Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

PSAT-BASED SENTIMENT ANALYSIS: FOR TEXT AND DATA MINING

IRFAN ALI KANDHRO,
Department of Computer Science, Sindh Madressatul Islam University,Karachi, Pakistan

FAYYAZ ALI,
Department of Software Engineering, Sir Syed University of Engineering and Technology, Karachi Sindh, Pakistan.

ASADULLAH KEHAR,
Institute of computer science, Shah Abdul Latif University, Khairpur, Sindh, Pakistan

RAJA SOHAIL AHMED LARIK,
School of Computer Science and Engineering, Nanjing University of Science & Technology, 210094, P.R. China

MUHAMMAD SAMEER AHMED SHEIKHA,
Department of Computer Science, Sindh Madressatul Islam University,Karachi, Pakistan

JAVERIA TAUFIQ,
Department of Computer Science, Sindh Madressatul Islam University,Karachi, Pakistan

Abstract

In this paper, we have developed a preprocessing and Sentiment analysis (SA) tool (PSAT) that will be used to analyze the sentiments of peoples, cricket audience, former cricketers, and other sports personality on abandoned tour of Pakistan by New Zealand reasoning security concerns. This paper focuses on data cleaning and analyzing the sentiments from textual data. The data is collected from different social websites, Facebook, and tweeter. In which the user can choose a topic and specify their preferences. The model uses recent linked tweets to detect the polarity (negative, positive, both and neutral) of the issue and displays the findings. Around 3000 Arabic tweets were randomly selected and evenly labelled to train the programmed. In this research, we offer a novel technique that uses a combination of parameters to apply sentiment analysis of cricket text tweets and comments. Those parameters are (1) the time of the tweets, (2) preprocessing methods like stemming and retweets, (3) removing whitespaces, (4) Capitalizing. The PSAT tool combined with Naive Bayes classifier a group of classification algorithms based on Bayes' Theorem. The Accuracy PSAT tool is 75% approx. and F1 Score is 69%. According to our experiment, The Naive Bayes machine learning approach is the most accurate at predicting topic polarity. The tool is excellent for intermediate and advanced users, and it can assist them in determining the ideal parameter combinations for sentiment analysis

Keywords: Sentiments Analysis, Cleaning, Tool, PSAT, Naïve Bayes, Confusion Matrix and F1 Score.

1. INTRODUCTION

The Internet has changed the way individuals share their contemplations and assessments today. It is presently achieved basically through blog posts, on the web gatherings, item audit sites, online media, and other comparative mediums. Recently many peoples utilized long range informal communication destinations like Facebook,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

Twitter, Google Plus, and others [31]. The research is based on the recent event of cancellation of New Zealand tour of Pakistan in 2021 One Day International. So, research will be to acquire the sentiments of peoples across the globe that what they think about cancellation of this tour. After collecting the data through scraping tool, we have developed the data cleaning tool using C# .net technology which will clean the data and preprocessing and it gives the results as positive view, negative view and neutral in the shape of graph and resulted data [34]. In PSAT, we will input the excel file and then system will load the excel file then there will be options of cleaning the data as removing stop-words, removing white-spaces, Capitalizing first letter and more etc.[22] Using this options one can preprocessed the data according to the need then the data will be further processed in which graphs of the result will be generated that what are the sentiments of the peoples in neutral, positive and negative value [33]. The Naive Byes performed better than other classifiers in terms of accuracy, and that their performance was better with balanced datasets [34].

The most popular message trait tool is sentiment analysis, which examines the approaching message and assumes that the proposed message is positive, negative, or neutral [36]. Sentiment analysis is message mining that seeks and focuses on emotional data from source material, observing web-based conversations to help an association understand the social feelings of its image, object, or organization. Allows [38]. In field of information mining, the sentiment analysis is usually concerned with deciding whether a particular message is sentimental or objective, and if summarized, regardless of whether it is frustrating or positive [40]. The most online media streams are limited to analysis, then, sentiment analysis and numerical measurement [37]. SA aims to understand and categorize these emotions [7]. It's like starting to expose the things below and getting rid of the superficial revelations that are just about to be discovered.

Analysis of sentiments Extremity (good, pessimistic, and nonpartisan) as well as sentiments and feelings (irate, cheerful, tragic, etc), earnestness (not prompt, quick), and surprisingly expectation (premium v. Not interest) [9]. You can characterize and tweak your classifications to address the issues of your passionate analysis, contingent upon how you need to decipher client criticism and queries [2]. The September and October 2021, the New Zealand cricket crew visited to Pakistan to play three One Day Worldwide (ODI) and five Twenty20 Global (T20I) matches. The whole visit was canceled when New Zealand Cricket (NZC) raised a security alert with on the morning of the principal ODI match, the Pakistan Cricket Board (PCB) and the Pakistan Government. New Zealand's cricket visit through Pakistan was canceled on Friday, not long before the Dark Covers' first Match in Pakistan in 18 years, because of safety worries that perplexed the hosts.[39] New Zealand Cricket says it got an administration security notice and selected to drop the match and drop the visit not long before the one-day global series in Rawalpindi was booked to start. The two groups remained at their inn [34]. The NZC declined to remark on the security danger. Experts are worried that the Taliban's triumph in adjoining Afghanistan could energize psychological oppressor bunches in Pakistan [38]. The motivation behind this task is to remove web-based media includes and examine client

