A Machine Learning Approach to Multilingual Sentiment Classification

Dr. Elena Petrova, Dr. Tomás Valverde, Dr. Aisha Mohammed
Department of Geology, Saint Petersburg State University, Russia;
Department of Earth Sciences, University of Seville, Spain;
Department of Environmental Studies, Cairo University, Egypt

ABSTRACT

Sentimental Analysis or Opinion Mining is increasing in today's world. As every industry, require the opinion or review from the end-user or consumers to improve the quality of their product or service. The previous research work has been done on sentimental analysis is only limited to English data analysis and secondly the on the enhancement of accuracy by using different algorithm. In this research we overcome both the issues by developing a language-independent system using Google translator API and proposed a solution with Stanford NLP which the modeling of the English language. The simulation of the proposed work has been done in two stages. First, all the tweets are translated into a single language. English is used as the target language here. To make the research database broader, we have used the Live Google translation API that will convert all the tweets into the English language. Then, emoticons, slang language, misspellings, email id, URLs, etc. are forced to preprocessing before feature extraction. In the second phase, the stop words, which do not contribute towards the sentiment of the tweets, are removed, and the tweets are converted into feature vectors. This feature vector is then used in the classification algorithm. The results are compared using two classification techniques, i.e., Naïve Bayes classification and RNN.

Keywords: *Opinion mining, Translation, Naïve Bayes, RNN.*

I. INTRODUCTION

With the advent of 21st century and massive boom on the Internet, more and more people are using social media platforms like Blogs, Facebook, Twitter and much more to express their views on topics of their choice. With this massive number of people continuallyshowing their opinions every second, it is impossible for any human to analyze the opinions physically. Techniques like Machine learning, artificial intelligence are now being used to analyze the views of the masses. Companies can use these opinions for market research for the need of a particular product or for checking public views on a newly launched product. Marketing intelligence for any company acts as a guiding path to understand the customer's needs and to further improve the planning, control, and implementation. There can be many different feature extraction techniques for collecting the customer's response from social media platforms. To pick the most accurate one has been the call for any marketing strategist. Hence, Sentimental analysis appears. It helps marketing strategist analyze the people's mindset about a product or an idea. It needs a large amount of data for analysis. Sentiments or opinions about an issue, person or event can be only as accurate as the number of persons participating. Hence, Data gathering and classification are two very crucial tasks in sentimental analysis. In this project, the data-gathering step is emphasized. It uses Twitter as a source of the required data as it has become the most favorite platform for people to express their views. With the limited set of 140 characters, users on Twitter use all kinds of emoticons, slangs to show them better in the limited set of available styles. This is a problem when you talk about sentimental analysis. However, the results are more accurate as the data is more authentic. Preprocessing is vital andhence it is divided in two phases. In the first phase, all slangs, emoticons are removed and tweet is converted into plain and simple text. In the next phase, stop words are removes, and the feature vector is extracted. The test set is generated manually so that polarity can be defined more accurately. The classification techniques hence give better results. There are two classification techniques used for this purpose, i.e., Naïve Bayes classification & RNN.

1.1 WHAT IS SENTIMENT ANALYSIS

Ethical analysis (emotional mining, emotion classification, opinion extraction, personality analysis, mining review, evaluation extraction and in some cases, polarization) analyzes computational analysis of opinion, emotion and self expression in the text. He intends to explore the position or opinion

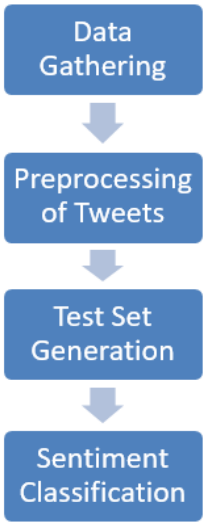
of the speaker or writer in connection with a specific subject or purpose. The situation can show his or her decision, opinion, assessment or emotional status (how does the writer feel during writing) or desired. [2] Define private state as the period of general coverage of opinion, assessment, emotions and predictions. There are three main types of self-expression: personal states (for example, "boil with anger"), and reference to events of speech (or writing) which are private states (e.g., speaker ") and self-expressive elements (see below).

