Proceedings of the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), 26-28 September, Dhaka, Bangladesh

Real Time Sentiment Analysis and Opinion Mining on Refugee Crisis

Md. Mahiuddin
Department of Computer Science and Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
mmuict@gmail.com

***Abstract*— It is acknowledged that Twitter is a micro blogging social site and millions of people share their thoughts, views and reactions, as well as commenting on particular subject as his opinion or fact for seeking attention from different categorical person. In the current analysis and experimentation, we investigate the public opinions, facts and sentiments on Refugee Crisis which is widely discussed topic in social media platform today. To analyze the public sentiment on this crisis, around 35,000 relevant Twitter data have been collected in five different languages such as English, Bangla, Turkish, Chinese and Urdu, and use them for sentiment analysis and decision mining in the way of data mining and data science for Refugee Crisis. For that, a new model of real time sentiment analysis on Refugee Crisis is presented in this paper to provide some prediction on political improvements. This paper will also be able to give end level decision of how much people are commenting for supporting Refugee and how much comments are posting against Refugee by binomial classification of positive and negative using supervised machine learning algorithms such as DT, RF, NB and KNN. Result shown that KNN algorithm perform the best polarity accuracy of 95% compared to DT, RF and NB classifier. Respective and responsible person for Refugee can get a better knowledge by having our analysis.**

Keywords*— *Refugee crisis; Sentiment analysis; Data science; Machine learning; Opinion mining

I. Introduction

Sentiments can be defined as emotions, opinions, feelings or ideas expressed by people. There are two types of sentiment such as facts and opinion-based information. Fact is just the briefing about a matter, event called as objective and opinion is the view, thought, intention and idea about an event or topic called as subjective.

Sentiment Analysis is the process of analyzing, experimenting, testing the opinion or fact of people after collected and pre-processed data in text form. It is also known as opinion mining, emotion gathering, decision mining, idea generating on particular matter of a person which can be subjective or objective [1]. Subjective has terminological meaning that is the own opinion of person and objective has also inside meaning that is the intention of sharing just facts of any incident.

In the recent time a huge amount of research is done by researchers on Sentiment Analysis and Opinion Mining on different subjects and events. As social media becomes a great source of data, Sentiment Analysis or Opinion Mining is able to give more impact on community service.

People of the today's world are experimenting on various sentiment data. New information and theory are gathered from there. They have used different analyzing protocols to differentiate the sentiment class. The actual meaning and classification of sentence is the most important part in decision making. There is a plenty of research regarding sentiment analysis but not much about Refugee Crisis.

Refugee can be defined as groups of dislocated persons, who could be dislocated either internally or migration. Internally dislocated persons are the people, who forced to leave their homes, but they failed to reach a neighboring country and the migrated people are the persons who left their home and took shelters on other country. Because of war nature has changed in the last few decades, the number of internally dislocated persons has increased.

Refugee Crisis is one of the biggest issues all over the world today. This problem originated many years ago in many countries like Afghanistan, South Sudan and Somalia. Very lately, people from Myanmar and Syria are the new victim.

Though millions of people have been suffering by this serious problem, still now, we can't see any effective decision and proper solution on this particular embedded problem.

In the present world social media is one of the biggest platforms to express people's opinion, thought, view, intention, reaction and idea. There are a lot of social media platform, Twitter is one of them. In addition to having a global coverage of issues, Twitter provides a media platform that enables sharing opinions easily using various content forms including text, images, links with the character restriction unlike many other social media platforms. More than 300 million people use Twitter all over the world. So, it has important impact in the real world [1].

Twitter sentiment mining can be helpful in different situations such as analyzing people's comment on different event, product, movie, song etc.

In this paper, we have proposed a new way of real time sentiment analysis on current Refugee Crisis to provide some prediction on polarity types for political improvement based

978-1-7281-4934-9/19/$31.00 ©2019 IEEE

on twitter data. The extracted and pre-processed data sets are used in various opinions mining algorithm such as Support Vector Machine (SVM), Naive Bayes (NB), K Nearest Neighbor (KNN), Random Forest (RF) and Decision Tree (DT) using python programming.

