BIU Journal of Basic and Applied Sciences 9(1): 82 – 96, 2024.
©Faculty of Science, Benson Idahosa University, Benin City, Nigeria
ISSN: 2563-6424

MACHINE LEARNING-BASED SENTIMENT ANALYSIS OF TWEET DATA

***ORIE, C. E.,[1] EGWALI, A. O.[2] AND AMADIN, F. I.[2]**
[1]Department of Computer Science, Benson Idahosa University, Benin City, Nigeria
[2]Department of Computer Science, University of Benin, Benin City, Nigeria
*Correspondent author: corie@biu.edu.ng

ABSTRACT

The idea of sentiment analysis is to assess and ascertain various opinions on specific data as it relates to various discuss on product review. These sentiments could be assessed in terms of products, human behaviours, events, or topics that are covered by reviews. Machine learning is used to classify data in line with various techniques or models developed. However, in the context of sentiment analysis, several issues, particularly with regards to text classification are yet to be resolved; this has been a major cause for concern to many researchers. Digital technologies, as well as social media platforms such as Instagram, Twitter, micro-blogs, etc are veritable sources for opinion mining. Today a lot of people and businesses often use the information in this media to make efficient decision. Most importantly the major challenge or problem with sentiment analysis is the fact that a lot of individuals classify the polarity of a text either at the document level, the sentence level and the feature aspect level as either a positive statement, a negative statement, or a neutral statement. In this study, we gave an overall information of the various sentiment analysis techniques, stating their strengths and weaknesses and the various challenges involve in classification of tweet. We also reviewed some literatures on sentimental analysis in line with the various machine learning techniques.

KEYWORDS: *Sentimental, Analysis, Opinion, Mining, Polarity, Data and Tweet*

INTRODUCTION

Sentiment Analysis of Tweet Data is very significant to all individuals as well as students, education, businesses, hospitals, various establishment and industries and even politicians. It can be a good source of getting important information that can be beneficial to individuals and businesses. However, the study of sentiment Analysis, if done properly, is exceptionally complex and is a field of study, not just a future in social media tool (Bing, 2012).

The subject matter of Sentiment Analysis (SA) is to statistically compute and ascertain people's opinions, attitudes and emotions toward an entity. These sentiments could be assessed in terms of individuals' dispositions, events or topics that are covered by reviews (Amreen and Madhuri, 2017). Sentiment analysis is widely used today across different establishments and industries as well as business domains. It involves deriving insights from any given text data, as well as understanding the emotional tone or sentiment polarities expressed by any individual. This technique helps to identify the orientation of a sentence thereby recognizing the element of positivity or negativity in (Andreas *et al.*, 2017).

According to Bing (2012), there are three general levels of Sentiment Analysis which include the Document Level where the Analysis tries to classify the overall sentiment of the entire document. This Level of analysis provides a very high level of understanding of the sentiment passed in the sentence or text data without getting into key details. The second level is the Sentence Level that involves analyzing the sentiment at a finer grained level thereby classifying the individual's sentiment within the text data (Panda *et al;* 2020). The Sentence Level more so recognizes the fact that the document can contain a combination of either positive, negative or neutral data. The third Level is the Aspect Level which tries to ensure the total sentiment of the persons statement or written document ensures the identification of the sentiment relating to some particular features, aspects or the individual entities relayed in the document being discussed. The Aspect Level provides key details into the various aspect in which the speaker or writer view the document making this level very useful as businesses and organization can have a good understanding of their customers perspective towards their products and provide avenue for any improvement on their product if there's any need to.

Today many individuals derive their sentiments regarding any subject or product of interest from social medial which has encouraged the rapid growth. There are various kinds of social medial expressions which can be either product ratings, product reviews or recommendations made by people which are the major sources of interest as well as information needed by most business trying hard to market their products and services in order to improve their social reputation.

Sentiment Analysis has various issues regarding the classification of text data and many researchers are still trying to solve these challenges. With the explosive growth of social media (e.g., reviews sites, discussions forum, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making (Amreen, 2017). Analyzing and classifying the sentiments expressed by individuals in online social media is an important area for researchers to look into. There are numerous issues with sentiment Analysis of tweet data. One major issue is an individual opinion statements that is supposed to be positive in one document but referred to as a negative statement in another document. A second challenge is that individuals don't continuously specify opinions within the same manner (Shukla and Mishra, 2016).

Although, one challenge with Sentiment Analysis is the classification of the polarity of any written statement at any of the three Sentiment Levels stating if the sentiment expressed in the document is either a positive sentiment, negative sentiment or a neutral sentiment.

METHODOLOGY

Literature Search Strategy

Sentiment Analysis research started in the early 2000s and various methods for the analysis and classification of opinionated text documents from social media sources like twitter have been researched on. Nowadays a special interest has arisen on the social networks such as twitter where people share their opinions about several topics (Fernandez *et al.,* 2016). Most of these efforts are based on two main approaches: The Semantic Orientation (SO) approach and the Machine Learning approach (Pang and Lee, 2004).

