Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary & Peer Reviewed
Volume: 4, Number: 2, Pages: 149- 152, Year: 2021

Sentiment Analysis Based on Reviews Using Machine Learning Techniques

Ameema Sattar*, Joddat Fatima

Software Engineering Department, Bahria University Islamabad, Pakistan
*Corresponding author: Ameema Sattar (ameema.sattar2017@gmail.com)

Abstract: **In our daily life, people's opinions and experiences are important sources of information. To measure the feeling of people's opinions the term used that is called sentiment analysis. The text is the main method of communicating on the Internet in modern digital time. Sentiment analysis captures the user's views, moods, and their opinion related to the specific services provided by the business organization in a real-time. This research focuses on Roman Urdu reviews. It has three basic classes: negative, positive, and neutral where reviews are classified. The proposed method is Analysis of different machine learning algorithms with different datasets has made and a comparison shows, SVM performs the best result on used data sets. We get the highest accuracy against dataset-1 that is 69%, dataset-2 is 73% and dataset-3 is 76%, a clear result in the form of accuracy, precision, recall, and f1 score shows the results against the specific techniques against the dataset.**

Index Terms-- Sentiment **analysis, Machine learning, social media, text classification**

I. INTRODUCTION

People's opinions and experiences are the main sources of information in our everyday life. People use social media platforms to comment on products and share opinions about specific products or services [1]. All big organizations or companies must know the feedback or sentiment of customers or users. As a result, companies are beginning to realize the importance of social media comments as part of their marketing schemes [2]. The social media data is used to know the customer feedback, opinions of users which is informative for companies, because peoples share own experience about a specific thing either they are satisfied or not [3]. Therefore, sentiment analysis is an important aspect that tells you about how popular business services are running. Maintaining customer loyalty is an important marketing approach as well. Meanwhile, social media platforms are full of customer feedback on whether they are at this time satisfied or not. The reviews or opinions of customers are written in a natural language. In sentiment analysis, the weightage of comments is measured to count the positive and negative words [4]. The analysis is a type of data mining that measures people's feelings and sentiments over natural language processing. Data mining is the study of collecting the data from any useful source, process on that data, then analyze, and gain useful information from data [5]. Data mining is mainly used by companies to turn raw data into useful information. Sentiment analysis allows the business to identify the customer sentiment concerning products, brands, or any online services, to determine the neutral positive, and negative text [6].

II. RELATED WORK

Many researchers proposed the analysis of feelings becomes important to identify positive or negative tweets and is responsible for the voice of the consumer [7]. The proposed approach likewise improves feeling arrangement execution [8]. The classification of feelings affects several preprocessing techniques, including deleting URLs, modifying rejection, repeating letters, deleting stop words, deleting numbers, and abbreviations [9]. there are various methods of pre-processing removing URLs, replacing negation, reverting repeated letters, removing stop words, removing numbers, removing white space between sentence's effects on sentiment classification [10]. Researchers proposed sentiments analysis on social media customer reviews for improving the level of services and quality of products [11]. The researcher implemented a multi-class sentiment examination [2]. Researchers worked on sentiment analysis of social and textual information to discover the force of sentiment of extremism [12]. Researchers

proposed a sentiment analysis classification approach useful based on a majority poll of many classification methods. They consider only the text of tweets and other information like the user who tweets them, and the factors are potentially useful [13]. Researchers prepare graphs before and after sentiment knowledge enhancers [14]. The researcher proposed a new approach that gives high precision, recall and FI score, according to the author's this model will not be applicable for further any other languages [15]. Researchers Just overview the different classifiers on the English dataset [16]. Researchers follow the three steps of pre-processing in sequence on Persian language data [17]. The researcher Not define any pre-processing method for cleaning the data [18]. The researcher proposed an analysis of feelings in Arabic, not define any pre-processing method [19]. The researchers' tweets first extracted and then apply the pre-processed method and then categorized them into positive, negative, or neutral sentiments [20]. The researcher discussed the effect of the pre-processing process on sentiment classification and use six pre-processing methods in this article by using two feature models [21]. Conduct sentiment-based filtering implemented for specific types using seed lists to reduce the loss of data [22]. Present the first extraction perform remove noise from data, then corpus development processed, then conversion of extracted data into Arff on roman Urdu language data [23].