Twitter is a word of social media that describes the content created by the user and can be shared online with others. Blog, Wiki, Social Networking commentary and various other types of platforms can be included [14]. Due to the increasing reach of broadband, availability of powerful street computers and new website technologies, there has been a significant increase in Twitter usage, which makes content sharing easier. Millions of visitors use the Twitter Platform. There are several thousands of social media platforms, each of which has its own purpose, pattern and demographic characteristics. Before deciding which platform to use, think about what your target audience is, and which platform is most suitable. You can do this by looking at the current data on the demographic usage of social media and just think about whether the audience will actually interact with the media type [11].

Social networking platforms can be used to promote research, events and other activities. However, this promotion should be done with caution: the new users will annoy Most online communities just advertising about things. Instead, encourage yourself by joining groups by providing interesting content (making multimedia, question questions from others in discussion forums).

1.2 MODULES OF SENTIMENTAL ANALYSIS

The sentimental analysis works in four modules, which is shown in the figure 1.



II.
Fig. 1. Block Diagram for sentimental analysis

III.

1.1.1 Data Gathering Module: In this module, we use LinqToTwitter API (.NET framework) to extract the data from the twitter. To access the twitter data, we have to create Twitter APP in dev.twitter.com. Once the app is creating successfully, Generate the credential for access the data of twitter social site. This credential is used in C# program to connect with twitter data.

1.1.2 Pre-processing of Tweets: The Objective of this module to obtain filtered data from the large amount of data. To obtain filtered data we used much operation like Hashtag removal, Regular expression, NLP stop word, Retweet, URL & username. After processing, we found the processed data.

1.1.3 NLP Stanford& Google Translator: In this module the module, we apply the Google translator API as similar we used linqtotwitter API (C#). Once GOOGLE translator applied on our dataset. Filter data convert into the English language. In this English data, we apply Stanford NLP library, these libraries is used to create a modelling learning for our machine.

1.1.4 Sentiment Classification: In this final module, Naïve Bayes and RNN is used as a classifier model as well as Predication module of both the algorithms is used. RNN is acquire better output as compare to Naïve Bayes while Naïve Bayes have less space and time complexity [10].

This paper is organized into the following sections. The first section contains the introduction. Section II: Review of literature: provides an overview of the rich morale and analysis of research in this field. Section III: Proposed Work: Introduces the proposed technique and describes a model and how to evaluate emotions and order. Our description of USP is for emotional analysis. Section IV: Implementation of proposed work. Section V: Analysis of Results: The results of our pilot results are presented with comparative current technologies compared to our proposed approach.

II. LITERATURE SURVEY

Shukla et al. provided a control tool on the quality of the text based on observations on scientific papers. Method of collecting annotation emotions in two approaches. Calculates all the annotations producing the documents and calculates the total numbers. His problem is that there is a connection between complex annotations. There should be a large knowledge base of technology for querying metadata [1].

Casper et al. The Internet Opinion Mining System was proposed for review of the hotel. Paper presented an evaluation system for online user reviews and comments to support quality control in the hotel management system. It is able to detect and recover online reviews and deal with German reviews. It has a multi-subject area and is based on a multi-polar assortment; the system can recognize neutrality, for example, neutral and multi-disciplinary conditions have been defined in its body to classify emotions as polarity [2].

Goodball, et al. analyzed the feelings of news and blogs. In the context of its specific work (sense analysis of news and blogs), the previous work is divided into two categories. In the first category, there is a technique and second class system to automatically create a dictionary of emotions, which reduces the morale of the whole documents [3].