Devendra *et al.* (2017) proposed an approach to utilize Twitter user-defined hastags in tweets as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. In their work they also made use of the K Nearest Neighbour machine learning approach to give sentiment labels by constructing a feautre vector relating to each of the avaliable examples in their training set.

Gongde, *et al.* (2003) in their research discussed techniques for preprocessing and information retrieval with the help of TF-IDF, SVM. In their work, they studied Support Vector Machine (SVM) which is use for text categorization and also helps in finding out the polarity of textual comment. At the end of their study they concluded that Support Vector Machine (SVM) acknowledge some properties of text which include High Dimensional feature space, few irrelevant feature, and sparse instance vector. Also their study reveals that several results showed that the Support Vector machine (SVM) generated an accurate performance on text categorization as compared with Artificial Neural Network (ANN).

Ranjeeta and Vaishali (2015) provided the workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large-scale computational analysis of user-generated textual content. They compared their results for both the Support Vector Machine (SVM) and the Naïve Bayes algorithms then they did evaluate a multi-label classifier which they used in detecting the engineering student problems. Comparing their final result showed a more accurate results of the Non-Linear SVM than the Naive Bayes classifier and Linear SVM. Although they did not take into consideration the analyzing of students ‘generated content and texts which include images and videos on social media sites.

Babaljeet *et al.* (2016) concluded that the combination of K-Nearest Neighbour and Support Vector Machine produced better results based on Accuracy, Precision, Recall and F-Measure. In their study, the K-Nearest Neighbour had an improved performance in the case of small reviews while the Support Vector Machine (SVM) also improved the performance in case of large reviews as both cases were serving as a single hybrid approach. They also had a two or more parameters that had positive and negative polarities. This showed that some of their reviewers had positive thoughts towards SuperFetch and the negativity percentage was also few, they considered the remaining other reviews as neutral. Therefore, their results indicated that SuperFetch has a good feature in the memory management system. But there was no transparency in their results and it proved quite difficult to be interpreted. Andreas *et al.* (2017) presented a tool that analyses microblogging messages regarding their sentiment using machine learning techniques. Their algorithm exploits the various hashtags and emoticons present in a tweet, referring to them as sentiment labels, and further proceeds to the classification procedure of the various sentiment types. They also utilized Bloom filters to compact the storage size of the required intermediate data, boosting the overall performance of the algorithm. According to them some classification algorithms was implemented in the Apache Spark cloud framework using Apache Spark's Machine Learning library, entitled MLlib. Through an extensive experimental evaluation, they also during their study proved that the system is efficient, robust and also scalable. They did not consider the various clusters to better evaluate Hadoop's and Spark's performance regarding time and scalability.

Arockia *et al.* (2015) in their work focused on the analysis of the sentiment of passenger reviews which they got from a popular airline forum. Their study suggested that Linear Support Vector classification gave a better accuracy than the Multinomial Naive Bayes. although, during the process of run time, Multinomial Naive Bayes classifies more faster than the Linear Support Vector. Further in their work, they noticed that the number of features reduced by 36.27% during the application of various pre-processing techniques. The major limitation of their work is that the system did not provide any relevant information about the scope of the sentiment. Also, the system requires manually training the various classified reviews which is quite different from the traditional lexicon-based analysis system.

The overview of the State-Of-The-Art Multilingual Sentiment Analysis methods was given by (Migual and Padmini, 2009). Their study also analyzed data pre-processing, typical features, and the main resources that can be used for multilingual sentiment analysis. They also gave a summary of the various approaches applied by their authors to English and other languages. A detailed classification of these approaches into corpus-based, lexicon-based, and hybrid ones was also discussed in their work. According to them, their results outperform other approaches although, the approach used by them is very costly and it has only been tested more on English Language Data.

The ensemble machine learning technique is now currently used for sentiment classification of tweet. It makes use of a boosting principle. A series of experiments on sentiment classification in social media text using ensemble learning methods were conducted by (Shukla and Mishra, 2016). In their work each of the base learner in the ensemble made use of meta-level feature extraction. During their study, they investigated three datasets in other to verify the effectiveness of the present approach across the various data. This results according to their experiment showed that such ensemble classifiers can minimize the error rate by avoiding a poor selection from the stand-alone classifiers, making it a better way of ensuring stability. Additionally, using the meta-level feature mitigated problems associated with the sparsity of the data.

Silva and Hruschka (2014) in their work on Tweet Sentiment Analysis made use of the ensemble techniques referred to as the Adaptive Boosting model and their results showed that AdaBoost provides good performance in sentiment analysis.

