relations between concepts. These dictionaries when employed in sentiment analysis have higher recall than word based match techniques with greater variety of relations among the elements.

Concepts in such commonsense networks may be broken down into word based relations or parsing the entire sentence to find assertions in sentences. Assertions comprised of two concepts, and words such as “Used For” “IsA” are used to demonstrate the relation between them. For example, consider the sentence “Banana is a fruit.” Here, the sentence banana is a fruit can be parsed into assertions thus: Banana/ isA /Fruit. The concepts mentioned here are “Banana” and “Fruit” with “IsA” specifying the relation between them. The division of such concepts has tremendous value in sentiment analysis projects, where sentiments can be related to each other on the basis of the concepts given in them. Assumptions can be made on the fact that semantically related words have similar sentiments related to them.

B. Emotion Models

Classifying a statement as a positive, negative or a neutral opinion will not suffice if one wants to gather more information about the user’s mood while tweeting. Identifying the basic emotion of the user will help organizations understand how the user really perceives his product. Performing emotion extraction on text is an uphill task because human emotions are very complex and subjective in nature. However, with the help of emotional models, one can broadly classify a piece of text into a predefined set of human emotions. Some significant emotional models are as shown below:

- Russell’s circumplex model:
  James Russell conceptualized emotions as being distributed in a two dimensional space, where the X axis represents valence emotions and the Y axis represents the arousal. Valence is related to the polarity, i.e. the attractiveness or aversiveness towards a particular subject or an event. Arousal deals with the intensity of the emotions felt. A more specific version of circumplex [9] exists for customer reviews as well, created in 2008 by Desmet. A selection of 24 relevant emotions is made based on valence and arousal, while eliminating the unnecessary emotions in the circumplex model based on reactions to a product’s appearance. This circular model was then used for product review evaluation software.

- Plutchik’s wheel of emotions:
  Robert Plutchik’s ‘Wheel of Emotions is yet another popular emotion model. It takes into consideration eight basic emotions, and the emotions obtained by varying the intensities of the basic ones. Additionally, various emotions in this model can be combined to obtain more complex emotions. The Plutchik’s model has been used in extracting opinions from text [7] and has been found computationally suitable for this task [8].

- OCC Model:
  The OCC model is similar in sense to many existing cognitive models in which it compares effect of an action (valence) on the desirability of the consequence of the action. The model separates and quantifies emotions based on their underlying strategic patterns of appraisal such as the consequences one would consider applying in any situation. Such responses are derived from a person’s primary focus at the time, what affects them and how they react in response to the subject in question.

Using these psychological models, we can map the text that we extract from the tweets, obtain the core emotion and also the intensity and decide which emotion is most significantly expressed in the tweet. These emotional models can be used in conjunction with neuro-fuzzy systems [10] to perform the task of emotion extraction.

III. PROPOSED SYSTEM

To overcome the drawbacks of the methods we have reviewed above, we propose a new model for sentiment analysis. In this model we combine many techniques to reach our final goal of emotion extraction. The steps for the process are documented below.

1. Retrieval of Data: Public Twitter data is mined using the existing Twitter APIs for data extraction. Tweets would be selected based on a few chosen keywords pertaining to the domain of our concern, i.e. product reviews. We have elected to use the Twitter API due to ease of data extraction.
2. Preprocessing: In this stage, the data is put through a preprocessing stage in which we remove identifying information such as Twitter handles, timestamps of the message and embedded links and videos. Such information is largely irrelevant and may cause false results to be given by our system.
3. Tweet Correction: As tweets are written for human perusal, they often contain slang, misspellings and other irrelevant data. Thus we correct the misspellings in the sentences and look to replace the slang in the sentences with words from standard english that may roughly relate to the slang in question. As slang itself can be used to display a wide variety of sentiment, often with greater emotional impact, this process is necessary so that slang words may be considered as part of the emotion expressed.

4. Polarity detection: In this step we begin the second phase of our proposed system, in which we try to identify the polarity of the sentence in question. If emoticons exist in the statements, they will be used as well to compute the overall polarity of the statement. We aim to find sentences where the polarity detection is not very clear or where the expressed sentiment may be low. We also try to isolate the opinion words in the sentence in relation to a given concept in the sentence.
    a. We train the system to understand the relation between words in various contexts. Pre-existing dictionaries like SenticNet can be used in this phase to segregate the emotion from the context it is in.
    b. Once the opinion words are identified with context, we can find the polarities of the words using NLTK-SentiWordNet.
    c. To help with detection of the concepts associated, we train our system on a large dataset that expresses a wide variety of complex and ambiguous emotions. The system is given this data in an unsupervised fashion and will proceed by clustering.
5. Emotion Extraction: Emotion models often map the core emotions to a computational scale from which we can broadly classify and detect the emotions expressed. For the purposes of our system, we consider the "Plutchik's Wheel of Emotion" which divides all emotions into an eight-point wheel which represents the intensity and complexity of human feeling as we move from the centre of the wheel to the outer rim. The central core is made of 8 basic emotions that decrease in intensity as we move away from the centre, often blending with one or more emotion to become increasingly complex. For example, the wheel may express the simple emotions "rage" and "loathing" at the centre, but the rims contain the harder to identify emotions of "contempt", "boredom" and "annoyance".
    a. Mapping: Once the emotional relation has been extracted, we map it to Plutchik's model using a neuro-fuzzy inference system. As ambiguous phrases contain a high probability of expressing two or more emotions together in order to create a complex feeling, a neuro-fuzzy system is designed so that the emotions may be computed to a membership function instead.
    b. Once the system calculates the degree of membership of the emotion or emotions expressed in the statements, we use it to determine the most significant emotions. This value is decided after comparing all the degrees of membership given by the opinion words in the statement.
6. A graphical representation is provided for the statement.