II. Related Work

There are some referable researches done in the field of Sentiment Analysis by predicting election, analysis of student feedback, analysis of movie and product review and Syrian refugee crisis etc.

Dhanalakshmi V. et al. use supervised learning algorithms to find the polarity of the student feedback based on pre-defined features of teaching and learning [2]. This paper presents a comparative performance study of the algorithms like SVM, NB, KNN and NN classifier. Their Result shows that KNN got the best precision result of 100%, NB got the best recall and accuracy result of 97.07% and 99.11% respectively.

Text analyzing tool to get tweets in Hindi language is used in [3]. They analyze 42,235 collected tweets that referenced on various political parties in India, during the campaigning period of elections in 2016 by using both supervised and unsupervised technique. They use Dictionary Based, NB and SVM algorithm to classify the data as positive, negative and neutral. Result shows that BJP have the chance of 78.4% to win elections due to the positive sentiment received in tweets. NB algorithm give the accuracy of 62.1%.

Huma Parveen et al. experiment the extraction of sentiment from a famous website Twitter where the people post their views and opinion [4]. They perform the sentiment analysis on movie review related tweet. They showed the results of their sentiment analysis as different sections presenting positive, negative and neutral sentiments.

The public opinions and sentiment analysis towards the Syrian Refugee Crisis was done in [5]. The results showed that there was significant difference between sentiment from Turkish tweet and English tweets. The results also show that the tweets posted by Turkish are more positive sentiments. On the other hand, the largest number of English tweets is neutral.

Asma Musabah Alkalbani et al. focus on experimenting feedback of customers on SaaS products by predicting reviewer's attitudes [6]. The goal of this paper is to predict the sentiment of SaaS reviews of customers. They proposed five techniques based on five algorithms such as SVM, NB, NB (Kernel), KNN and the DT algorithm to predict the attitude of SaaS reviews. In their experiment they got 92.37% accuracy by using SVM algorithm which proved that this algorithm is to be able to give better result on sentiment of the online reviews compared with the other models.

Opinion mining for newspaper headline is work done in [7] using SentiWordNet. They separate the adverb-adjective combination exists in the statements. They also analyzed the news headline whether it is a part-of-speech tag.

Sentiment analysis on movie review is done in [8]. A new technique is proposed with heterogeneous features such as machine learning based and Lexicon based features and classification algorithms like NB and LSVM used to build the system model. 250 training dataset and 100 testing datasets are used in the experiment. 89% for NB and 76% for SVM of accuracy is found. 84% for NB and 79% for SVM accuracy are calculated for 300 training dataset and 150 testing datasets.

A study is worked done on a dataset of tweets for 6 major US Airlines and performed a multi-class sentiment analysis [9]. Seven different classification strategies are used such as DT, RF, SVM, KNN, LR, NB and AdaBoost. Based on the results obtained, they calculated the accuracies to draw a comparison between each classification approach and the overall sentiment count was visualized combining all six airlines.

From above discussion we observe that there is no sufficient research work is done on refugee crisis. In this paper, we have completed a study on refugee crisis.

III. Explanation of Proposed Model

Our proposed model is starts with collecting the tweet data and end with the evaluation of final results. There are several steps in our model. Details model is depicted in Fig.1.

To facilitate the description of whole model the model is partitioned into four major parts such as data collection, preprocessing, classification and implementation shown in Fig. 2.

A. Data collection

Collection of data from social blog site Tweet is one of the hardest jobs. In fact, it is interesting too. There is so many ways to extract or collect Tweet data.