Bhavitha et al.focuses on the several machine learning techniques which are used in analyzing the sentiments and in opinion mining. Sentimental analysis with the blend of machine learning could be useful in predicting the product reviews and consumer attitude towards to newly launched product and presents a detail survey of various machines learning techniques and then compared with their accuracy, advantages and limitations of each technique. From the survey, they can conclude that supervised learning methods like Naïve Bayesian and Support Vector Machine are considered as standard learning method. Support Vector Machine provides excellent accuracy as compared to many other classifiers Lexical based approaches are ideally aggressive because it requires manual work on document. Maximum Entropy also performs better but it is suffered from over fitting. Many researches implemented opinion mining different techniques but still there is a need of automated analysis, which addresses all the challenges of sentimental analysis simultaneously. A more innovative and effective techniques required to be invented which should overcome the current challenges like classification of indirect opinions, comparative sentences and sarcastic sentences. proposed method includes the creation of a word of the word discrimination to manual selection and to tag words with several thousand letters [4].

Corpus-based approaches examine to include primary words based on large sets of text or looking for relevant labels by looking at local barriers based on Argamen et al. Instead, people find encrypted knowledge in WordNet as relationships (synonyms, contrast, and hypnosis) and meaning. Self-discovery research assumes that in the polarization of emotions often require input documents according to many instruments and applications. Decisions need to be made on the given document in which the personal information is included or not, or to identify parts of the self-document [5].

Hailong Zhanget et al. proposed that Opinion word is used in many emotions classification tasks. Positive opinion words are used to express some desired states, while negative thoughts are used to express unwanted states. There are rhetoric sentences and idioms, which are said to be together as the dictionary of opinion. There are three main approaches to compiling or collecting list of opinion words [6].

Atzioni, et al. examined the effect of the character's ability and ability to self-scale bulk scale. The goal that tells us about a particular sentence is whether it is personal to judge the qualities that appear in that sentence [7].

The personal sentence expresses some emotions, thoughts or beliefs. Instead of personal words, each sentence is parsed in a particular document and verified for being subjective. Where necessary, subjective sentences can be classified as positive or negative. Pang and Lee use a self-detector to remove the objective sentence from a particular document. Then, using the minimum deduction formula, they add relevant information at the wholesale-level level with traditional bag case facilities. They report significant improvements in a basic vector word workbook. Researchers presented recurrent neuroscience models: Vector representation and classification. They used the semantic relationship model to analyze and evaluate online emotions. Their style provides a multi-disciplinary field. However, emotion requires extensive training and assessment resources, which are under supervision [8].

There was a search of opinions on Jane and Ho, so how to use social media to bully anyone. The idea of research in global social networking platform crawling like YouTube, has the ability to search content and conversations with the aim of radicalizing those who have little or no interest in violent jihad. His work examines the approach that is truly fruitful. He has collected a large data set from a YouTube group, which has been identified as a radical agenda [9].

By automatically naming data collected from web sites on the Internet, the researchers addressed the relevant work of discovering the polarity of morale in review through supervised learning methods. The interesting thing is that our main experience of this work is that human beings cannot always have the best meaning of choice for discriminatory words. While experimenting with a series of different features in Brody and Elahad's previous research, his primary focus was not on engineering facilities [10].

III. PROPOSED WORK

The aim of this thesis work is to investigate and discover efficient algorithms and a method that can use to convert maximum data into once language so that we can fill the research gap. One more thing we added here for enhancement of accuracy, which is RNN algorithm for machine learning.

a. Investigate feature selection methods for text classification.
b. Apply Google translator for maximum data analysis.
c. Apply Stanford NLP for modelling training.
d. Apply RNN and naïve Bayes for Classification

Data can be removed from Twitter in relation to a particular domain, but this data will take a long time to rate. Another disadvantage of available Twitter data sets is that there is noise data that should be removed. The following research methodology is used to complete the research work.