III. MATERIALS AND METHOD

This section provides implemented approach that is shown in Fig 1. It describes how sentiments detect in the acquired dataset. Then apply machine learning techniques on the dataset to find the accuracy and find maximum best results. First of all, data was collected, for this purpose already labeled datasets were acquired that are publicly available on the internet. Then pre-processing techniques apply to the data to refine the data. For pre-processing we Remove all special characters, all single characters from the start of comments, Replacing multiple spaces with a single space, Take Off prefixed 'b', and Convert uppercase into lowercase from review or comments and result in store in-process comments.

Dataset-1: The dataset-1 contains data about the user or customer reviews in the Roman Urdu language. The data set consists of 14647 rows (number of records) and two columns one contains comments, and the others contain sentiment. And comments contain 14131 unique values. Each row or record contains two string datatype values. Tagged of sentiments are positive, neutral, and negative. SVM performed better results than others and gain the highest accuracy which is 69% shown in Fig. 2.

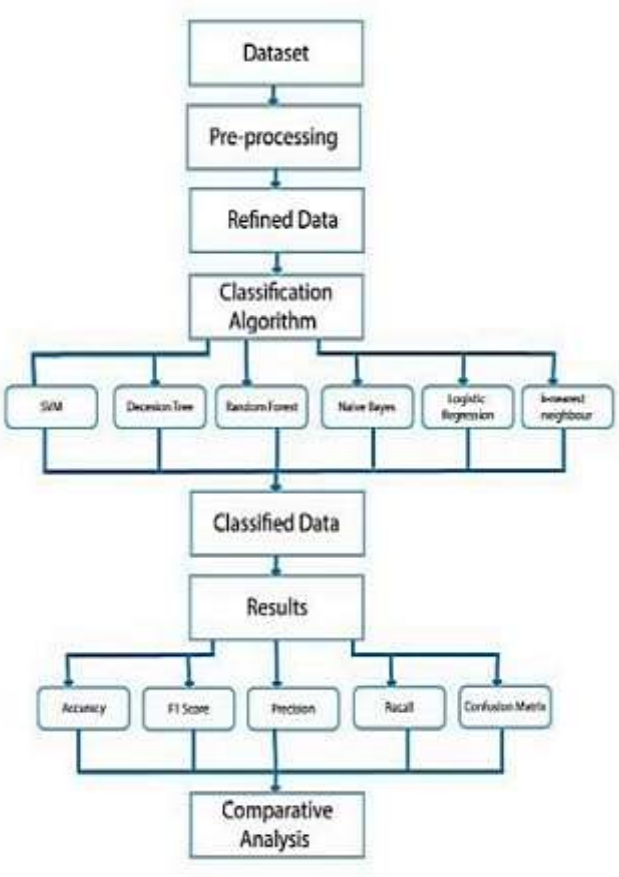

FIGURE 1: Proposed methodology

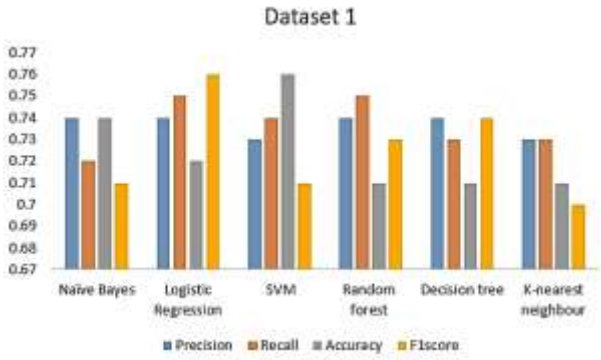

FIGURE 2: Comparative Metrics for Performance of Twitter dataset

Dataset-2: The dataset-2 that we used, obtained from "Kaggle". It contains 1,600,000 tweets. It handles the sentiment analysis of a text that includes more than one language, English and Roman Urdu. In this phase, we clean the data as well as label the data, 0 for negative, 2 for neutral,4 for positive. SVM performed better results than other classifiers and gain the highest accuracy which is 73% shown in Fig. 3