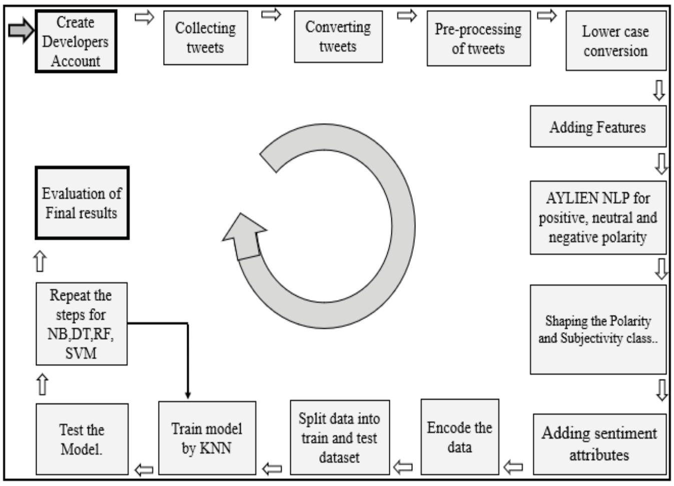

Fig. 1. Proposed Model of real time sentiment analysis and opinion mining

Fig. 2. Process strategy

Our proposed model is implemented with real time data. A well-known and remarkable software called RapidMiner which can help to extract tweets from Twitter with authorized gateway created by RapidMiner and Twitter developer community. It is noticeable that RapidMiner can also be able to handle any data science project and modeling with supervised machine learning (SML) algorithm in proper and dependable way.

We collect tweets using required query which can be easily matched with our searched tag with tweets posted by Twitter user. If our searched tag found in Twitter user's tweet then it will be extracted in our document file. Two different types of operator from RapidMiner called as "search Twitter" and "write excel" are used. Fig. 3 shows the operator that used in RapidMiner for collecting tweets.

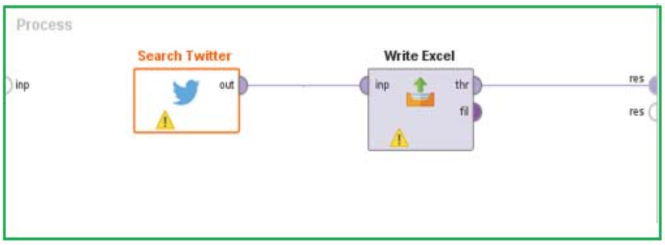

Fig. 3. Data Collection Process

Around thirty-five thousand rows of tweets are collected from authorized Twitter community as a Twitter developer with five different languages such as English, Bangla, Turkish, Urdu and Chinese. It is needed to convert from various languages to global language. In this paper, Google translator is used for tweets conversion. Data was not noise free and congested with removable resources. Its included duplicated tweets, lots of URL links, special symbol, emojis, mentioned name that is needed to be cleansing and cleaning.

B. Data pre-processing

Data which extracted from Twitter contains different non-sentiment contents such as duplicate tweets, website links, emotions, etc. which have to remove before processing our tweets so that the sentiment can generate accurately. Pre-processing steps is depicted in Fig. 4.

Every day peoples are sharing huge number of tweets. When extract their shared tweets from social media platform, sometime one tweet comes multiple times. This is what called duplicated tweets and that should be removed for our desired analysis (Fig. 5).

Extracted Twitter data so called tweets consist of different type of information that is called as URL. URL should be removed from tweets. The URL found here many as YouTube and Facebook's video link.

There are different types of unnecessary symbols called as special symbol used by the social media user. These types of symbol have to remove as punctuation mark (!), full stop (.) mentioning symbol (@), single quote (""), single quote ('') etc. which does not contain any sentiment. We removed all the special symbols from our tweet dataset by following the apidMiner provided operator with the help of regular expression. Fig. 6 shows it well.