a) We have collected a corpus of positive, negative and neutral tweets with the help of LINQ to Twitter API from Twitter. The size of our corpus can be enormously large.
b) We then apply all pre-processing step in order to filter dataset.
c) Apply Stanford NLP for data modelling and training.
d) Apply RNN & Naïve Bayes for machine learning

The proposed algorithm for data processing is as follows

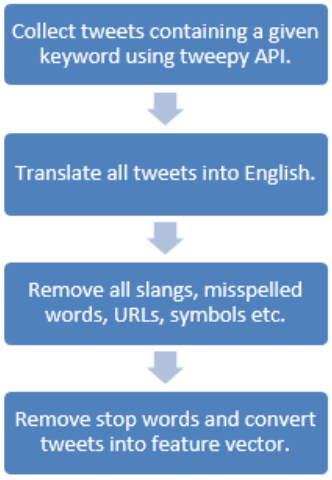


Fig. 2. Algorithm for preprocessing of data

3.1 Performance Measure

To calculate the accuracy of classifier we required measure on which accuracy can be obtained. There are two measures on which accuracy can be dependent:

a) Precision
b) Recall
c) F-measure
d) Accuracy

Precision: It gives the exactness of the classifier. It is the ratio of number of correctly predicted positive reviews to the total number of reviews predicted as positive.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Recall: It measures the completeness of the classifier.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure: It is the harmonic mean of precision and recall.

$$\text{F} - \text{measure} = \frac{(2\times\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}}$$

Accuracy: It is one of the most common performance evaluation parameter.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

IV. IMPLEMENTATION

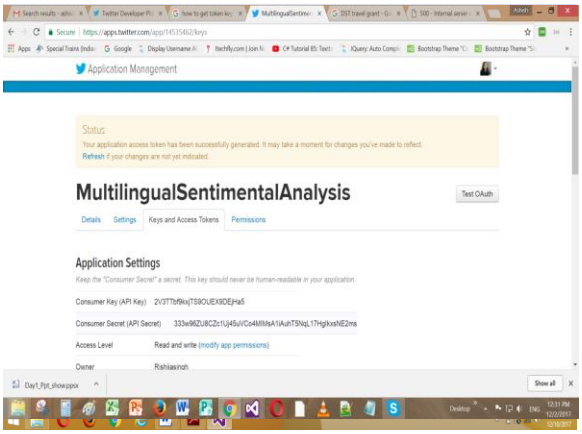
In this part of the message are the execution steps that are done on C # (.NET Framework 3.5) which have been used and used to perform more accurate analysis of the identities. The building initially analyzes emotions based on modeling, but the approach is later changed into existing learning styles. Analysis was made in a flexible way so that it could be used, even if the topic of analysis changed. This makes modeling training data more difficult than using training data on the subject.

4.1 Data Retrieval

When you see the difference between social media platforms and how to extract data from them, getting data from Twitter can be the fastest and easiest option for the timer. Twitter has a great API for developers, and a single feed contains lots of information. Tweet is a short message sent using Twitter, which can include text and / or media. A tweak can be up to 140 characters long, but metadata is very high. Tweet is usually an embedded text that contains two extra pieces of primary data, two units and space. Entries can be associated with a tweet that is user (user), tick tag (#), URL (http: // ...), and media. Places are real-world locations like school, cafeteria, city or country.

Fig. 3: Twitter4j libraries

We are extracting tweets from the twitter with the help of the Java API called Twitter4j. It consists various number of libraries that are used in the extraction. At first we have added this library into our java project. Then with the help of twitter app we have obtained Consumer Token Key and Access Token Key. Further, extraction of tweets will be start only after when we generate Access Key. Generation of Access Key needed every time for the extraction of the tweets. The twitter4j containing libraries are shown in figure 3. The purpose of using Twitter API is to select tweets from the broadcast platform, where a specific label or word is used. OAuth with REST allows you to read and write program data in practically. Due to the nature of the analysis, Stream API is required, which continuously provide new feedback to REST API queries. REST API responses are available in the JavaScript object notation (JSON) format. We have made twitter app to generate the consumer token key and access token key. figure 4 shows the generation of consumer key and figure 5 shows the generation of access token key.



