International Journal of Research in Engineering, Science and Management
Volume-1, Issue-10, October-2018
www.ijresm.com | ISSN (Online): 2581-5792


# Sentimental Analysis Using Machine Learning

S. Muthamilselvan[1], Arka Raha[2], Sunny Kumar[3], Sayan Ghoshal[4]
[1]*Asst. Prof., Dept. of Computer Science & Engg., SRM Institute of Science and Technology, Chennai, India*
[2,3,4]*Student, Dept. of Computer Science & Engg., SRM Institute of Science and Technology, Chennai, India*

***Abstract*—Sentimental analysis is an ongoing field of research in text mining field. It deals with the identifying and classifying opinions or sentiments expressed in source text. Nowadays, social media is booming in exponential and is generating a vast amount of sentimental rich data in form of tweets, updates, blog posts, etc. this data generated by user is very useful for knowing the opinion of the crowds. Compared to general sentimental analysis twitter sentimental analysis is difficult due to the presence of bad words, misspellings and short forms. Machine learning approach is one of the strategies for analysing sentiments from the text. In this paper, we try to analyse the twitter posts about from mobile, laptops, pc, etc. using machine learning. We present a new feature for differentiating the tweets as positive, negative and extract peoples' opinion about products.**

***Index Terms*—Emotion Detection, Lexicons, NLP Techniques, Stochastic Agreement Regularization algorithm.**

## I. Introduction

The best businesses is to understand sentiment of the customers – what people are saying, what they are saying and what they actually mean it. Sentiment Analysis is the domain of understanding these emotions with software and it is must to understand for developers and business leaders in a modern workplace. As with many other fields, advancement in Deep Learning have brought Sentiment Analysis into the foreground of cutting-edge algorithms. Now a days we use natural language processing and text analysis to know and identify the sentiment of text into positive, negative or neutral categories.

Sentimental Analysis is also used for Brand Monitoring. Now-a-days sentimental analysis is used to get full view of how brand, product or company is viewed by the customers and stake holders. Multi-national companies can also use sentimental analysis to measure the impact of their new product in market or social media.

Sentimental analysis is also used for advancement of customer service. As customer service becomes more and more automated through Machine Learning, understanding the sentiment of a given case becomes increasingly important.

Sentimental Analysis is also used for marketing research and analysis. It is used in business intelligence to understand the choice of interest of consumers that the things they are or are not responding to something about their product and service. They check why customer are not buying their product or what was there experience after using their products or the customer service support meet their expectations or not ??... Sentiment analysis can also be used in the areas of political science, sociology, and psychology to analyse trends, ideology, opinions, gauge reactions, etc. There are also some challenges in front of sentimental analysis. Sentiment Analysis runs into a similar set of problems as emotion recognition does, before deciding the sentiment of a given sentence. Actually, we need to figure out what "sentiment" is in the first place. Sentiment can be split into clear brackets like happy, sad, angry, or disappointed? There are many layers of meaning in any human sentence. People express themselves in complex ways; like sarcasm, irony, and implied meaning can mislead sentiment analysis. There is only way to understand these devices or system are through the help of context: knowing how a paragraph is started can strongly impact the sentiment of later internal sentences. Today microblogging has become a very popular communication tool among Internet users. Billions of messages, post, etc. are appearing daily in popular web-sites that provide services for microblogging such as Twitter, Facebook, Instagram. Peoples shares opinions on variety of topics and discussion the current issues. Due to the free format of messages and an easy accessibility of microblogging platforms the users are tend to shift from traditional communication ways to microblogging services.

So, basically what actually we are going to do is?? We use a dataset of collected messages from twitter. Twitter contains a huge number of short messages created by the users of this social/microblogging platform. The body of the messages vary from personal opinions to public statements.

As the users of blogging/social platforms and services grows every day, therefore the data from these sources can be used in opinion mining and sentiment analysis tasks.

## II. Related Works

This technology is used to collect and analyse opinions and reviews about product, service or a brand. However, the main use of this sentimental analysis is on social networks. It allows to gain an overview of the public opinion behind certain topics and emotional reaction to a document, interaction or event. Now-a-days, the applications of sentiment analysis are wide and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the globe. Indeed, the ability to quickly understand consumer attitudes and reaction accordingly is something that businesses take advantage of. These businesses listening to that feedback

and adjusting accordingly are being active and are current winning companies.