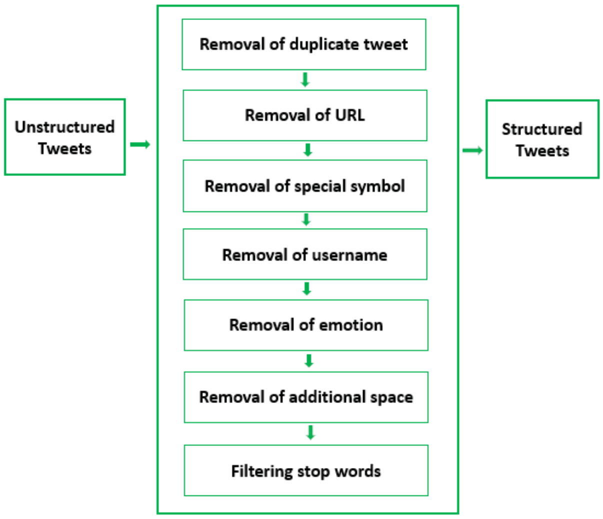

Fig. 4. Data Pre-Process steps

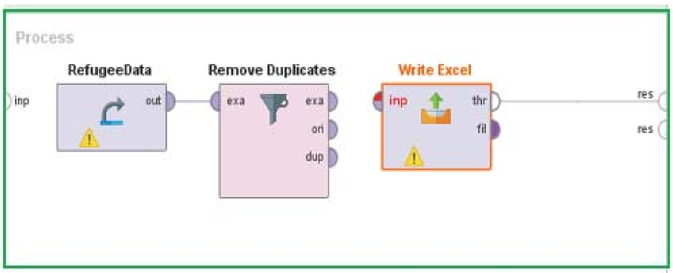

Fig. 5. Duplicate Tweets Removal diagram

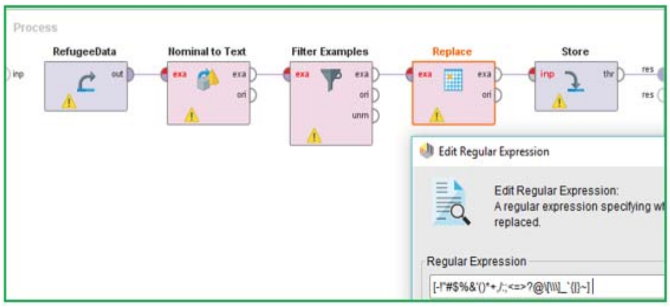

Fig. 6. Special symbols removal diagram

One user can use one username and that is should be unique by following the guide line of Twitter; anything is posted by a user there is his/her username proceeding by @ which is used as proper nouns. For example, @someones_username. This also removed from our dataset for effective analysis.

There may be extra white space in the data and it needs to be removed. By removing white spaces, the analysis is to be done more efficiently.

"Stop words" means some common words that don't carry useful information of sentence like "the," "a," and "and". Removing stop words means that Model won't be able to see these words and will be trained on a clean Dataset. Text

Analysis Platform (TAP) recognizes stop words based on a manually created, comprehensive list, but since what defines a stop word can vary depending on context. Since shorter documents like Tweets contain such little text, for training a model filtering stop word is needed. Fig. 7 depicts the process of how filter the stop words from extracted data set.

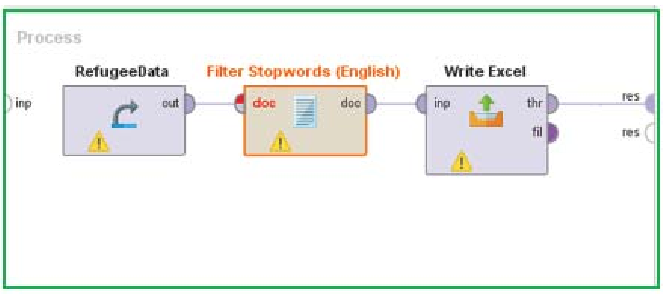

Fig. 7. Filtering of stop words

C. *Classification*

After completing our pre-processing steps, we got 8788 structured tweets from around 35,000 unstructured tweets. We have analyzed our tweets by AYLIEN extension which is natural language programming (NLP) platform provided by RapidMiner. At first, we are going to determine the sentiment of each tweet whether they are Positive, Negative or Neutral. Analyzed Sentiment Operator is added in our process and selected "text" as Input Attribute in RapidMiner software. (Fig. 8)

Tweets data are processed and analyzed to find the polarity such as positive, neutral and negative, and Subjectivity (subjective & objective). Table I shows the polarity after processed and analyzed tweet data.

Usually in sentiment analysis negative words make the negative class and positive word makes the positive class. So, each tweet is reviewed as negative, positive and neutral polarity class. If there found any supporting words or hash tag likes help, #help, save, #save then read it manually and classified as support class otherwise classified as against class in polarity attribute. Fig. 9 describe the support class and against class.