V.

Fig. 4. Consumer Token Key

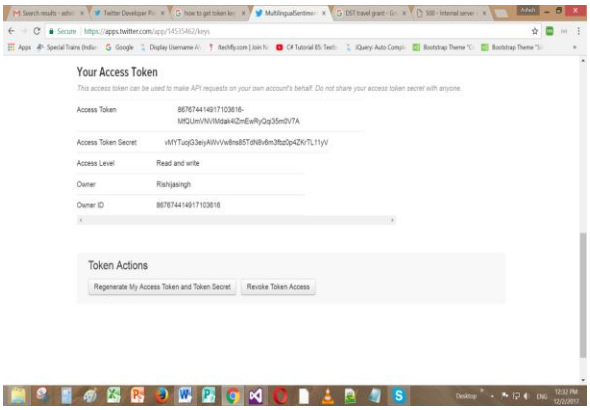


Fig.5 Generation of Access Token Key

Consumer Token Key will be provided by the twitter app. There is a unique key for every app and that key known as Consumer Token Key. In order to obtained tweets, we have to apply consumer token key and access token key into the .NET code. In figure 6, we develop tweet download module, which use twitter access key and call the twitter URL with twitter access. In return, twitter provides unfiltered data, which we handle in, excel file. We define the required no of tweet in the C# code. According to requirement program generate request to the twitter and twitter return for the same. Once the define limit meet, Program show above message as alert that download complete.

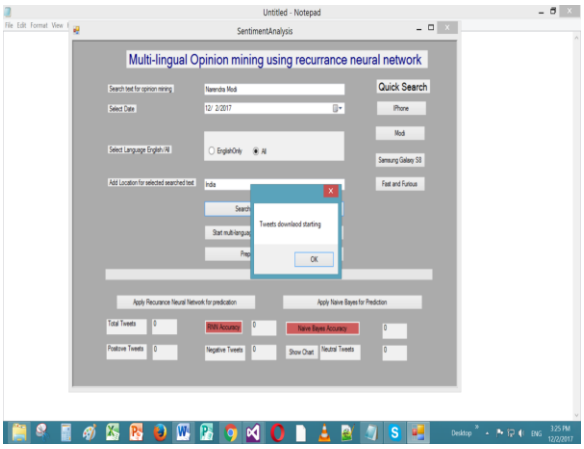


Fig.6 Tweet Download

When the incoming JSON represents, the total content of the feed in the incoming feed is more than 5K. This means that the amount of data is 40 times more than the 140-character tweets.