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

input as positive, negative, or neutral [35]. The Social media users sometimes express opinions about different topics. Some of the decisions and like a normal people or audience views about anything rely on these viewpoints [10]. Before making a purchase, purchaser can utilize these techniques to research properly about the thing they are purchasing. E.g., Kindle marketers can utilize this to find out what the public thinks about their company and products, as well as to assess customer happiness., for example, Election polls [3]. This can also be used by businesses to get critical feedback on issues with newly released items. Take, for example, big business-like Puma etc. [19]

2. LITERATURE REVIEW

In past the people have done some work initially on this which were some of the people have used sentiment analysis using corpus in which the language problems were the limitation it works on multi-language processing using Salience and ME in which Salience give better accuracy. [1] Secondly, Sentiment Analysis has been done using SVM and maximum entropy on teacher’s review as data set in which limitations where it was less effective on negative comments and SVM has given better results [3]. Some have done Sentiment analysis using deep feed forward ontology where twitter reviews were taken as dataset.

Before, some work has been finished with these procedures and instruments, for example, distinguishing disdain discourse on Twitter, which is fundamental for applications, for example, eliminating hazardous occasions, making artificial intelligence chatbots, and suggesting of content. To do, and passionate analysis [10]. Task is characterized as the capacity to arrange a tweet as bigoted, chauvinist or not [34], troublesome because of the intricacy of the normal language structure [12]. To deal with this intricacy, they do broad exploration with different profound learning designs to gain proficiency with the semantic inserting of words. In the benchmark dataset of 16K explained tweets, and profound learning techniques outflank the most recent four-word engram calculations by 18 F1 focuses [18]. Passionate analysis is a kind of shallow printed semantic analysis [36], and then analyzed the individuals' considerations, sentiments, and mentalities about various things of interest. For instance, one may be keen on shopper view of commodities, citizen perspectives toward ideological groups, or securities exchange assumptions. [19] With the ascent of the web and web-based media, enthusiastic analysis has gotten a great deal of consideration since its origin during the 2000s [32]. An assortment of literary data (for instance, news, websites, audits, Facebook remarks, posts Twitter, and so forth) With its far and wide accessibility, numerous techniques for investigating feelings have arisen [21]. AI and lexicographic procedures are the two most normal techniques for huge scope enthusiastic analysis. [1] Feelings in a text are first determined utilizing a mix of passionate words distinguished in the text. In the subsequent model, the grouping of feelings is first demonstrated, utilizing a lot of feeling named text, and afterward applied to a progression of unlabeled text. [3] The model appears as a capacity that changes over text highlights into enthusiastic marks (which typically have discrete qualities: negative, impartial, or positive). [5] The two

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

techniques expect basically a lot of human information first. People should have the option to distinguish the importance of their feelings in a solitary word or sentence. [20] This characterization of feelings is language, area, and issue explicit. [34] Dodd, and so forth Give an illustration of a word reference-based procedure that includes grouping an immense human feeling of words. [5] He gathered a positioning of 5 million human feelings of 10,000 well known words, each named multiple times in 10 dialects [36] Sent WordNet, a semi-robotized enthusiastic word reference with more than 100,000 terms yet restricted to English, is another notable passionate word reference. In this review, inspect a bunch of over 1.6 million Twitter postings for human mediators' feelings in 13 European dialects [30]. Enthusiastic classifiers in numerous dialects are prepared to utilize tweets named as preparing information. Gives a best-in-class outline of Twitter's enthusiastic analysis. [27] A new outline of word reference and AI strategies, as well as their blend, is currently accessible. [24] We center around the amount and nature of named tweets, as well as their impacts on the presentation of feeling classifiers. Used to decide the nature of tweets marked understanding between human mediators [20]. The fundamental reason of the article is that the interpreter's agreement offers a significant degree of execution for the classified and one incorporates manual translation and quality evaluation of Twitter posts as well as dataset highlights. Another theme is preparing, execution, and correlations for feeling characterization. [8],[31] The third line looks at the nature of the named information to the presentation of the grouping and supports their essential presumption. [5]. the attention is on careful testing of informational collections. They perceive how they have changed after some time as extra tweets are labeled, and the way that positioning exhibition has changed over the long run. [23] In this additionally go through the impacts of elective preparation and circulation of utilization datasets. [25] The outcomes give short-and long-haul research direction with answers for research issues. [30] The strategies area contains all the data connected with the initial two lines of investigations and the information, translations and enthusiastic appraisals utilized in their outcomes. [15] In the space of entry level position arrangement and AI, they depict four indicative measures. For each dataset, estimations are utilized to decide self and between interpretive arrangements. [29] From these ends they can presume that human translators comprehend the association of classes of feeling. [14] They talk about comparative work on Twitter's feeling characterization calculations and openly accessible marked datasets. [32] They use a conventional statistical test to compare the performance of six different classifiers. The assessment approach is outlined in depth, as well as the typical Twitter pre-processing steps. [13] In the following section they provide a review of related work on the automatic emotion classification of Twitter posts. [6] They review published label sets that were used to train classification models, as well as machine learning methods that were used. [4] Contributions they offer a large collection of emotionally labeled vets in different languages and quality levels. [2] They make freely available collections of over 1.6 million carefully categorized tweets as this is by far the largest dataset published in the literature. [27] The corpus should prove to be a useful and realistic testing ground for various classification techniques. They use four different methods of diagnosis and show that two of them are better for diagnosing emotion classifiers. [26] The same criteria are also used to review the quality of training data,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