In our dataset we have several attributes. Various source of text found in our dataset. These are Twitter Web Client, Twitter for Android, Twitter for iPhone, Twitter Lite, Twitter for iPad, Facebook, Google, LinkedIn. Our provided attributes are as follows:

Id; Name; Tweet date; Source of text; Language; Retweet; Subjectivity; Subjectivity confidence; Polarity; Polarity confidence.

Id, Name and Tweet date contains no sentiment value. We removed them from our dataset. Especially, Tweet date is removed because we are working in real time. So, tweet date is redundant in this case of analyzing.

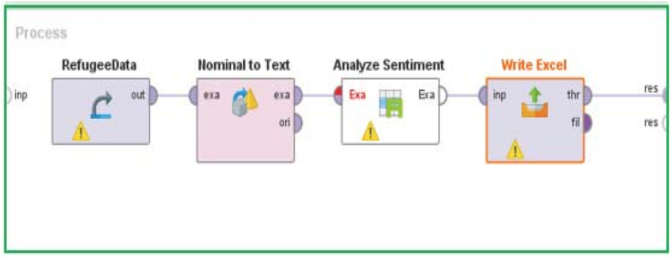

Fig. 8. Sentiment classification operator's diagram

TABLE I. SENTIMENT CLASSIFICATION

polarity	subjectivity	polarity_confidence	subjectivity_confidence
positive	subjective	0.731	1
positive	subjective	0.606	1
neutral	subjective	0.877	1
neutral	objective	0.783	1.000
negative	subjective	0.947	1
negative	subjective	0.938	1
neutral	objective	0.798	1
negative	objective	0.838	0.685
neutral	objective	0.783	1.000
negative	subjective	0.731	1
negative	objective	0.901	1.000

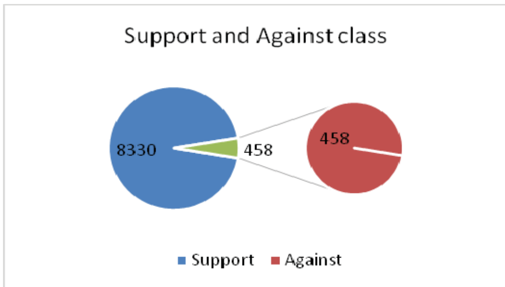

Fig. 9. Visualization of polarity class

Many classification algorithms can't deal with string data. We transformed our attributes into numerical value for getting algorithmic performance with the help of python programming. This is what just form of presentation. We encoded the source attribute of each tweet into different labeled of numeric value which is done by Label Encoder class from python. We transformed our language attribute by dummy encoding with the help of One-Hot-Encoder class from python. In subjectivity attribute, subjective is what we called it opinion presented by 1 and objective is what we called it fact is presented by 0. In polarity class support is presented by 0 and against presented as 1. Table II shows the value of polarity after conversion of string data to numerical value using python programming.

TABLE II. TRANSFORMATION OF ATTRIBUTES

Filter (8,788 / 8,788 examples): all

Retweet	subjectivity	Subjectivity...	PolarityConf...	polarity
2	1	1	0.900	0
3	0	1	0.500	0
10	1	0.800	0.700	0
0	1	1	1	0
1	1	1	0.900	1
0	0	1	0.500	0
14	1	1	0.500	0
1	1	1	0.500	0
0	1	1	0.900	0
2	1	1	0.400	1

D. Implementation

We implemented our SML algorithm by python programming. We have done our project by going through OOP concept with Python. Details implementation structure is depicted in Fig. 10.

At first, we imported three large packages of python namely numpy, matplotlib. pyplot and pandas. Dataset are imported and separated the dependent and independent variable as we do in our typical mathematical problem. Then performed label encoding for our categorical data by Label Encoder class from scikit learn. One hot encoding or binary presentation for our required attribute by One-Hot-Encoder class from scikit learn is formed and avoid dummy variable trap of One-Hot-Encoder.

Dataset is split into training and testing data set with the help of scikit learn and cross validation, and complete fit and transformation by Standard Scaler class from scikit learn.