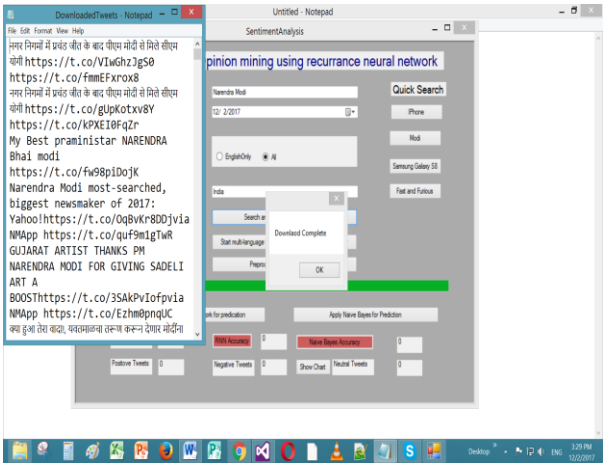


Fig. 7. Display Multilingual Data

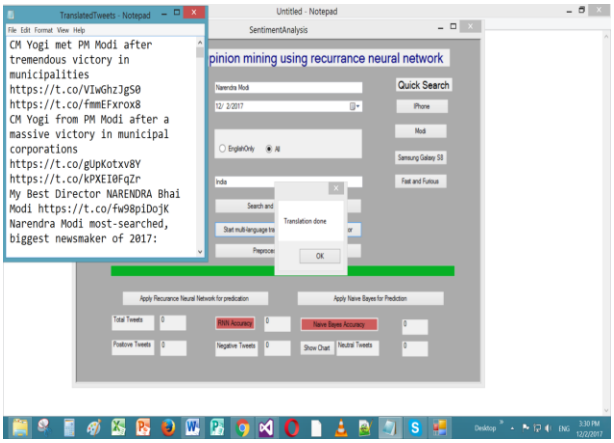


Fig. 8. Apply Google translator

In this module, we apply google translator, which is actually unique work in our project. Actually downloaded data was in multiple language and we cannot develop NLP for every language. NLP for every language will increase the ambiguity and increase searching time, which is actually not feasible. Therefore, as alternate we used google translator API that translate all the data into English. Once the data converted into English then we apply NLP for further processing.

4.2 Preprocessing of Data

After the data collection, R language is used for the preprocessing as shown in figure 9.

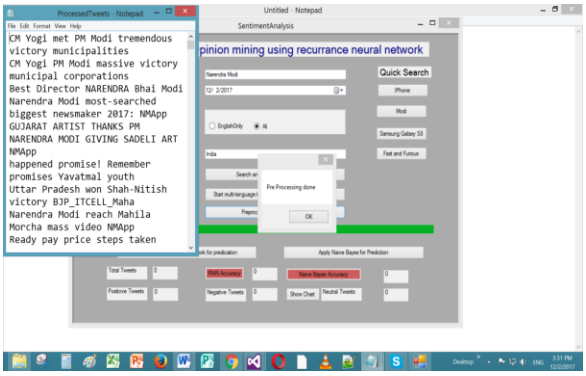


Fig. 9. Preprocessing of data

4.3 Apply Classification

Choosing the right workbook and classifying selected text attributes are the last part of the training process. As the previous studies have shown, there are several possible options for text libraries and the selection is entirely dependent on the data that requires analysis. When data is a social media, where people use words like smileys, abbreviations and colloquial words, for example, books are different from reviews.

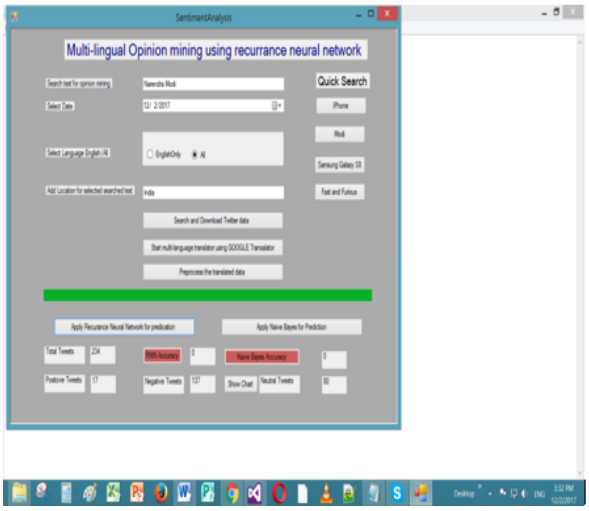


Fig. 10. Apply RNN

The text includes more typos and things that some people cannot guess. An example of such a statement is "Yolo" used by teenagers or a shortcut for some TV programs such as "TVOF" (Voice of Finland). In this research, classification has been divided into three separate chapters; Positive, Neutral and Negative in addition to these basic emotional categories, there were chapters for more specific feelings such as happiness, laughter, cry / tears, ridicule, and hatred. There is a classification for small groups in the background, but the ambition is that one day the text can be classified into more precise categories of morale. In the phase RNN is applied, RNN will provide the classification result of the data, which is the objective of this thesis. RNN classify the data into Positive, Negative and Neural. Which help us to find the actual behavior of the data. Apply Naive Bayes in above mention screenshot. This is similar to RNN in which we trying to find the classified output of the data.