which can be used to monitor the interpretation process. [28] Instead, they compare the neutral six emotional classifications using the same basic parameter values. [8]

In this article, they have seen how deep neural network architecture can be used for hate speech. [20] The best values of accuracy are obtained when deep neural network models are combined with a gradient boosted decision tray. [17] They intend to explore the role of user network features in this work in the future [9]. They view many manually labeled tweets in multiple languages, use them as training data, and create automated ranking models. [13] The classification models rely significantly more on the quality and size of the training data than the type of trained model. [18] The aftereffects of the analyses show that there is no critical factual contrast in the presentation of high-grade models. [17] They utilize various interpretive understanding strategies to evaluate the nature of preparing information and to track down the most fragile focuses in an assortment of datasets. [11] They show that while the preparation set is adequately huge, the exhibition of the model agrees. [16] They show that people see three passionate classes (pessimistic, nonpartisan, and positive) as order available marked datasets.[1]

Most of past exploration has zeroed in on ways of learning portrayal, either by extricating a manual component or by joining straight arrangement. [29] Notwithstanding, profound learning techniques have as of late been shown by an enormous number of creators in the 1c 2017 Global Internet Meeting Advisory group (IW3C2), distributed under Innovative. [25] They offer an analysis of over 1.6 million genuinely commented on Twitter postings, the biggest assortment of all time. [30] Named datasets are utilized to prepare feeling characterization models, and they center around four key angles: quality, sum, and examining of preparing information, and order execution. - [27] their fundamental decision is that the sort of positioning chose is less significant than the preparation information, which significantly affects the outcomes. [10].

The table discuss the comparison of previous research on some of the related models, the techniques on which those models work, and the type of data set, which is used in that model, also describes the limitations of that model, based on the accuracy of their results. The major aspect of the table is the limitations and the result of previous models by which we can easily determine that how much works is being done on the previous models and how accurate their results are so that we can make a model which gives the more accurate result and decreases the limitations of the previous models Table 1.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

Table 1. Comparison of Sentimental Analysis previous work.

Ref	Techniques	Dataset	Size	Limitation	Results
[1]	Sentiment Analysis using Corpus	Twitter	2500	It works only on English Language.	Salience and entropy techniques used for accuracy in which Salience has better accuracy.
[2]	GC1 - Model-based Collaborative Governance and GC2 - Data-powered Collective Intelligence and Action	Reviews on lot.	1500	It does not detect spam reviews.	It has good accuracy on non-spam reviews or positive reviews.
[3]	SVM and entropy	Teachers' reviews.	850	Less effective in negative reviews than positive reviews.	It has 89% Accuracy in SVM.
[4]	Using machine techniques KNN, boosting and support vector machines.	Analysis of Linguistic	1100	Less effective on Verbs.	Support Vector Machine technique give maximum accuracy.
[5]	Lexicon Technique	Student Comments	3000	Less effective when dataset is not clean	85% Accuracy
[7]	Distant Supervision (Maximum entropy)	Twitter Data	2000	Less effective in Parts of Speech	75% Accuracy Of maximum entropy.
[8]	Using Emoticon to reduce Machine Learning Dependency. SVM, ME and Naïve Bayes.	General Reviews from peoples.	800	Work good in Noun only.	SVM has high accuracy on comparison with ME and Naïve Bayes.
[17]	Lexicon In Thai using SVM, Naïve Bayes and ID3.	Student Reviews of Loei University.	1148	The small dataset of students was utilized to teacher evaluation	The student comments corpus result should be 0.67
[18]	Deep feed forward neural network	Twitter Database	2000	Limited sample	75%
[31]	Concept Net based ontology	Restaurant Worker Review	900	Less effective on analyzing parts of speech	60%

3. PROPOSED METHODOLOGY

The content associated to the impact of cancelation cricket tour and evaluated in this study to obtain user sentiments from social media. Figure 1 shows the flow and significant steps of the proposed method. The study also trains Nave Bayes machine learning classifier to identify content based on the sentiment of the user based on the recent event of cancellation of New Zealand tour of Pakistan in 2021 One Day International. So,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

research will be to acquire the sentiments of peoples across the globe that what they think about cancellation of this tour. After collecting the data through scraping tool, we have developed the data cleaning tool PSAT using C# .net technology which will clean the data and preprocessing it and giving the results as positive view, negative view and neutral in the shape of graph and resulted data. In this PSAT, we will input the excel file and then PSAT will load the excel file then there will be options of cleaning the data as removing stop-words, removing whitespaces, Capitalizing first letter and more etc. Using these options one can preprocessed the data according to the need then the data will be further processed in which graphs of the result will be generated that what are the sentiments of the peoples in neutral, positive, and negative value.