Sentiment analysis is the automated process of determining an opinion about a given subject from written or spoken language. It's estimated that 80% of the world's data is unstructured and not organized in a pre-defined way. Most of this comes from text data, like emails, chats, social media, surveys, articles, and documents. These texts are usually hard, time-consuming and expensive to analyze, understand, and sort through. Sentiment analysis systems allows companies to make sense of this huge sea of unstructured text by automating business processes, getting actionable insights, and saving hours of manual data processing, in other words, by making teams more efficient. There are many methods and algorithms to implement sentiment analysis systems, which can be differentiated as:

Rule-based systems that perform sentiment analysis based on a set of crafted rules.

Automatic systems that depends on machine learning techniques to learn from data.

Hybrid systems that combine both of the above mentioned methods.

### A. *Rule-based Approaches*

Usually, rule-based approaches is defined as a set of rules in some kind of scripting language that identify subjectivity, polarity, or the subject of an opinion.

The rules may use a variety of inputs, such as the follows:

Classic NLP techniques like stemming, tokenization, part of speech tagging and parsing may be used for this Also, other resources, such as lexicons (i.e. lists of words and expressions).

An example of a rule-based implementation would be the following:

First define two lists of polarized words (e.g. negative words such as bad, worst, ugly, poor, etc and positive words such as good, best, beautiful, etc.).

Analyse the given text as follows,

First, count the number of positive words that appear in the text.

Secondly, count the number of negative words that appear in the text.

Upon comparing if the number of positive word appearances is more than the number of negative word appearances then return a positive sentiment, else, return a negative sentiment. Otherwise, return neutral. This system is much unsophisticated since it doesn't take into account how words are combined in a sequence. A more advanced processing can be made but these systems get very complicated quickly. They can be very difficult to maintain as new rules may be needed to add support for new expressions and vocabulary. Besides, addition of new rules that may have undesired outcomes as a result of the interaction with previous rules. As a result of which, these systems require important investments in manually tuning and maintaining the rules.

## III. Existing System

Tweepy is an open-sourced, which enables Python to communicate with Twitter platform and use its API. It supports OAuth authentication. Authentication is handled by the tweepy. AuthHandler class.

TextBlob is a Python (2 and 3) library for the processing of textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, differentiation, translation, and more. TextBlob's attribute of polarity tells the attitude of phrase on the scale [-1.0 to 1.0] where -1.0 stands for negative and 1.0 stands for positive. TextBlob's attribute of subjectivity is a float within the range [0.0 to1.0] where 0.0 is very objective and 1.0 is very subjective.

## IV. Proposed System

In the proposed system, we will use scikit learn, a popular machine learning library. Scikit-learn has a couple of dependencies, those are numpy and scipy. It provides several vectorizers to translate the input documents into vectors of features (or feature weights). It needs to be given appropriate weights to different words, and TF-IDF is the most common weighting schemes used in such applications. Scikit-learn have a number of different classifiers already built-in. In these experiments, different variations of Support Vector Machine (SVM) are used. Scikit-learn comes with a number of distinct classifiers already built-in. In these experiment, we use different variations of Support Vector Machine (SVM), which is commonly used in classification applications.

## V. Concepts Used

Sentiment analysis is sometimes considered as an NLP task for discovering opinions about an entity; because there is some ambiguity about the difference between opinion, sentiment and emotion, they defined opinion as a transitional concept that reflects reaction towards an entity. The sentiment is the part that refers to feeling or emotion while emotion refers to attitude.It was argued by Plutchik that there are eight basic emotions: joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. Emotions Detection (ED) can be considered a SA task; SA is concerned with specifying positive or negative opinions, but ED is concerned with detecting various emotions from text. As a Sentiment Analysis task, Emotion Detection can be implemented using ML approach or Lexicon-based approach, but Lexicon-based approach is more frequently used.

Lu and Lin proposed a web-based text mining approach for detecting emotion of an individual event embedded in English sentences. Their approach was based on the probability distribution of common mutual actions between the subject and the object of an event. They integrated web-based text mining and semantic role labeling techniques, together with a number

of reference entity pairs and hand-crafted emotion generation rules to recognize an event emotion detection system. They did not use any large-scale lexical sources. They showed that their approach revealed a satisfactory result for detecting the positive, negative and neutral emotions. They proved that the emotion detecting problem is context-sensitive.