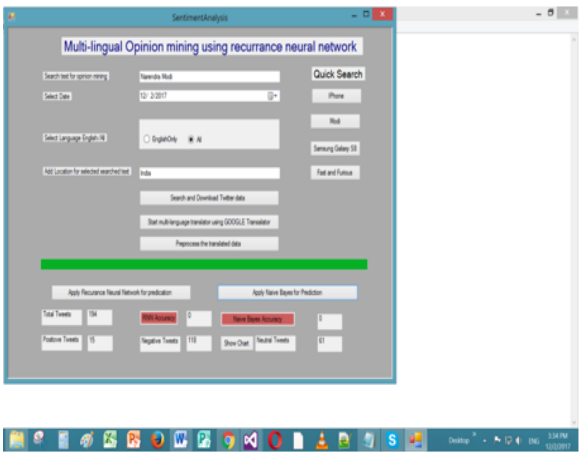


Fig. 11. Apply Naive Bayes

V RESULT ANALYSIS

In this section, the numerical results of two proposed algorithms are presented and discussed. For analysis and performance measurement, commercially available Visual studio software is used. Mainly we have considered accuracy, recall, precision and f-measure to judge the performance of the algorithms.

5.1 Results of First Proposed Algorithm (RNN).

In the figure 12, Pie chart is representing the classification result in which red one is represented with Positive, Blue is represented as Negative and Maroon is representing as Neural. Misclassified data is representing by Navy Blue.

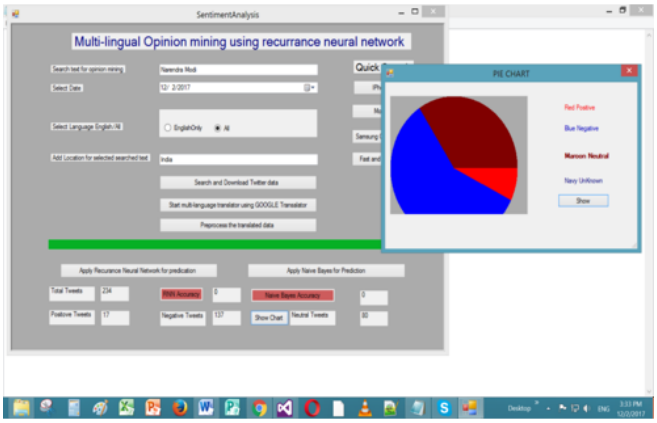


Fig. 12. PIE Chart representation for RNN

Table 5.1 show parametric result of RNN with respect to accuracy, precision, recall and f-measure.

TABLE 1. RESULT OF PARAMETERS FOR RNN

Classification of Twitter Data					
Algorithm Used	**Dataset**	**Accuracy(%)**	**Precision**	**Recall**	**F-Measure**
RNN	2000	96.15	92	90.2	90.98

5.2 Results of Second Proposed Algorithm (NB).

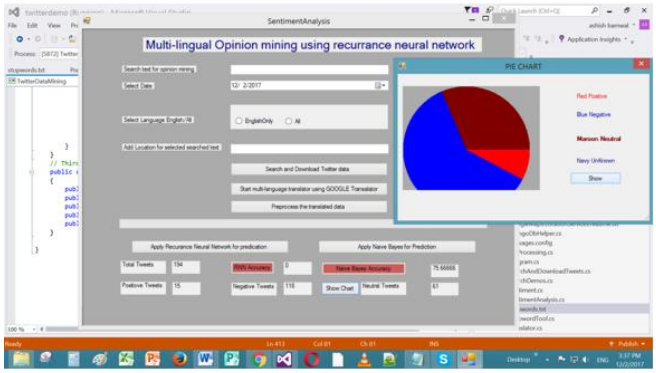


Fig. 13. PIE Chart representation for NB

TABLE 2. RESULT OF PARAMETERS FOR NB

Classification of Twitter Data					
Algorithm Used	**Dataset**	**Accuracy(%)**	**Precision**	**Recall**	**F-Measure**
RNN	2000	77.16	78.12	74.5	76.15