In the step of training with classifier, trained the dataset by using Decision Tree Classifier, K Neighbors Classifier, SVM, Random Forest Classifier, Gaussian NB class from scikit learn. Finally, build the confusion matrix from scikit learn with computed desired results.

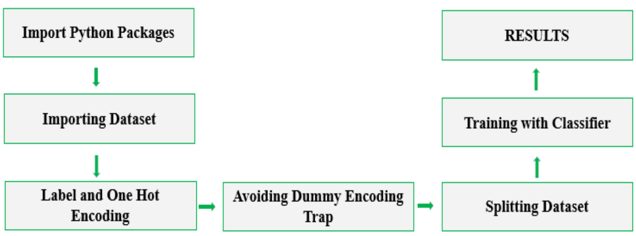

Fig. 10. Implementation diagram

IV. RESULTS AND DISCUSSIONS

In this section, we have completed our experiment with our dataset by taking one-fourth of data as testing and three-fourth of data as training data from 8788 tweets dataset. Some graphical visualization of experimented dataset in various perspectives has been presented in the following sub section.

A. Visualizing of NLP Results

Results stored in a document but in order to make them more presentable, it is necessary to visualize the results. From Fig. 11, it is shown that among 8788 tweets, 8330 is the support class and 458 are against class.

After analyzing natural language processing, it is noticed that in our dataset, we got 5129 opinion sentence and 3659 fact or so-called objective sentence from tweeter user. Graphical view is shown more exclusively in Fig. 12.

Fig. 13 depicted the subjectivity and polarity. It is also shown that among 3659 facts, 126 are against class and 3533 is the support class. For the opinion subjectivity, 332 are against class and 4797 are support class among 5129 opinion subjectivity.

Tweet versus re-tweet also observed in Fig. 14. Re-tweet means when a Twitter user posted, number people comment on particular post.

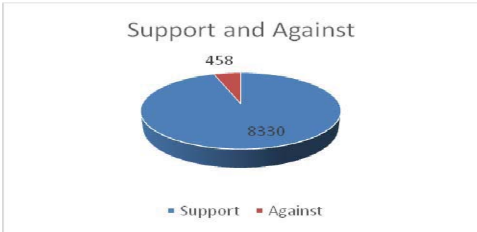

Fig. 11. Overview of Polarity Class

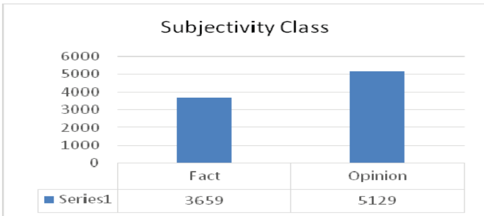

Fig. 12. Overview of Subjectivity Class

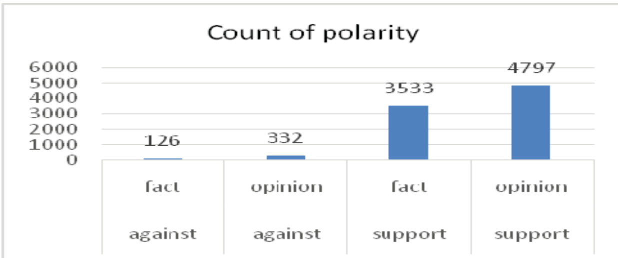

Fig. 13. Subjectivity and Polarity Diagram

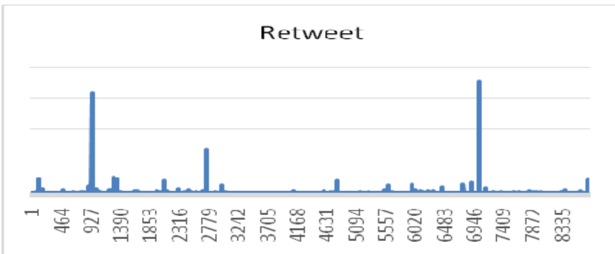

Fig. 14. Retweet Visualization

B. Algorithmic performance and result visualization

Algorithmic performance can be calculated and measured using confusion matrix. In the following section we described our used domain.