Using both ML and Lexicon-based approach was introduced by Balahur et al. They proposed a method based on commonsense knowledge stored in the emotion corpus (EmotiNet) knowledge base which said that emotions are not always expressed by using words with an effective meaning i.e. happy, but by describing real-life situations which are detected. They used SVM and SVM-SO algorithms to reach their goal. They showed that the approach based on EmotiNet is the most appropriate for the detection of emotions from contexts where no affect-related words were present. They proved that the task of emotion detection from texts such as the ones in the emotion corpus ISEAR (where little or no lexical clues of affect are present) can be best solved using approaches based on commonsense knowledge. They showed that by using EmotiNet, they obtained better results compared to the methods that employ supervised learning on a much greater training set or lexical knowledge. Affect Analysis (AA) is a task of identifying emotions elicited by a certain semiotic modality. Neviarouskaya et al have suggested an Affect Analysis Model (AAM). Their AAM consists of five phase: symbolic cue, syntactical structure, word-level, phrase-level and sentence-level analysis. This AAM was used in many applications introduced in Neviarouskaya work. They developed a system that relied on the compositionality principle and a novel approach dealing with the semantics of verbs in attitude analysis. They worked on 1000 sentences. Their evaluation showed that their system achieved reliable results in the task of textual attitude analysis.

### *A. Building resources*

Building Resources (BR) aims at creating lexica, dictionaries and corpora in which opinion expressions are interpreted according to their polarity. Building resources is not a SA task, but it could be used to improve SA and ED as well. The main challenges that confronted the work in this category are ambiguity of words, multilinguality, granularity and the differences in opinion expression among textual genres

Building Lexicon was presented by Tan and Wu; in their work, they proposed a random walk algorithm to construct domain-oriented sentiment lexicon by utilizing sentiment words and documents from both old domain and target domain simultaneously They coordinated their experiments on three domain-specific sentiment data sets. Their experimental results indicated that their proposed algorithm improved the performance of automatic construction of domain-oriented sentiment lexicon. Building corpus was introduced by Robaldo and Di Caro who proposed Opinion Mining-ML, a new XML-based formalism for tagging textual expressions conveying opinions on objects that are considered relevant in the state of affairs. It is a brand new standard beside Emotion-ML and WordNet. Their work consisted of two parts; first, they presented a standard methodology for the annotation of affective statements in the text that was strictly independent from any application domain; second, they considered the domain-specific adaptation that relied on the use of ontology of support which is domain-dependent. They started with the data sets of restaurant reviews applying query-oriented extraction process. They evaluated their proposal by means of fine-grained analysis of the disagreement between different annotators and their results indicated that their proposal represented an effective annotation scheme that was able to have both high complexity while preserving good agreement among different people.Boldrini et al. have focused on the creation of EmotiBlog, a fine-grained annotation scheme for labeling subjectivity in nontraditional textual genres. They focused on the annotation at different levels: document, sentence and element. They also presented the EmotiBlog corpus; a collection of blog posts composed by 270,000 token about three topics in three languages such as, Spanish, English and Italian. They checked the robustness of the model and its applicability to NLP tasks. They examined their model on many corpora i.e. ISEAR. Their experiments provided satisfactory results. They applied EmotiBlog to sentiment polarity classification and emotion observation. They proved that their resource improved the performance of systems built for this task. Building Dictionary was presented by Steinberger et al. In their work they proposed a semi-automatic approach to creating sentiment dictionaries in many languages. They first produced high standard sentiment dictionaries for two languages and then translated them automatically into a third language. Those words that are found in both target language word lists are likely to be useful because their word senses are likely to be similar to that of the two source languages. They addressed two issues during their work; the morphological inflection and the subjectivity involved in the human annotation and evaluation effort. They worked on news data. They compared their triangulated lists with the non-triangulated machine-translated word lists and verified their approach.

### *B. Transfer Learning*

Transfer learning extracts knowledge from auxiliary domain to upgrade the learning process in a target domain. For example, it transfers knowledge from Wikipedia documents to tweets or a search in English to Latin. Transfer learning is considered as a new cross domain learning technique as it addresses the various aspects of domain differences. It is used to enhance many Text mining tasks like text classification sentiment analysis. Named Entity recognition part-of-speech tagging, etc. In Sentiment Analysis- transfer learning can be applied to transfer sentiment classifications from one domain to another domain or building a bridge between two domains. Tan and Wang proposed an Entropy-based algorithm to pick out high-frequency domain-specific (HFDS) features as well as a weighting model which weighted the features as well as the

instances. They assigned a smaller weight to HFDS features and a larger weight to instances with the same label as the involved pivot feature. They worked on education, stock and computer reviews that come from a domain-specific Chinese data set. They proved that their proposed model could overcome the adverse influence of HFDS features. They also showed that their model is a better choice for SA applications that require high-precision classification which have hardly any labeled training data. Wu and Tan have proposed a two-stage framework for cross-domain sentiment classification. In the first stage a bridge between the origin domain and the target domain was built to get some most confidently labeled documents in the target domain. In the second stage the intrinsic structure was exploited, to reveal these labeled documents to label the target-domain data. They worked on books, hotels, and notebook reviews that came from a domain-specific Chinese data set. They proved that their proposed approach could improve the performance of cross-domain sentiment classification.