Sentiment Analysis Techniques

The Sentiment analysis technique can be referred as the various methods used to analyze or classify the sentiments present in each text or data. They are categorized into three and they include the machine learning, the lexicon based and the hybrid approach. In the case of the machine learning approach, it applies some machine learning steps to the linguistic feature. While the Lexicon based approach greatly depends on a sentiment lexicon as well as a collection of already known sentiment which can be used to analyzed the text. The Lexicon approach is further divided into the dictionary-based approach and corpus-based approach. These two approaches use the statistical or the semantic methods for finding the sentiment polarity. The hybrid approach is a combination of the two approaches and it is often common with sentiment lexicons making a vital role in the majority of methods. Figure1 shows the various Sentiment Analysis and classification techniques.

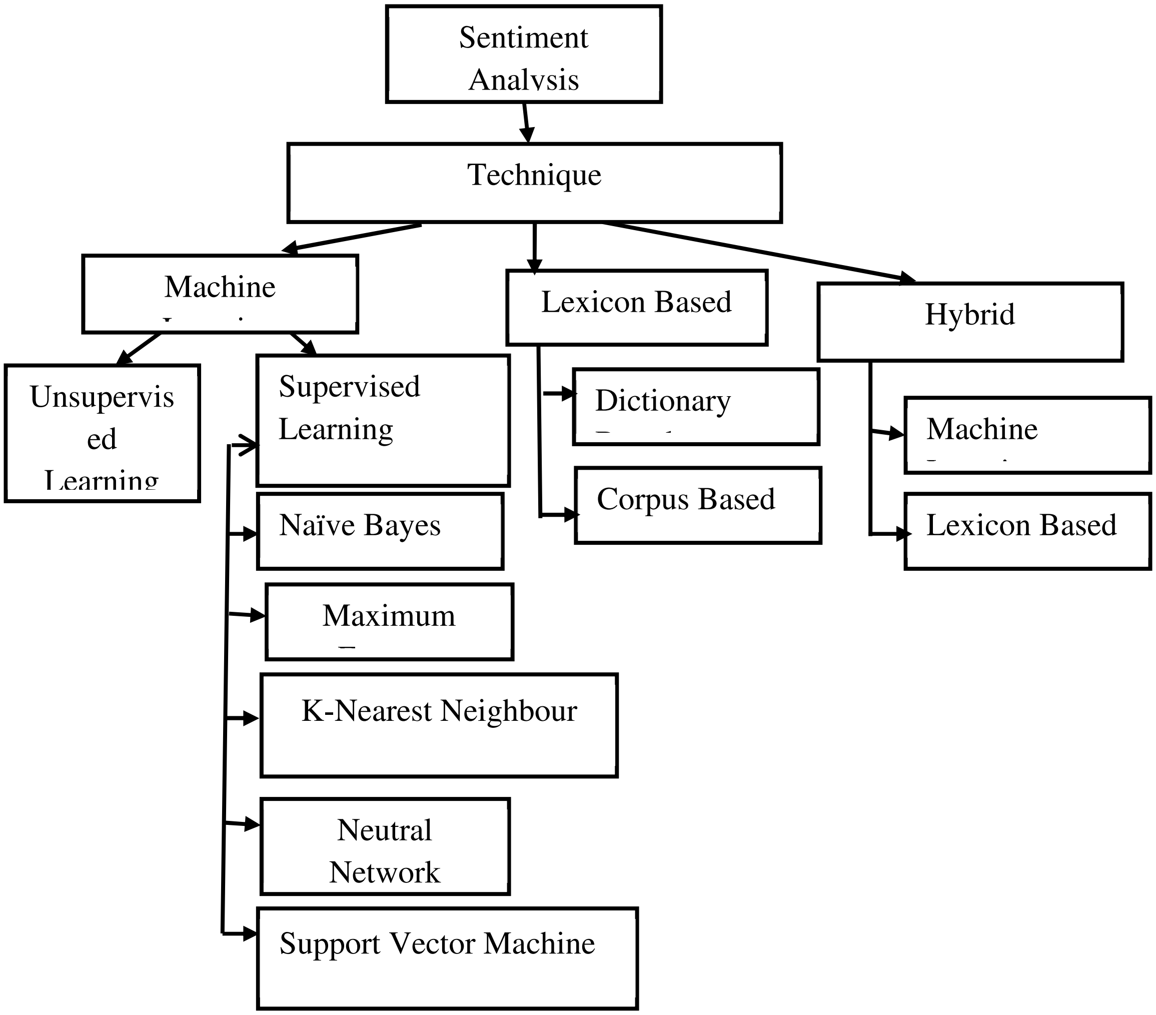


Fig. 1: The various Techniques for Sentiment Analysis

Machine Learning Approach

The machine learning approach involves the use of different machine learning steps and machine learning techniques that can be used to automatically classify textual data which can include peoples reviews or sentiments to online products, tweets or comments from social media platforms. Machine learning approach can be classified into two which are the Supervised Learning and the Unsupervised Learning (Enas *et al.,* 2021) as shown in Figure 1.