Confusion matrix is presented as 2x2 matrix which is the combination of actual class and predicted class. By confusion matrix we can get the prediction of model result by comparing with actual testing dataset. (Table III).

TABLE III. CONFUSION MATRIX

Confusion Matrix		Predicted Class	
		yes	**no**
Actual Class	**yes**	True Positive (TP)	False Negative (FN)
	no	False Positive (FP)	True Negative (TN)

Accuracy is defined as the ratio between addition of TP, TN and addition of TP, TN, FN, FP. Simply we can write as follows.

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

Recall is defined as the ratio between TP and addition of TP, FN. That is shown in the following.

$$\mathrm{Re}\ call = \frac{TP}{(TP + FN)} \quad (ii)$$

Precision is defined as the ratio between TP and addition of TP, FP. We can write as follows.

$$\mathrm{Pr}\ ecision = \frac{TP}{(TP + FP)} \quad (iii)$$

For calculating f-measure first we multiplied 2, recall & precision with each other and then product divided by addition of recall, precision.

$$F - measure = \frac{2 \times \mathrm{Re}\ call \times \mathrm{Pr}\ ecision}{\mathrm{Re}\ call + \mathrm{Pr}\ ecision} \quad (iv)$$

We have experimented our structured data twice by selecting seventy five percent training dataset and by selecting eighty percent training dataset where twenty five percent dataset was for testing and twenty percent dataset was for testing respectively.

It is noticed that KNN machine learning algorithm shown the best accuracy by only giving five percent error and NB shown the lowest accuracy by eighty three percent. In Fig. 15, we have presented a pie diagram which contains accuracy of our used KNN, NB, DT and RF classifier.

Table IV presents Recall, Precision and F-measure with respect to our used classifier. It is also observed that performance of KNN is better than RF, NB and DT classifiers with respect to Recall, Precision and F-measure. Graphical representation of Recall, Precision and F-measure for KNN, RF, NB and DT is also depicted in Fig. 16.

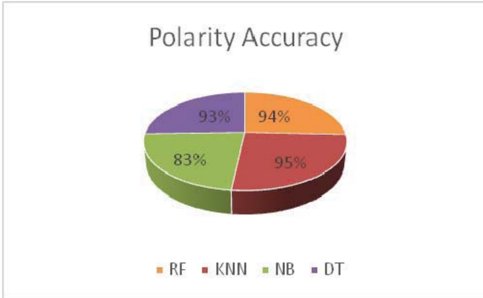

Fig. 15. Accuracy of predicting Polarity

TABLE IV. RECALL, PRECISION AND F-MEASURE OF POLARITY CLASS

Classifier	Recall	Precision	F-measure
KNN	0.99	0.95	0.98
RF	0.99	0.95	0.97
NB	0.86	0.95	0.90
DT	0.98	0.95	0.96

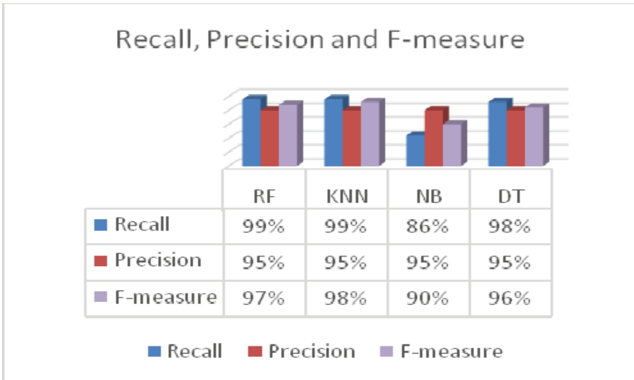

Fig. 16. Recall, Precision and F-measure of Polarity Class

C. Comparisons with existing works

We compared our accuracy result with related work. It is concluded that DT, KNN and RF performed better predicting result after calculating the accuracy from confusion matrix for the proposed method. Table V represents the details results of comparisons with related works. Accuracy of proposed model for DT is 93% compared with exiting is 91%. For KNN, proposed model also show the better accuracy 95% compared with related works such as 91% and 87%.

We also compare the value of Recall, Precision and F-measure of proposed method with existing works. With respect to Recall, Precision and F-measure, our proposed method gives the better results. Details are presented in Table VI.