1) White space removing
2) Upper to lower case conversion
3) Stop words removal.

There are two arrangements of words. One is totally positive, while the other is both positive and negative. The calculation scans the text for terms that match the arrangement of models. The calculation then, at that point, figures out which word types are more in the text. The text is called positive extremity assuming there are more positive words. The issue with calculations in view of guideline is that, when they produce a few outcomes, they need adaptability and precision needed to be genuinely useful.

1.1 Algorithm

Input:

D (Taken as Whole Dataset),

T= (t1, t2, t3, tn) // Tested D values variable.

Analyze the tested D trained dataset);

Calculate the average and S.D

Repeat

By using Gauss probability of t1 in each.

Until the whole variables to be predicted (t1, t2, t3,, tn) has its probability calculated;

Calculated the value for each class.

Until

Greatest value resulted.

Output:

Tested D Class.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

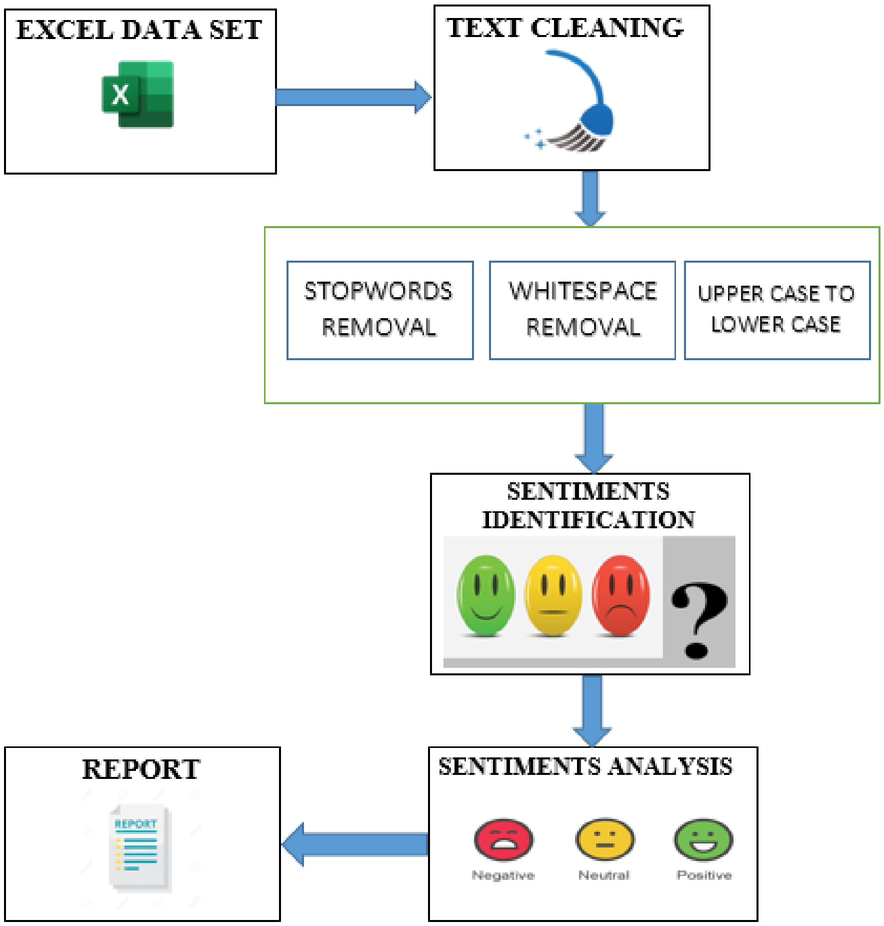

Figure 1. System Architecture of PSAT

1.2 STEPS OF PROPOSED ARCHITECTURE:

1) Excel data set or scraped dataset will be input in the PSAT.
2) Data cleaning function will ask what parameters to clean in the file.
3) PSAT will identify the sentiments in the backend.
4) Result show as a graph in the shape of negative, positive, and neutral comments analyzed.
5) Result will be downloaded in excel file.

1.3 CLASSIFIER USED

Naive Bayes classifier is a group of classification algorithms based on Bayes' Theorem. It refers to a group of algorithms that all work on the same principle: each pair of features being categorized is independent of the others. Each feature in naïve bayes is in the same

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

way importance. Knowing merely the values example humidity, for example, is insufficient to accurately forecast the outcome. None of the characteristics are unimportant, and they are all expected to have an equal impact on the outcome.