Supervised Learning

In the supervised machine learning method of classification, the algorithms tend to learn from the training data that are already labeled before making any predications. The trained model can also be able to predict situations where there are new or unseen textual data. The various machine learning classification techniques are examples of the supervised learning. They are the Naïve Bayes, Maximum Entropy, K-Nearest Neighbour, Neural network, Support Vector Machine etc.

Naïve Bayes

Naive Bayes is a classifier that works using the Bayes theorem mathematical approach. The way it works is calculating the probability of a given event occurring when given the probability of other related events. It is very efficient for text classification both for either a large data set or small data set (Ranjeeta and Vaishali, 2015). Let's take a look at Some advantages of Naive Bayes:

1. Naïve Bayes can work well even with a small amount of training data in other to estimate parameters.
2. It is very fast and also incremental; it can also deal with discrete and continuous attributes. the Naive Bayes does not give accurate results for uneven class data sets. Complement Naïve Bayes addresses this problem and has been proven to give higher results than Naive Bayes when the classes are uneven (Gokulakrishnan *et al.*, 2012). Complement Naive Bayes estimates the probability of a class using the parameters of all the classes excluding the class itself (Ranjeeta and Vaishali, 2015).

Maximum Entropy

The Maximum Entropy classifier is referred to as the MaxEnt. It is almost similar to the Naïve Bayes classifier as it also makes use of the probabilistic approach. The difference is that the Maximum Entropy does not make use of independent features instead the model attaches weights to the features using the search-based optimization. The Maximum Entropy does not make any assumptions about the relationships between the features making it a little different from the Naïve Bayes and Support Vector Machine (SVM). One drawback is that it is not very realistic in many practical problems, as real datasets contain random errors or noise which create a less dataset (Groot, 2012). According to Phillips *et al.* (2006), a challenge with maximum entropy machine learning methods is that it is not a frequently used statistical method as there are fewer guidelines for its use alongside fewer methods for deriving the amount of error in a giving prediction this makes it a rear machine learning technique. Maximum entropy technique is not available in most standard statistical packages. An example of how maximum entropy can be applied in sentiment analysis is using maximum entropy to classify reviews made by various individuals regarding a movie. Each review has to be classified based on the polarities expressed by these individuals as either positive or negative. Maximum entropy techniques model can be trained to predict these various individuals' sentiment. Various features will be considered for this classification and they include:

1. The availability of specific words that are related to the individuals positive or negative reviews.
2. The frequency (number of times) in which these specific words occur in the review.
3. The context in which the specific words are used. If a positive word is used in a negative review this can change the overall structure of the review.
4. Information about the structure of sentences, like the presence of exclamation marks, stop words or emoticons in the review.

After considering all these features training of the maximum entropy model will be done. Here the model learns the weights for the features in the training set. New movie reviews not present in the dataset can be classified using the trained model based on the already learned feature weights. If the dataset is also large there could be complexity of training the model which is another drawback of this machine learning technique.

K-Nearest Neighbour

K-NN classifier is a case-based learning algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measure's (Gongde *et al.,* 2003). This method has been applied to many applications, as it is effectiveness, non-parametric and easy to implementation properties, however the classification time is long and difficult to find optimal value of k. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification but make boundaries between classes that are less distinct.

Neural Network

A neural network classifier is a machine learning technique that is built to function like the human brain. It can be used for classification documents with a wide range of tasks. A good example for this explanation is when classifying a text document, you assign term weighs to the input units; making the activation of the units to be propagated forward through the network. The value taken by the outputs units determines the categorization of the decision. Some researchers use the single layer perception, due to its simplicity of implementation (Albaugh *et al.,* 2017).

Cheng and Soon, (2009) proposed a model using back-propagation neural network (BBNN) and modified back-propagation neural network (MBPNN) for documents classification.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a popular machine learning classification method mainly used for classification, regression, and other learning tasks. The SVM is a good tool for sentiment analysis as it is a good resistant to data set filled with noise meaning irrelevant details and can also handle a large feature data set. SVM performed better than Naïve Bayes and Maximum Entropy for sentiment classifications (Ranjeeta and Vaishali, 2015). Support Vector Machine (SVM) is used for binary classification problems, with the intention of separating the data into two main classes. It can be used to classify textual data into positive and negative polarities.

Unsupervised Learning

Unsupervised learning is a machine learning where the steps practice the patterns and structures from the input data without explicit labels. In this learning technique the dataset that is given to the learner is not labelled. Clustering is also used in the unsupervised learning. Thus, it is commonly used for finding hidden patterns and grouping.

Lexicon Based Approach

The lexicon-based approach analysis technique relies on reference to finding the feeling in the vocabulary to accurately analyze the content. It is a method that involves using a defined set of phrases to classify the information from the given text data. This is done by using applied mathematics or linguistics techniques.

Dictionary Based Approach

The main cost for dictionary-based approaches is the time spent developing and refining a dictionary (Shiva, 2016). This approach proceeds with a defined dictionary that consists of positive and negative polarities vector of words. The various independent words in the dictionary are giving a score based on its importance to the present task.