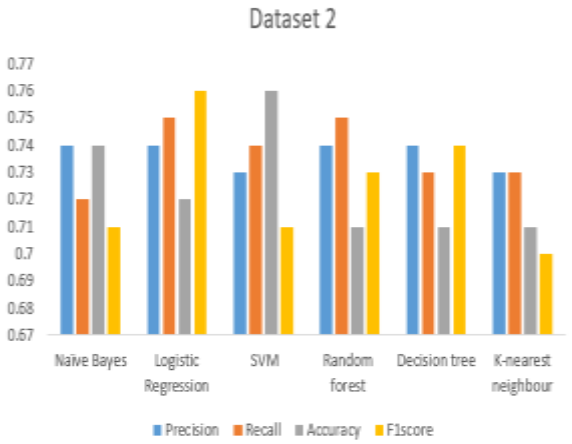

FIGURE 3: Comparative Metrics for Performance of Bilingual dataset.

Dataset-3: Dataset-3 contains 40,000 records, it's Roman Urdu dataset, reviews obtain in five areas: Drama, Web Reviews, Movie and Telefilm, Politics, and assorted. Tagged of sentiments are positive, neutral, and negative. Logistic regression performed better results on this dataset rather than another classifier and gain the highest accuracy which is 76% shown in Fig. 4.

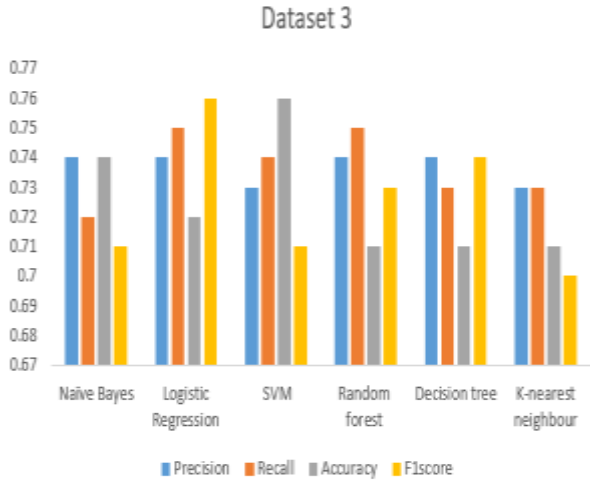

FIGURE 4: Comparative Metrics for Performance of daraz dataset

IV. CONCLUSION

This research performed sentiment analysis in the roman Urdu language using different machine learning classifiers. First, we collected roman Urdu data. Developed the algorithm for pre-processing to remove the noise from the data. To perform pre-processing filter the noise from textual reviews.

In this process, input reviews are processed and get output process reviews. Further, we have analyzed data using machine learning techniques applied to process data. The accuracy in the dataset as well as to measure the precision, recall, and F-measure score of data on the different classifiers. Highest accuracy against dataset-1 is 69%, dataset-2 is 73% and dataset-3 is 76%. We get a different result by comparing the classifier result SVM and Logistic Regression to obtain the highest accuracy on a given dataset.

ACKNOWLEDGMENT

First of all, I thank Allah Almighty for his mercy, and the source of all knowledge, for giving me the courage and knowledge to complete this document. Then also deeply thankful to my parents and my sister for their constant support and belief in me.

REFERENCES

[1] S. B. Moralwar and S. N. Deshmukh, “International Journal of Computer Sciences and Engineering Engineering Open Access Different Approaches of Sentiment Analysis,” no. 3, 2015.

[2] K. Mehmood, D. Essam, and K. Shafi, “Sentiment Analysis for a Resource-Poor Language — Roman Urdu,” vol. 19, no. 1, pp. 1–15, 2019.

[3] R. Khan and S. Urolagin, “Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data,” vol. 9, no. 6, pp. 380–388, 2018.

[4] Z. Sharf and S. U. Rahman, “Performing Natural Language Processing on Roman Urdu Datasets,” vol. 18, no. 1, pp. 141–148, 2018.