1.4 Bayes' Theorem

Bayes' theorem calculates the probability of a match (can be anything which occurred) text which can be based on the probability of a previous match.
The mathematical expression of Bayes; theorem is:

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)}$$

In this x and y are matches P (d) ≠ 0 essentially, we're searching for the chance of occasion an on the off chance that occasion b is valid. Proof is likewise alluded to as Occasion d.
The priori of 'c' is p (c). The proof is a worth allotted to an obscure example's characteristic (occasion d).
P (c|d) is the deduced likelihood of b the chance of an occasion after it has been noticed.
Presently, we can apply Bayes' hypothesis in after way by utilizing our dataset:

$$P(e|w) = \frac{P(w|e)\,P(e)}{P(w)}$$

Whereas w=class variable and e= vector (of size n):

$$e = (e_1, e_2, e_3, \dots \dots, e_n)$$

An example for clearing the class variable and vector is e = input w = No
P (w|e) represents the probability of input

1.5 Naive assumption

The Naïve theorem will be applied in Bayes theorem Now, if c and y are independent, Then, P (c, d) = P (c) P (d) Hence, we have the result:

$$P(w|e_1 \dots, e_n) = \frac{P(e_1|w)P(e_2|w)\dots P(e_n|w)\,P(w)}{P(e)P(e_2)\dots P(e_n)}$$

Which can be expressed as:

$$P(w|e \dots, e) = \frac{P(e)\prod_{i=1}^{n} P(e\,|w)}{P(e)P(e)\dots P(e_n)}$$

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

The denominator values or variable remains constant for a given input, and we can remove that term:

$$P(w|e_1 \ldots, e) \alpha \; P(w) \Pi_{i=1}^{n} P(w \; |e)$$

Then, we create a classifier model. To do so, we calculate the probability of a given set of inputs for all possible values of the class variable s, then choose the output with the highest probability. The mathematical representation for this is:

$$w = argmaz\_\{w\}P(w)\prod\_\{i = 1\}\Lambda\{n\}P(e\_i|w)$$

Finally, P(w) and P (ei | w) must be calculated.

P(w) is also referred to as class probability, and P (ei | s) is referred to as conditional probability.
So, using this Naïve classifier we are going to classify our data set which is going to be based on New Zealand tour of Pakistan and returning to New Zealand over security concerns.

EXPERIMENTAL RESULTS

In this section the experiment has taken place in which we take some input as dataset and then experiment it.

The dataset has been taken from some social media website like Facebook, Instagram, and cricketing website like cricinfo and cricbuzz. We have scraped the information about what are the sentiments of former cricketers, Test cricketers and other fraternity on New Zealand cancelling the tour of Pakistan without playing any ball. This data is being scraped in excel sheet. For Experiment, some of the data (5% of data) has taken in different fields or manually analysis data will distribute to check the accuracy of the PSAT. Experiment of some data. The data scraped from websites will be then go for the experiment using confusion Metrix and python's script of Confusion Matrix which analyze the accuracy of PSAT. First, The Positive data has been taken for the experiment and found out that it was 100% showing the result as the positive comment will analyzed as positive by the PSAT. Similarly, the Negative data taken for the experiment and PSAT analyzed it as negative. Experiment on whole data. Now the whole excel file of 3000 comments from different websites has been taken for experiment and PSAT analyzed it. After the analyzation, Confusion Matrix found that some data should be analyzed as positive, but PSAT analyzed as negative or neutral and some data rightly analyzed, and some data or comments are negative, but it analyzed them as positive or neutral similarly

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

the neutral data analyzed as positive or negative by it this experiment is taken placed using confusion matrix.