Corpus Based Approach

The Corpus Based Approach is a machine learning approach that assist in the solving of problems relating to searching the opinionated words which include all the context specific orientation words. The Corpus Based Approach is mainly used for very large data set. A Corpus can be referred to as an organized collection of textual data set.

Hybrid Machine Learning

This machine learning is a combination of multiple techniques to improve performance and accuracy of the model. It leverages on the strengths of each independent techniques to improve on the limitation of the model. The combination of hybrid machine learning approach greatly depends on the required attributes relating to the problem giving and the strengths of the independent machine learning models to be combined.

A combination of the various machine learning techniques and the lexicon combination of sentiment analysis in the hybrid machine learning approach helps to immensely increase the strengths and overall accuracy of the model’s performance. Sentiment Lexicon Construction also is vital. This is because sentiment lexicon ensures a better classification of text by matching words with sentiment scores. Also, Sentiment lexicons help in balancing the representation of different sentiment classes by providing additional information that might be scarce in the training dataset making sentiment construction very important.

Table 1: Techniques of Sentiment Analysis Classification

SENTIMENT CLASSIFICATION APPROACHES		**FEATURES/ TECNIQUES**	**STRENGTH**	**WEAKNESS**
Machine learning	1.Naive Bayes 2.Maximum Entropy 3.Neural Networks 4.Support Vector Machine	1.Term presence and frequency 2.Part of speech information 3.Negations 4.Opinion words and phrases	NAÏVE BAYES 1.Model is easy to interpret. 2.Fast and efficient computation. 3.Not affected by irrelevant features MAXIMUM ENTROPY 1.Provides proper distribution. 2. Do not assume statistical independence of random variables. NEURAL NETWORK 1.Capable of learning almost any relationship between input and output variable. SVM 1. Very good performance 2. Data set dimensionality has low dependency. 3. Produces accurate and robust classifications	NAÏVE BAYES 1.Naïve Bayes Assume independent attributes MAXIMUM ENTROPY 1.Requires more of the human efforts in the form of additional resource or annotations. NEURAL NETWORK 1.Neural network Requires more time for execution. 2.Flexibility depends on enough training data need. 3.It is somewhat considered as complex 'black box" SVM 1.Lack of transparent of results. 2.Difficult interpretation of resulting model.
Lexicon based	1.Dictionary based 2.Corpus based approach	1.Manual construction, 2.Corpus-based 3.Dictionary based	wider term coverage	finite number of words in the lexicons and the assignation of a fixed sentiment orientation and score to words
Hybrid	1.Machine learning 2.Lexicon based	1.Sentiment lexicon Construct using public resources for initial sentiment detection 2.Sentiment words as features in machine learning method	In lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level is less sensitivity to changes in topic domain	noisy reviews

Sentiment Analysis of Tweet Data Challenges

Today people express their sentiments/opinions in different and very complex ways that is usually difficult to be analyzed and classified. This can likely result to issues when making use of the sentiment analysis classification techniques. In this section, we discussed the challenges common with Sentiment Analysis.

Named Entity Extraction

Named entities refers to specific noun phrases that represent different types of individuals, such as business organizations, persons, special dates, etc.

The importance of the named entity extraction is to locate the presence of textual mentions of the named entities in a written document. Named entity recognition is a well acceptable task suited to the type of classifier-based approach like sentiment analysis. The process below explains the various process of Named Entity Extraction:

Tokenization

This refers to the process of breaking a stream of text up into independent words, symbols and other meaningful elements called "tokens" (Panigrahi *et., al* 2017). This is the first stage in classifying the review. Let's use an example from the tweets we got from twitter.

The Nokia mobile phone is a good phone for users.

This will be tokenized as:

The, Nokia, Mobile, Phone, is, a, good, phone, for, users.

Part Of Speech Tagging

The individual tokens are giving a POS tag (Part of Speech) which also tells its grammatical category that could either be a noun, verb or adjective tokens. This is very necessary for proper understanding of the tokens used in each review.

Classification

The entities are further classified into different categories. Entities can either be classified as Names, which can be individuals, as companies which can be various business types etc.

Below are sentiments from twitter on mobile phone:

Example 1:

i. The Nokia mobile phone is a good phone for users.
ii. It is not difficult to use and it is having a perfect quality.
iii. The pictures are very clear and bold.
iv. The face recognition feature is powerful which makes the images even brighter than it was before.
v. The Audio quality is pretty perfect although the hard disk quality is very low.

In the following examples above the details stated relating to the Nokia phone brand is named entity.

Information Extraction

Information is usually in different dimension and styles. Natural language complexity can result to the difficulty of accessing the information in the opinionated document. The different tools present in Natural language processing (NLP) are still not able to develop a general-purpose representation of meaning available in written documents. One of the important forms is the presence of structured data where there is a regular and predictable organization of entities and relationships. Another is the presence of a large volume of unstructured data present on the Internet. Information Extraction can be applied in large areas like, business intelligence, media analysis, legal compliance, sentiment detection, patent search record, and email scanning. In the sentiment analysis application, the information that is to be extracted are the various polarities values of the individual's opinions.