The Stochastic Agreement Regularization algorithm that deals with cross-domain polarity classification is a probabilistic agreement framework based on minimizing the Bhattacharyya distance between models trained using two different views. It regularizes the models from each view by constraining the amount by which it allows them to disagree on unlabeled instances from a theoretical model. The Stochastic Agreement Regularization algorithm was used as a base for the work presented by Lambova et al. Which had discussion of the problems of cross-domain text subjectivity classification. They proposed three new algorithms based on multi-view learning and the co-training algorithm strategy constrained by agreement. They worked on movie reviews and question answering data that came from three famous data sets. They showed that their proposed work give improved results compared to the Stochastic Agreement Regularization algorithm. Diversity among various data sources is a problem for the joint modeling of multiple data sources; Joint modeling is important to transfer learning; that is why Gupta et al. have tried to solve this problem. In their work, they proposed a regularized shared subspace learning framework, which can exploit the mutual strengths of related data sources while being unaffected by the effects of the changeability of each source. They worked on social media news data that come from famous social media sites such as Blogspot, Flicker and Youtube and also from news sites as CNN, BBC. They proved that their approach achieved better performance than others.

## VI. System Architecture

We aimed pre-processing different advances that are taken as stop word evacuation, treatment of nullification, and shortened form of extension, incorrectly spell rectification, stemming, positive word arrangements of each tweet, and negative word arrangements of each tweet. Watchmen calculation is utilized for stemming.

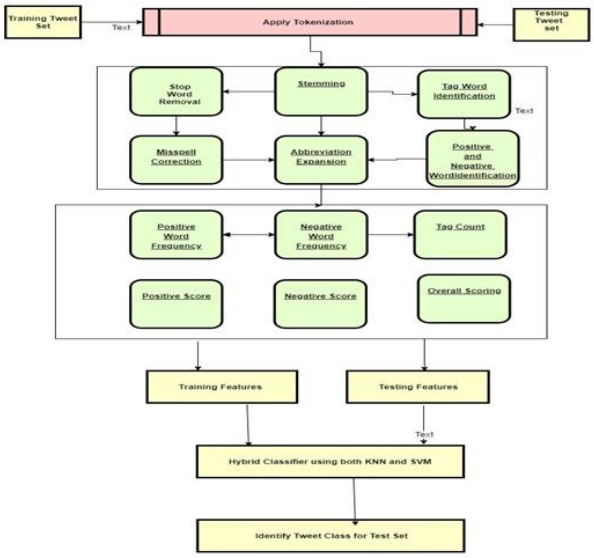


Fig. 1. Tweets subsequent to preprocessing

The definition and displaying of a design devoted to the exercises of investigation of enormous information, as the ones created by interpersonal organizations as Twitter, is presently still at a beginning period of its advancement and solidification.

Not at all like conventional information stockroom or business knowledge frameworks, is whose engineering intended for organized information, frameworks committed to huge information work rather with semi-organized information, or alleged "crude information", i.e. without a specific structure. It ought to likewise be brought up that such frameworks ought to be ready to permit preparing and examination of information in cluster mode, as well as in a genuine constant form.

These days a tremendous measure of information, day by day created by interpersonal organizations, can be prepared what's more, and examined for various purposes. These information are furnished with a few highlights, among which:

- measurement;
- idiosyncrasies;
- source;
- unwavering quality;

When the need to get the data and the manner in which this data must be handled has changed. As of not long ago it was imagined that the information ought to be first prepared and in this manner made accessible, paying little heed to the time angle. This sort of handling is normally called group preparing. These days the measure of information is expanded exponentially and now continuous handling is expected to get the most favorable circumstances from this information, in various fields.