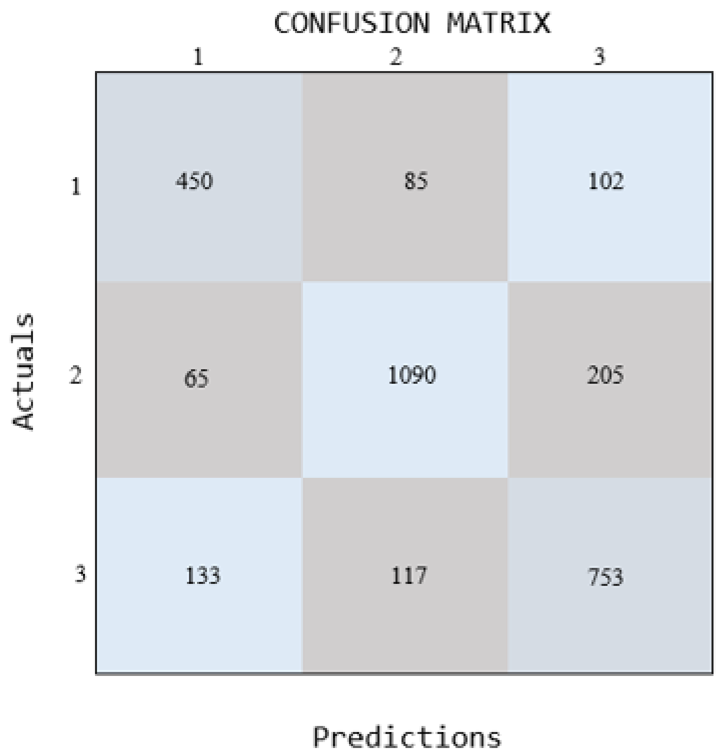

Figure 2. PSAT Tool Class wise True and False Prediction

So, in the figure 2 it has been shown that the data which is Negative has been predicted Negative by 450 and predicted Positive was 85 and Predicted Neutral was 102. So 70% data was predicted right as Negative and 30% falsely predicted. The data and comment which should be predicted Positive was Predicted Negative in 65, predicted Positive in 1090 and Predicted Neutral in 205 comments So, the percentage of passing in Positive was 80%. Similarly, the data which should be predicted as Neutral was Predicted Negative by PSAT in 648, Predicted Positive in 1292 and predicted Neutral in 1060 comments and the passing percentage is 75% approximate. So, the total accuracy of the PSAT is 75% approx. The PSAT analyze 75% of dataset rightly and 25% wrong analysis in confusion matrix.

We have then calculated the F1 Score of PSAT to identify the PSAT's accuracy. For this purpose, we have the F1 score's python's library to analyze F1 Score of PSAT figure 3.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

$$F1 = 2 \text{ x } \frac{precision \text{ x } recall}{precision + recall}$$

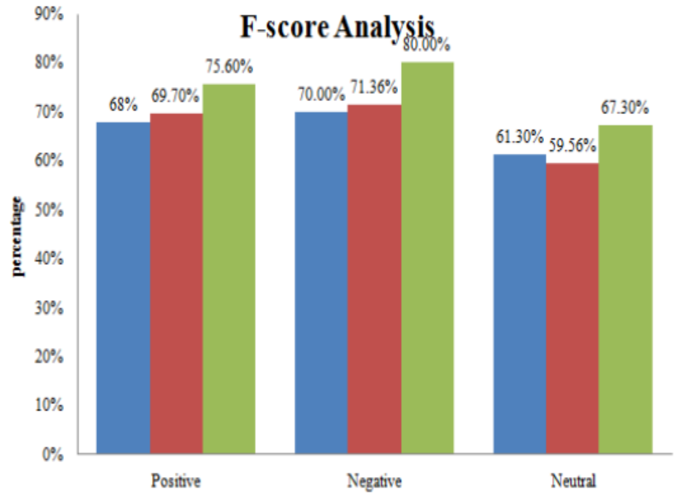

Figure 3. Positive, Negative and Neutral F Score Percentage.

In figure 3, F1 Score calculated using the Machine learning approach which determined the accuracy of PSAT using python script of F1 score.
This script work on the following formulas to calculate the positive negative and neutral analysis of dataset give.

$$Precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F1 = 2 \text{ x } \frac{precision \text{ x } recall}{precision+recall} = 2 \text{ x } \frac{0.70 \text{ x } 0.69}{0.70+ 0.69} = 0.69$$

So, the accuracy calculated using the F1 Score is 69% which means that PSAT will give 69% accurate result according to F1 Score experiment.

After analyzed the whole reviews combine. We have analyzed the sentiments reviews data of cricket analyst, cricket fans in Pakistan and fans of other countries.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

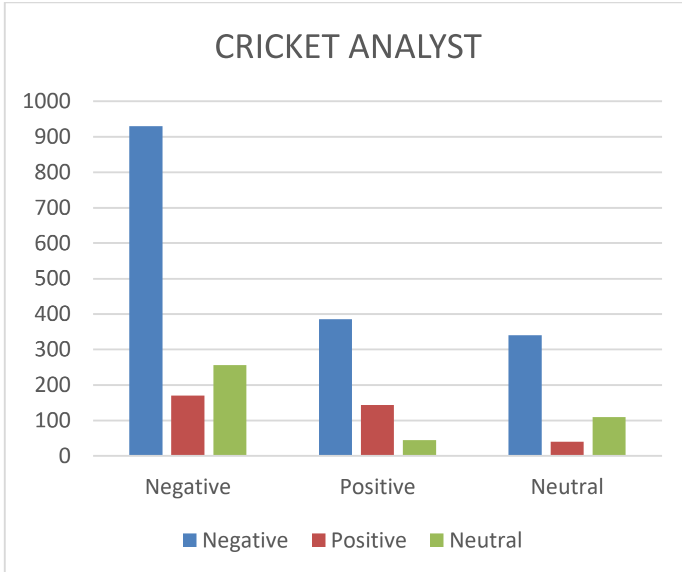

Figure 4. Cricketer point of view

So, first the data of Cricket analyst reviews been analyzed through PSAT, so the results are that the data of negative were bigger that positive and neutral, so the system accuracy was 78% and the cricket analyst reviews were negative in Figure 4 and 5.