The block diagram for the proposed system is given below in Figure 1:

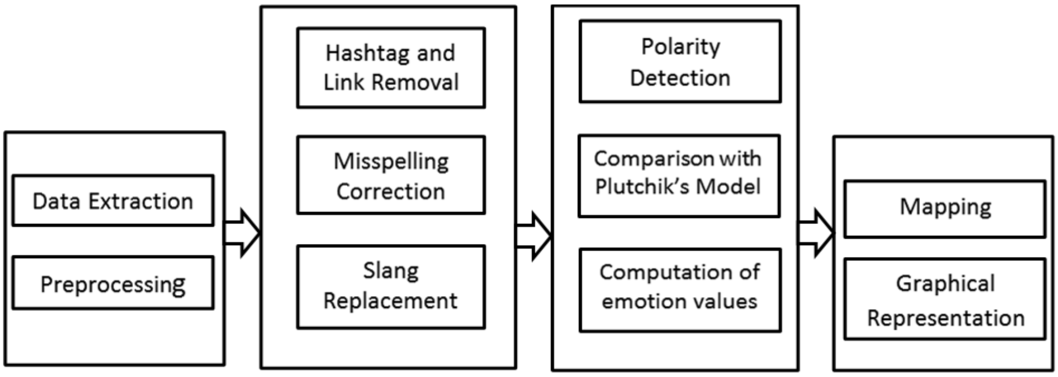

Fig 1: Model of Proposed System

IV. CONCLUSION AND FUTURE WORK

Performing sentiment analysis on data obtained from Twitter is a huge challenge because of the amount of ambiguity involved. Due to the widespread usage of slang, wrong spellings, emoticons etc. it becomes difficult for automatic detection of emotions from tweets. This project is a small step towards the efficient automation of sentiment analysis by focusing on ambiguous statements. The system proposed by us also attempts to extract actual emotions from tweets. Such a system will be very useful for various marketing teams to gain actual and detailed feedback from their users.

At present, we have only proposed a system to perform the extraction of emotions from ambiguous tweets. The implementation has to be done and the system must be trained. At this stage, the project is limited to product reviews aired by users on Twitter. In the future, the system can also be extended to analyze sentiments about politics, finance and other affairs. Complete removal of ambiguity is an uphill task indeed. Therefore, interpretation and classification of sarcastic sentences are not a part of the current scope. However, in the future, the scope can be extended to accommodate the same. Finally, the project can be extended to work for natural languages other than English.

ACKNOWLEDGMENTS

The authors are grateful to Dwarkadas J. Sanghvi College of Engineering and the Computer Engineering Department for their support and encouragement.

REFERENCES

[1] S. M. Vohra and J. B.Teraiya, "A Comparative Study of Sentiment Analysis Techniques", Journal of Information, Knowledge and Research in Computer Engineering, 2012.

[2] Ahmad Kamal and Muhammad Abulaish, "Statistical Features Identification for Sentiment Analysis using Machine Learning Techniques", International Symposium on Computational and Business Intelligence, 2013.

[3] Neethu M.S and Rajasree R., "Sentiment Analysis in Twitter using Machine Learning Techniques", Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013.

[4] Erik Cambria, "An Introduction to Concept-Level Sentiment Analysis", 2013.

[5] Bhuta S., Doshi A., Doshi U. and Narvekar M., "A Review of Techniques for Sentiment Analysis Of Twitter Data", International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014.

[6] Shuigui Huang, Wenwen Han, Xirong Que and Wendong Wang, "Polarity Identification of Sentiment Words based on Emoticons", Ninth International Conference on Computational Intelligence and Security, 2013.

[7] Dhanashri Chafale and Amit Pimpalkar, "Sentiment Analysis on Product Reviews Using Plutchik's Wheel of Emotions with Fuzzy Logic ", ABHIYANTRIKI An International Journal of Engineering & Technology (AIJET), 2014.

[8] Alastair J. Gill, Robert M. French, Darren Gergle and Jon Oberlander, "The Language of Emotion in Short Blog Texts", Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work, 2008.

[9] NaseemAhmadpour, "OCC model: application and comparison to the dimensional model of emotion ", International Conference on Kansei Engineering and Emotion Research, 2014.

[10] Kuo-Kuang Fan, Shuh-Yeuan Deng, Chung-Ho Su, Fu-Yuan Cheng, "Theory of Variable Fuzzy Sets for Artificial Emotions Prediction", Mathematical Problems in Engineering Volume 2015, 2015