For stopwords, shortened form expansion1
what's more, incorrectly spell correction2 database is made.
*Sifted list:* contains tweets after tokenization and applying the above composed channels Label Filtered list: contains the separated rundown with @ labels evacuated. The @ labels are utilized in highlight age as label check in each tweet.

*Negative List:* contains negative descriptive words in each tweet.
*Positive List:* contains positive descriptors in each tweet.

### A. Features Generation

A rundown of adjectives is utilized for highlights age. This rundown contains positive score, negative score, by and large rating of a descriptive word among different traits.
Depicting different traits of a descriptive word
*Properties Description*
Id - Numeric Unique id to all descriptive words
Descriptive word - Stores the printed data to speak to the genuine descriptive word
P-score - Positive score, to speak to the positive adequacy of a descriptive word Lies between 0 and 1
F-score - Negative score, Lies between 0 and 1
Score Overall score of descriptive word lies between - 1 and 1
+ve - values for +ve modifier
- ve - esteem for – ve modifier

Different highlights utilized for taking in the classifiers are:
Word include: Total words each tweet after filtration
Label include: Total @ labels utilized each tweet Negative word include: Total negative words each tweet
Positive word include: Total positive words each tweet
Positive score: Total positive score acquired by including the positive scores of every positive descriptive word.
Negative score: Total negative score acquired by including the negative scores of each negative modifier.
Score: Positive score-Negative score for each tweet
Message class 0: for negative tweets
1: for non-positive tweets
2: for positive tweets

## VII. Application and Future Work

I recommend that a future work on the Repetitive Neural Systems can give better outcomes. Since, they propose that is one of the better techniques for this sort of issue. Likewise, with the rise of specific equipment and the capacity to prepare with substantial datasets that does not fit in memory, utilizing Inclination Drop, makes the neural systems the most encouraging technique.

To enhance the general execution of this process, more examination must be done on learn with less changes. This is on account of, all the content changes have a misfortune on the amount of data we process, as more changes less data, and in the event that we have less data the likelihood to complete a wrong characterization expands.

At last, work at character level can have a high effect on the vocabulary measure on the grounds that can be diminished from 50 thousand to, almost, 3 hundred characters that can speak to the most words in the dialect. Additionally, is the manner in which that the people read. The issue is that, today, the machine learning strategies have issues preparing the halfway importance behind the time reliance among words and, similarly, the reliance between characters. In Twitter, for instance, a sentence has around 20 words that implies that the machine needs to take in the reliance between 20 time steps and decipher the concealed conclusion while perusing. In the event that we work at character level the machine needs to take in more than 300 hundred time steps. More examination is expected to influence the machines to learn like people perusing the sentences at character level.

## VIII. Conclusion

The task of sentiment analysis, especially in the domain of micro-bloging, is still in the developing stage and far from complete. So we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance. Right now we have worked with only the very simplest unigram models; we can improve those models by adding extra information like closeness of the word with a negation word. We could specify a window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be put into the model if it lies within that window. The closer the negation word is to the unigram word whose prior polarity is to be calculated, the more it should affect the polarity. For example if the negation is right next to the word, it may simply reverse the polarity of that word and farther the negation is from the word the more minimized ifs effect should be.

Apart from this, we are currently only focusing on unigrams and the effect of bigrams and trigrams may be explored. As seen from the literature review section when bigrams are used along with unigrams the performance is usually enhanced.

In this research the main focus is on general sentiment analysis. There is potential of work in the field of sentiment analysis with partially known context. For example we noticed that users generally use our website for specific types of keywords which can divided into a couple of distinct classes, namely: politics/politicians, celebrities, products/brands, sports/sportsmen, media/movies/music. So we can attempt to perform separate sentiment analysis on tweets that only belong to one of these classes (i.e. the training data would not be general but specific to one of these categories) and compare the results we get if we apply general sentiment analysis on it instead.

## References

[1] Mart´ın Abadi, Ashish Agarwal, Paul Barham, et al. “TensorFlow: Large-scale machine learning on heterogeneous systems”, 2015.
[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel,,et al. “Languages for Data Mining and Machine Learning”, 2013.
[3] P. Raghavan C.D. Manning and H. Schuetze. “Introduction to information retrieval” 2008.
[4] Alec Go, Richa Bhayani, and Lei Huang. “Twitter sentiment classification using distant supervision,” 2009.
[5] Ozan Irsoy and Claire Cardie. “Opinion mining with deep recurrent neural networks”, 2014.
[6] Andrej Karpathy. “The unreasonable effectiveness of recurrent neural networks”, 2015.