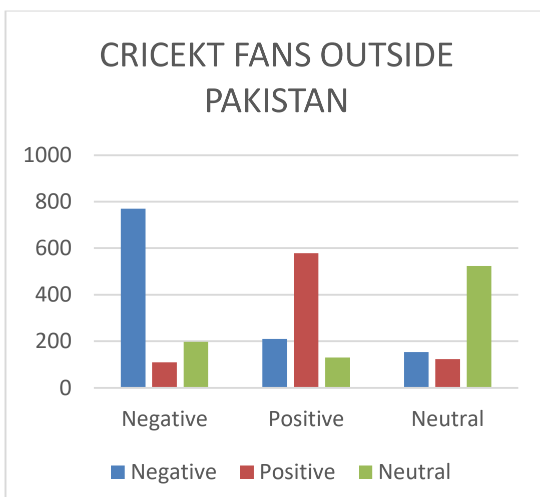

Figure 5. Cricket Fans outside the Pakistan

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

After cricket analyst data, the data of cricket fans outside Pakistan taken to analyze, the result was there were mix of negative, positive, and neutral data as 40% were negative, 32% were positive and 28% were neutral. There was not much difference as in Cricket Analyst data.

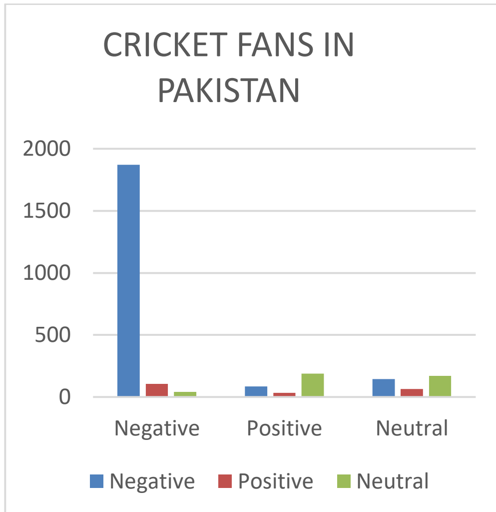

Figure 6. Cricket Fans in Pakistan

In the last, the sentiments reviews data of cricket fans across Pakistan taken to analyze, the results were negative as the reviews were negative there is least reviews of positive and negative and PSAT analyzed it negative Figure 6.

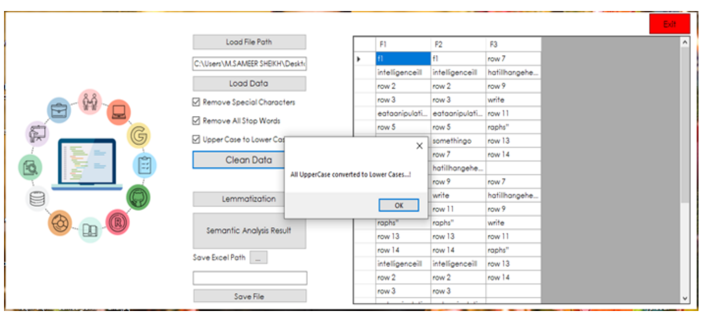

Figure 7. Preprocessing and Sentiment Analyzer (PSAT)

CONCLUSION

PSAT Tool is developed for analyzing the sentiment of peoples on the recent event of New Zealand cancelled the tour to Pakistan in 2021.This tool used to collect, manage, preprocess, and analyze the sentiments Figure 7, Furthermore, it was also used to build Naive Bayes classifier. The correct predication of system is 75% and incorrect predication is 25%. And the accuracy calculated using the F1 score. There is hardly any research that has contributed to this field. Due to the limitation of data in this field, labelled training data for machine learning classifiers is limited, and classifiers often perform better when fed with a significant amount of data. Furthermore, the dictionary also has limitations as

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

it struggles to differentiate simple and multi word units. Although these limitations, the model has achieved 69.01% accuracy using Naïve Bayes classifier