TABLE V. COMPARISONS OF ACCURACY WITH EXISTING WORKS

Classifier	Existing Works	Proposed Method
NB	83% [6]	83%
DT	91% [6]	93%
KNN	91% [6] 87% [10]	95%
RF	--	94%

TABLE VI. COMPARISON TABLE OF RECALL, PRECISION AND F-MEASURE WITH EXISTING RELATED WORKS

	Existing works [9]			Proposed Method		
Classifier	Recall	Precision	F-measure	Recall	Precision	F-measure
KNN	0.592	0.590	0.593	0.99	0.95	0.98
RF	0.865	0.856	0.865	0.99	0.95	0.97
NB	0.647	0.642	0.646	0.86	0.95	0.90
DT	0.646	0.630	0.645	0.98	0.95	0.96

D. Discussions

In this section we have shown the experimental results of proposed model with compare to the different machine learning approach. From NLP results, it is observed that most of the tweets are in support to refugee. For the subjectivity, number of facts is less than the opinion. Polarity for the five different languages also analyzed. From polarity results, it is shown that most of the tweets are in English language. From five different languages, it is also observed that most of the tweets are in support to refugee. To take decision about refugee, we have implemented different types of supervised machine learning algorithms. In those results, KNN algorithm shows the best performance compared with other algorithms in terms accuracy. Comparison with existing related works shows that our proposed model performed better than existing works with respect to accuracy, recall, precision and f-measure.

V. CONCLUSION

In this paper we have shown a different approach of text mining and sentiment analysis on refugee crisis. Thirty five thousand tweets data have been collected for our experiment. After preprocessing we have only 8788 tweets data for analysis. Different types of supervised machine learning algorithms are implemented on the preprocessed data. Proposed model have performed on the data set with the most accurate and best result while in classification of support and against class. It is also shown that KNN algorithm give the best results compared with other algorithms. We can decide that our proposed model is better model while compare with other existing models. In future, we will try to work with more social media platform like Facebook and YouTube.

References

[1] Haseena Rahmath P, Opinion Mining and Sentiment Analysis Challenges and Application. International Journal of Application or Innovation in Engineering & Management (IJAIEM), vol. 3, no. 5, 2014.

[2] Dhanalakshmi V., Dhivya Bino, Saravanan A. M., Opinion mining from student feedback data using supervised learning algorithms. 3rd MEC International Conference on Big Data and Smart City, Muscat, Oman, 2016.

[3] Parul Sharma, Teng-Sheng Moh., Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter. 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016.

[4] Huma Parveen, Prof. Shikha Pandey.,Sentiment Analysis on Twitter Data-set using Naive Bayes algorithm. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)., Bangalor, India,2016.

[5] Nazan Ozturka, Serkan Ayvaz,Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. Science Direct, Telematics and Informatics, Volume 35, No 1, P 136-147, 2018.

[6] Asma Musabah Alkalbani, Ahmed Mohamed Ghamry, Farookh Khadeer Hussain, Omar Khadeer Hussain.,Predicting the sentiment of SaaS online reviews using supervised machine learning techniques.2016 International Joint Conference on Neural Networks (IJCNN)., Vancouver, BC, Canada, 2016.

[7] Apoorv Agarwal, Vivek Sharma, Geeta Sikka, Renu Dhir., Opinion Mining of News Headlines using SentiWordNet. 2016 Symposium on Colossal Data Analysis and Networking (CDAN)., Indore, India, 2016.

[8] Rachana Bandana, Sentiment Analysis of Movie Reviews Using Heterogeneous Features. 2018 2nd International Conferene on Electronic, Materials Engineering & Nano Technology (IEMENTech), Kolkata, India, 2018.

[9] Ankita Rane, Dr. Anand Kumar, Sentiment Classification System of Twitter Data for US Airline Service Analysis. 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 2018.

[10] Ch. Nanda Krishna, Dr. P. Vidya Sagar, Dr. Nageswara Rao Moparthi, Sentiment Analysis of Top Colleges. 2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, India, 2018.