[5] M. J. C, J. M. R, V. J. L. Lucero, and S. C. B, “Sentiment and Opinion Analysis on Twitter about Local Airlines,” 2017.

[6] M. Taboada, J. Brooke, and K. Voll, “Lexicon-Based Methods for Sentiment Analysis,” no. December 2009, 2011

[7] [Online] https://en.wikipedia.org/wiki/Lexical_analysis.

[8] [Online]https://en.wikipedia.org/wiki/machineLearning_analysis.

[9] Z. Jianqiang and G. U. I. Xiaolin, “Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis,” *IEEE Access*, vol. 5, pp. 2870–2879, 2017.

[10] I. G. Cahya and P. Yasa, “Sentiment Analysis of Snack Review Using the Naïve Bayes Method,” vol. 8, no. 3, pp. 333–338, 2020.

[11] A. Rane, “Sentiment Classification System of Twitter Data for US Airline Service Analysis,” *2018 IEEE 42nd Annu. Comput. Softw. Appl. Conf.*, pp. 769–773, 2018.

[12] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, “Precise Tweet Classification and Sentiment Analysis,” pp. 1–6, 2013.

[13] Y. Wan, “An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis,” 2015.

[14] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, “Precise Tweet Classification and Sentiment Analysis,” pp. 1–6, 2013.

[15] E. Prabhakar, M. Santhosh, A. H. Krishnan, T. Kumar, and R. Sudhakar, “Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach,” vol. 7, no. 01, pp. 1–3, 2019.

[16] S. Geetha and V. K. Kaliappan, “Tweet Analysis Based On Distinct Opinion of Social Media Users ’,” *2018 Int. Conf. Soft-computing Netw. Secure.*, pp. 1–6, 2018.

[17] E. Asgarian, M. Kahani, and S. Sharifi, "The Impact of Sentiment Features on the Sentiment Polarity Classification in Persian Reviews," 2017.
[18] F. Noor, M. B. B, and J. Baber, *Sentiment Analysis in E-commerce Using SVM on Roman Urdu Text*. Springer International Publishing, 2019.
[19] R. M. Duwairi, "Arabic Sentiment Analysis using Supervised Classification," 2014.
[20] D. D. Das, S. Sharma, S. Natani, and N. Khare, "Sentimental Analysis for Airline Twitter data,"
[21] H. Isah, P. Trundle, and D. Neagu, "Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis."
[22] E. M. G. Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools : An Empirical Study," vol. 112, no. 5, pp. 44–48, 2015.
[23] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naı ¨ ve Bayesian, Decision Tree, and KNN classification techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, 2016

V. COMPARATIVE ANALYSIS

Table 1: Comparative Analysis on the different dataset using machine learning Techniques

Sr: no	About dataset	Approaches used	Precision	Recall	Accuracy	F1 Score
1.	**Social media reviews dataset (Roman Urdu dataset)**	Naïve Bayes	0.64	0.62	0.62	0.61
		Logistic Regression	0.64	0.62	0.62	0.61
		SVM	**0.69**	**0.68**	**0.69**	**0.68**
		K-nearest neighbor	0.64	0.64	0.64	0.64
		Decision tree	0.64	0.61	0.61	0.59
		Random forest	0.64	0.62	0.62	0.61
2.	**Daraz reviews (Roman Urdu dataset)**	Naïve Bayes	0.74	0.72	0.72	0.71
		Logistic Regression	0.74	**0.74**	0.74	0.73
		SVM	**0.75**	0.72	**0.73**	**0.74**
		K-nearest neighbor	0.73	0.72	0.71	0.70
		Decision tree	0.73	0.71	0.71	0.69
		Random forest	0.74	0.72	0.72	0.71
3.	**Roman Urdu and English dataset (Bilingual)**	Naïve Bayes	0.74	0.72	0.72	0.71
		Logistic Regression	**0.75**	**0.75**	**0.76**	0.76
		SVM	0.73	0.74	0.75	0.73
		K-nearest neighbor	0.73	0.72	0.71	0.70
		Decision tree	0.74	0.71	0.71	**0.79**
		Random forest	0.74	0.72	0.72	0.71