References

1. A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326
2. David Osimo and Francesco Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis".
3. Keith Norambuena, Brian &Lettura, Exequiel& Villegas, Claudio. (2019). Sentiment analysis and opinion mining applied to scientific paper reviews. Intelligent Data Analysis. 23. 191-214. 10.3233/IDA-173807.
4. S.Padmaja et al, "Opinion Mining and Sentiment Analysis – An Assessment of Peoples' Belief: A Survey", International Journal of Ad hoc, Sensor & Ubiquitous Computing IJASUC, Vol 4 No 1, Feb 2013.
5. G. Vinodhini et al, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 2, Issue 6, June 2012.
6. Wan, X..“A Comparative Study of Cross-Lingual SentimentClassification”. In Proceedings of the The 2012 IEEE/WIC/ACMInternational Joint Conferences on Web Intelligence and IntelligentAgent Technology-Volume 01 (pp. 24-31).IEEE Computer Society.2012
7. Alec Go, Richa Bhayani, Lei Huang, "Twitter Sentiment Classification Using Distant Supervision".
8. Dr. Ritu Sindhu, Ravendra Ratan Singh Jandail, Rakesh Ranjan Kumar, "A Novel Approach for Sentiment Analysis and Opinion Mining", International Journal of Emerging Technology and Advanced Engineering (IJETAE), Vol 4, Issue 4, April 2014.
9. S. Batra and D. Rao, ”Entity Based Sentiment Analysis on Twitter”, Stanford University,2010 .
10. ZHU Jian , XU Chen, WANG Han-shi, “" Sentiment classification using the theory of ANNs”, The Journal of China Universities of Posts and Telecommunications, July 2010, 17(Suppl.): 58–62 .
11. Rudy Prabowo, Mike Thelwall, "Sentiment analysis: A combined approach .", Journal of Informetrics 3 (2009) 143– 157.
12. Go,Lei Huang and Richa Bhayani , "Twitter Sentiment Classification using Distant Supervision", Project Report, Standford,2009
13. G. Mishne. Experiments with mood classification in blog posts. In 1st Workshop on Stylistic Analysis Of Text For Information Access, 2005.
14. B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135, 2008.
15. J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics
16. Melville, Wojciech Gryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD"09, June 28–July 1, 2009, Paris, France. Copyright 2009 ACM 978-1-60558-495-9/09/06.
17. Cross-Lingual Sentiment (CLS) dataset. P. Prettenhofer and B. Stein. http://www.uniweimar.de/cms/medien/webis/researc h/corpora/webis-cls-10.html, 2010.
18. Long-Sheng Chen , Cheng-Hsiang Liu, Hui-Ju Chiu , "A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics 5 (2011) 313– 322.
19. Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification",Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.
20. Bai, and R. Padman, "Markov blankets and metaheuristic search: Sentiment extraction from unstructured text," Lecture Notes in Computer Science, vol. 3932, pp. 167–187, 2006.
21. Go, R.Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper ,2009

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/ABXN4

22. G. Vinodhini et al, “Sentiment Analysis and Opinion Mining: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 2, Issue 6, June 2012
23. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. P. Turney. Proceedings of the Association for Computational Linguistics (ACL), 2002.
24. Thumbs up? Sentiment Classification using Machine Learning Techniques. B. Pang, L. Lee and S. Vaithyanathan. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.
25. Mining and Summarizing Customer Reviews. M. Hu; B. Liu. Proceedings of KDD, 2004. [12] Opinion Observer: Analyzing and Comparing Opinions on the Web. B. Liu, M. Hu and J. Cheng Proceedings of WWW, 2005.
26. Sentiment Analysis and Subjectivity. B. Liu. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.
27. Opinion Summarization of Web Comments. M. Potthast and S. Becker. Proceedings of the European Conference on Information Retrieval (ECIR), 2010.
28. Basant Agarwal, Namita Mittal, Pooja Bans al, Sonal Garg “Sentiment Analysis Using Common-Sense and Context Information.”Journal of Computational Intelligence and Neuroscience, 9 (2015) (2015)
29. U.T. Gursoy, D. Bulut, C. Yigit“Social Media Mining and Sentiment Analysis for Brand Management.” Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology, 3 (1) (2017), pp. 497-551
30. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia
31. Basant Agarwal, Namita Mittal, Pooja Bans al, Sonal Garg “Sentiment Analysis Using Common-Sense and Context Information.”Journal of Computational Intelligence and Neuroscience, 9 (2015) (2015)
32. U.T. Gursoy, D. Bulut, C. Yigit“Social Media Mining and Sentiment Analysis for Brand Management.” Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology, 3 (1) (2017), pp. 497-551
33. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia
34. Dr. Ritu Sindhu, Ravendra Ratan Singh Jandail, Rakesh Ranjan Kumar, “A Novel Approach for Sentiment Analysis and Opinion Mining”, International Journal of Emerging Technology and Advanced Engineering (IJETAE), Vol 4, Issue 4, April 2014.
35. Andrea Esuli and Fabrizio Sebastiani, “Determining the semantic orientation of terms through gloss classification”,Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.
36. Cross-Lingual Sentiment (CLS) dataset. P. Prettenhofer and B. Stein. http://www.uniweimar.de/cms/medien/webis/researc h/corpora/webis-cls-10.html, 2010.
37. P. Turney, M. Littman, “Measuring praise and criticism: Inference of semantic orientation from association” in ACM Transaction on Information Systems Journal, volume 21, issue 4, pp. 315-346, 2003.
38. [Esuli A. and Sebastiani E. “SentiWordNet: A publicly available lexical resource for opinion mining” in proceedings of Language Resources and Evaluation (LREC), 2006.
39. R. Sauri, “A factuality profiler for Eventualities in Text” PhD Thesis, 2008. [19] R. Quirk, G. Sidney, G. Leech, J. Svartvik, “A comprehensive Grammar of the English Language”, 1987.
40. Go, A., Bhayani, R., Huang, L., “Twitter sentiment classification using distant supervision”, Technical Project, Stanford Digital Library Technologies Project, 2009.