Sentiment Determination

Sentiment determination refers to the task that assigns a sentiment polarity either positive or negative to a word, a sentence, or a document. The use of the sentiment Lexicon is a traditional way for polarity assignment. In Opinion Mining, the adjectives of a sentence are very important since they have better probability of carrying information during sentiment analysis problem. Another thing that is important and helpful, is the availability of any other words in the opinion lexicon while finding the sentiment polarity. There are two major approaches which are Dictionary based approach and Corpus based approaches which are used to build up the opinion lexicon.

Co-Reference Resolution

Co-reference resolution is performed at the feature aspect level and the entity level. In a situation where opinionated text is present, we can see comparative texts. These comparative texts can include coreferences. The references must be able to produce results that are accurate and effective. Let's consider some opinionated text examples below:

Example 2: Comparing Lenovo to its competitor the Techno Phantom, it takes lovely pictures.

Here two named entities are mentioned and they are Lenovo and Techno Phantom. The pronoun 'it' in the text refers to 'Lenovo'. In case where the co-referring words are not sorted out, there won't be effective sentiment analysis. Co-reference resolution is important as it provides more information in the Information retrieval tasks. There are several anaphora resolutions factors that help in the task. Constraints and preferences are considered while carrying out this task. The scope of the resolution task is also to be defined. The scope can be a sentence, nearby sentences, or a document etc. The coreference resolution is important to the sentiment analysis problem and very complex task. The resolution problem itself is not solved yet in NLP.

Relation Extraction

Relation extraction is the task of finding the syntactic relation between words in a sentence. The semantics of a sentence can be found out by extracting relations between words and this can be done by knowing the word dependencies. This is also a major research area in NLP and serious research are going on to solve this problem. Textual analysis like POS tagging, shallow parsing, dependency parsing is a pre-requisite for relation extraction. These steps are prone to errors. Many of the problems in NLP are not fully solved because of the unstructured nature of text. Relation extraction also belongs to the group of challenging problems. The place of relation extraction in sentiment analysis is very high and thus this challenge is to be met and solved.

Domain Dependency

A sentiment classifier that is trained to classify opinion polarities in a domain may produce miserable results when the same classifier is used in another domain. Sentiment is expressed differently in different domains. For instance, consider two domains, digital camera and car. The way in which customers express their thoughts, views and prospective about digital camera will be different from those of cars. But some similarities may also be present. So, Sentiment analysis is a problem which has high domain dependency. Therefore, cross domain sentiment analysis is a challenging problem that must be unfolded.

Sentiment Analysis Steps for the Classification of Tweet

Performing sentiment analysis involves several steps which are key success to any of the machine learning techniques model. The quality and features of the labeled data is also important for the required model to achieve an accurate result. The various steps are as follows:

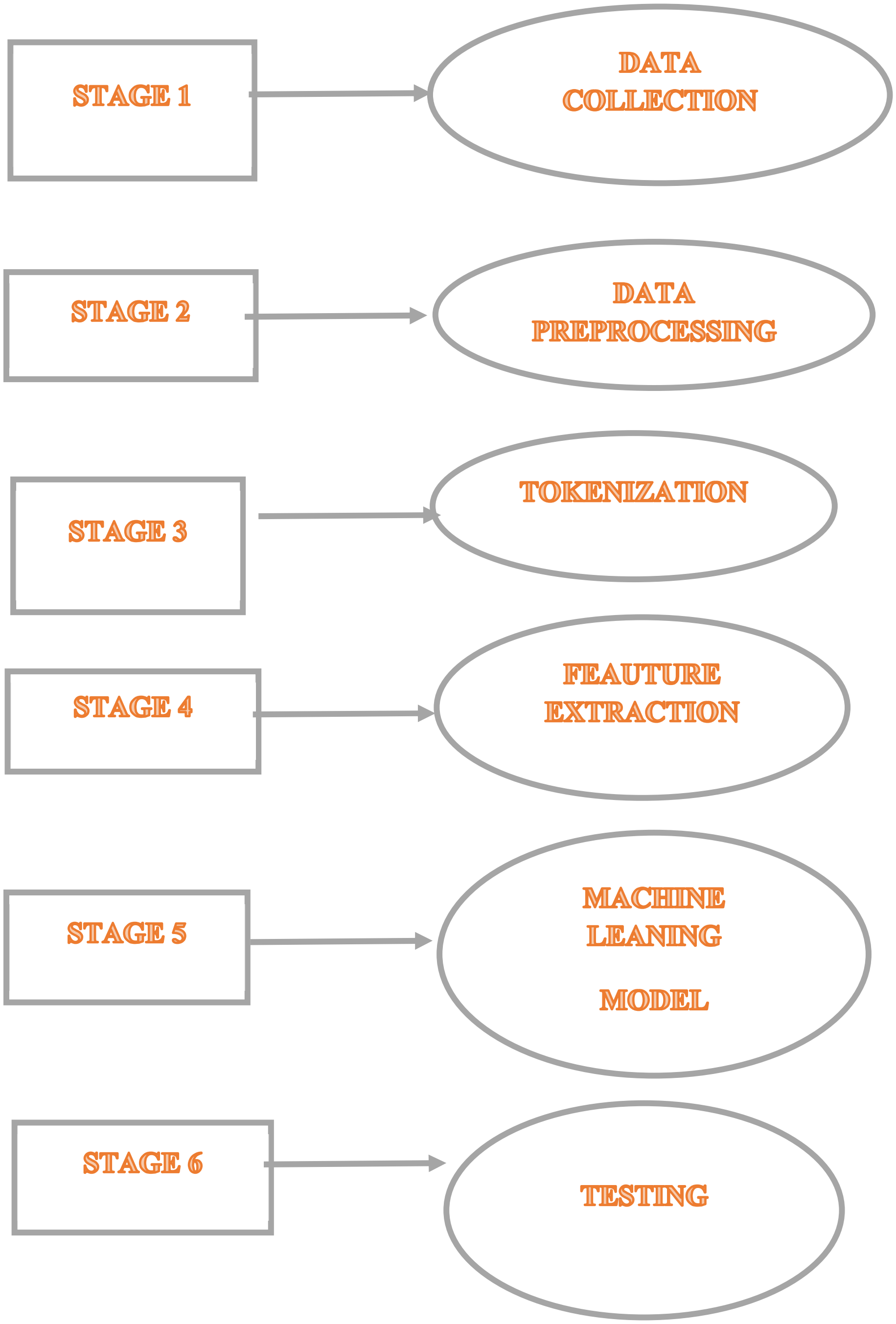


Fig. 2: Sentiment Analysis Steps for Classification of tweet

Data Collection

Data collection is the first and very important aspect in sentiment analysis. Tweets from any social media platform like twitter can be collected to make up the data set. These tweets are sometimes labeled already as either positive, negative or neutral.

Example 1: Here is a tweet on mobile phone product from twitter (this tweet was labeled as negative)
The design is very odd, as the ear "clip" is not very comfortable at all!

Data Preprocessing

Once the data has been collected, they are then preprocessed. Another word for preprocessing is cleaning, this is done to remove every form of irrelevant details which can include symbols, special characters, words not in English dictionary and stop words.

From Example 1 after preprocessing the tweet will become
Example 2: The design is very odd, as the ear clip is not very comfortable at all

Tokenization

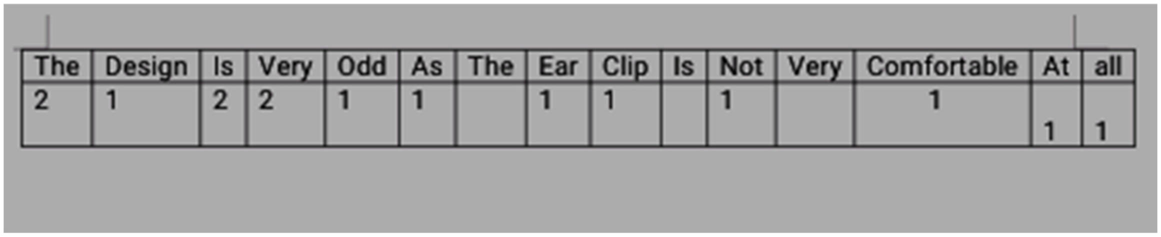
The pre-processed data is further Tokenize that is splitting the data into individual vector of words or tokens.

From Example 2: after tokenization the tweet will become;
Example 3: The, design, is, very, odd, as, the, ear, clip, is, not, very, comfortable, at, all
Making a total of 15 vector of words in the data set.

Feature Extraction

Each of these vectors of words will be numerically represented using techniques like the Term Frequency-Inverse Document Frequency or the word embeddings like the Word2Vec and the GloVe

From Example 3:



The	Design	Is	Very	Odd	As	The	Ear	Clip	Is	Not	Very	Comfortable	At	all
2	1	2	2	1	1		1	1		1		1	1	1

Machine Learning Model

The data is then ready to be analyzed using any of the preferred machine learning algorithms**.**

Testing

New tweets not present in the existing vocabulary can be trained using the chosen machine learning model to make new predictions.

SUMMARY

Sentiment Analysis has so many challenges relating to machine learning and this has been a major area for research for many years. In this study we tried reviewing several of the works done in Sentiment Analysis especially those that tried solving the various challenges. During the research we noticed that several works have been done in this area but there is no evidence of any that can be referred to as a fully automated system and that is also greatly efficient. This is as a result of the very unstructured nature of natural language. Natural language vocabulary is very large that things become very difficult. Our study focused on the various existing works relating to sentiment analysis and their numerous techniques. Different sentiment analysis techniques methods were also reviewed where we stated their strength and weakness and also include the major issues involve with sentiment analysis. These challenges are Named Entity Recognition, Coreference Resolution, domain dependency etc.