VI. IMPLEMENTATION

In our study, we tried to classify the emotional analysis of twitter tweets using automated teaching techniques. We implemented two algorithms called Naive Bayes (NB) and RNN algorithms. The results obtained are also compared to the results of the previous implementation of these algorithms. Notes show that the RNN class improves other workbooks in the emotion prediction work on approximately 96% accuracy. Although the time and space complexity is more in case of RNN. In this parameter, naive Bayes can be considered as the good algorithms. RNN learning model is great and performing is increase due to presence of Stanford NLP library. This Stanford NLP library having rich knowledge of English modeling. This modeling help our model to perform as compare to any other possible model which we are review in the thesis. As we discussed at the starting of the thesis that this project focus multilingual data analysis, in order to solve the problem, we apply Google translator which perform really good, around 97% we classify which is a very good jump in case of sentimental analysis.Twitter has a limit of 140 characters per tweet and is used by a large no of people to express their views so it provides result of better quality. Other than Twitter, we intend to expand our research work for other social media platforms too like Facebook, Instagram etc. Also, other steps can be included in pre-processing like considering emoticons and domain name of URL etc. If quality of data is improved, classification algorithms will be able to produce better quality of results. Future work will include studying other methods of pre-treatment treatment since it should be filtered more accurately, accurately, and so much more accurately.

VII. REFERENCES

1. Shukla, "Sentiment analysis of document based on annotation," arXiv preprint arXiv:1111.1648, 2011.

2. W. Kasper and M. Vela, "Sentiment analysis for hotel reviews," in *Computational linguistics-applications conference*, 2011.

3. N. Godbole, M. Srinivasaiah and S. Skiena, "Large-Scale Sentiment Analysis for News and Blogs.," *Icwsm,* vol. 7, pp. 219-222, 2007.

4. Bhavitha B K, Anisha P Rodrigues, Dr. Niranjan N Chiplunkar “Comparative Study of Machine Learning Techniques in Sentimental Analysis” International Conference on Inventive Communication and Computational Technologies (ICICCT 2017) 978-1-5090-5297-4/17/$31.00 ©2017 IEEE, pp. 216-221.

5. S. Argamon-Engelson, M. Koppel and G. Avneri, "Style-based text categorization: What newspaper am I reading," in *Proc. of the AAAI Workshop on Text Categorization*, 1998.

6. Z. Hailong, G. Wenyan and J. Bo, "Machine learning and lexicon based methods for sentiment classification: A survey," in *2014 11th Web Information System and Application Conference*, 2014.
7. B. Pang, L. Lee and others, "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, vol. 2, pp. 1-135, 2008.
8. B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004.
9. W. Kasper and M. Vela, "Sentiment analysis for hotel reviews," in *Computational linguistics-applications conference*, 2011.
10. H. Kanayama, T. Nasukawa and H. Watanabe, "Deeper sentiment analysis using machine translation technology," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004.
11. W. Jin, H. H. Ho and R. K. Srihari, "A novel lexicalized HMM-based learning framework for web opinion mining," in *Proceedings of the 26th annual international conference on machine learning*, 2009.
12. G. Jeh and J. Widom, "Mining the space of graph properties," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.
13. B. Heerschop, P. Iterson, A. Hogenboom, F. Frasincar and U. Kaymak, "Accounting for negation in sentiment analysis," in *11th Dutch-Belgian Information Retrieval Workshop (DIR 2011)*, 2011.
14. M. A. Hearst, "Direction-based text interpretation as an information access refinement," *Text-based intelligent systems: current research and practice in information extraction and retrieval,* pp. 257-274, 1992.
15. V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, 1997.