CONCLUSION

The problems of Sentiment Analysis of Tweet data have to be handled separately and also new researchers can improve on the independent machine learning approach by utilizing an ensemble boosting machine learning approach. This is a combination of the machine learning either supervised or unsupervised with any of the various boosting machine techniques to solve the problem of tweet classification.

REFERENCES

Albaugh, Q., Soroka, S., Joly, J., Loewen, P., Sevenans, J. and Walgrave, S. (2014). Comparing and Combining Machine Learning and Dictionary-based Approaches to Topic Coding.

Amreen, S. and Madhuri, R. (2017). Survey on Sentiment Analysis. *International Conference on Emanations in Modern Technology and Engineering (ICEMTE-2017)*, 5(3): 234-236.

Andreas, K., Nikolaos, N., Spyros, S., Athanasios, T., Dimitrios, T. and Giannis, T. (2017). Large Scale Implementations for Twitter Sentiment Classification. Computer Engineering and Informatics, pp. 2-21.

Arockia, X. A., Vignesh, M. and Sree, H. V. (2015). Sentiment Analysis Applied to Airline Feedback to Boost Customer's Endearment. *Full Paper Proceeding MISG-2015*, 2: 219-232.

Babaljeet, K. and Naveen, K. (2016). A Hybrid Approach to Sentiment Analysis of Technical Article Reviews. I.J. Education and Management Engineering, pp. 1-11.

Bing, L. (2012). Sentiment Analysis and Opinion Mining. Morgan and Claypool Publishers.

Cheng, H. L. and Soon, C. P. (2009). An efficient document classification model using an improved back propagation neural network and singular value decomposition. Expert Systems with Applications, 3208- 3215.

Devendra, K., Ninad, B., Harikrishnan, R. and Salil, S. (2017). Sentiment Analysis of product reviews. *International Journal of Engineering Sciences and Research Technology*, 6(1): 456-460.

Enas, A., Khalil, H., Enas, M., El Houby, F. and Mohamed, H. K. (2021. Sentiment Analysis Tasks and Approaches. *International Journal of Computer Science and Information Security*, 19(10):

Fernandez, G. M., Alvarez, L. T., Juncal, M. E., Costa, M. and Javier, G. C. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58: 57-75.

Gaurangi, P., Varsha, G., Vedant, K. and Kalpana, D. (2014). Sentiment Analysis Using Support Vector Machine. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1): 2607- 2612.

Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N. and Perera, A. (2012). Opinion mining and sentiment analysis on a twitter data stream. Advances in ICT for emerging region.

Gongde, G., Hui, W., David, B., Yaxin, B. and Kieran, G. (2003). KNN Model-Based Approach in Classification. ODBASE.

Groot, D. (2012). Data mining for tweet sentiment classification. Master Thesis, Utrecht University.

MIgual, E. R. and Padmini, S. (2009). Automatic Text Categorization Using Neural networks. *Advances in Classification Research*, 8: 267-286.

N'adia F. F., Eduardo, S., Hruschka, R. and Hruschka Jr., S. R. (2014). Tweet Sentiment Analysis with Adaptive Boosting Ensemble. Proceedings of the 8th International Workshop on Semantic Evaluation, pages 129–134

Panda, S., Gupta, S., Kumari, S. and Yadav, P. (2020), Sentiment Analysis Techniques and Approaches. *International Journal of Engineering Research and Technology*, ISSN N0: 2278-0181

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Annual Meeting on Association for Computational Linguistics.

Panigrahi, S., Momin, S., Patil, P. and Kshirsaga, P. (2017). Sentiment Analysis of Application Reviews on Google Play store. *Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2(2): ISSN: 2456-3307

Phillips, S. J., Anderson, R. P. and Schapired, R. E. (2006) Maximum entropy modeling of species geographic distributions

Rana, A., Adrian, O., Humphrey, S. and Cathal, H. (2015). Improving Sentiment Analysis Through Ensemble Learning of Meta-level Features. Computer Science, Western Gateway Building.

Ranjeeta, R. and Vaishali, K. (2015). Analysis of Students Emotion for Twitter Data using Naïve Bayes and Non-Linear Support Vector Machine Approachs. *International Journal on Recent and Innovation Trends in Computing and Communication*, 3(5): 3211-3217.

Shiva, G. (2016). Sentiment Analysis for Movie Review. *International Journal and Magazine of engineering, Technology, Management and Research*, 4(1): 10-24.

Shukla, R. and Mishra, N. (2016). Framework for Sentiment Analysis of Twitter Post. *International Journal of Innovative Research in Science, Engineering and Technology*, 5(3):