Migration Letters

Volume: 21, No: S10 (2024), pp. 631-643

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)

www.migrationletters.com

A Mechanism Of Text Pre-Processing For Sentiment Analysis On The Basis Of Comparative Study

Muhammad Javed[1], Arslan Ali Raza[2], Muhammad Ahmad Jan[1], Syed Muhammad Ali Shah[1], Fareedullah[3], Husnain Saleem[1], Maria Zuraiz[1]

ABSTRACT

Sentiment analysis, also known as opinion mining, plays a crucial role in comprehending the sentiments and attitudes expressed in textual data. With the rapid expansion of social networking platforms in today's global society, effective communication has become paramount. This technology can be leveraged for business development, evaluating social activities, and capturing novel ideas through sentiment analysis. Consequently, preprocessing is an essential task in sentiment analysis. People's reviews hold significant importance in this process. Text preprocessing involves cleansing the textual data and preparing it for efficient information processing within the model. The accuracy and effectiveness of sentiment classification models are greatly influenced by text preprocessing. This study aims to investigate and propose a tailored text preprocessing mechanism designed specifically for sentiment analysis, focusing on a comparative study of existing techniques. The proposed mechanism consists of seven phases: data extraction, noise reduction, Text Preprocessing, language identification (Urdu and English), language translation, sentiment analysis, and scorin[1]g module. Our mechanism is applied to a dataset comprising 8000 tweets related to product reviews in Urdu and English. Several experiments are conducted using an unsupervised lexicon-based approach. This research includes a comparative study of various preprocessing techniques, comparing them with our approach. The results show a significant improvement in accuracy from 61.7% to 93.45%.

Keywords: *Sentiment Analysis, Preprocessing, Preprocessing, English and Urdu Language.*

1. Introduction

Sentiment analysis, also referred to as opinion mining, aims to determine the sentiments and opinions conveyed in textual content. The term sentiment analysis is often used interchangeably with opinion mining by many authors. Social media platforms encompass a vast amount of sentiment data, including posts, tweets, stories, articles, and movie reviews. Sentiment analysis models are widely employed for analyzing movie reviews and product reviews on platforms like Amazon and others. Social media has provided internet users with the opportunity to express and share their emotions and thoughts on various subjects and events. Over the past few years, the influence of social websites such as Facebook and Twitter have significantly grown, playing a pivotal role in the dissemination of information. Users can access substantial amounts of data through these social networks. Twitter, in particular,

[1]Department of Computing and Information Technology, Faculty of Computing, Gomal University, D.I. Khan, K.P.K, Pakistan.
[2]Department of Computer Science, COMSATS University Islamabad, Veharhi 45550, Pakistan.
[3]Department of computer science, University college of Zhob,(BUITEMS),New Appozai,Sambaza Road,zhob, Balochistan, Pakistan.

facilitates early identification of conversations on various topics compared to traditional information channels. Internet users can easily communicate with each other by posting on social media or commenting on someone's post to provide feedback on products or services.
The primary task of sentiment analysis is to classify information as positive or negative using various approaches and techniques (Hardeniya & Borikar, 2016). When it comes to movie reviews, there is a wide range of mechanisms available for expressing emotions, which can be positive, negative, or neutral. In terms of movie reviews, Apache Hadoop outperforms NBC (Narendra, et al., 2016). Sentiment analysis is an NLP task that aims to identify emotional information in a text and determine the sentiment conveyed, whether it is negative, positive, or neutral. Numerous machine learning techniques can be employed to analyze reviews. Supervised learning techniques exhibit 85% accuracy compared to unsupervised learning techniques (K, Rodrigues, & Chiplunkar, 2017).

Text preprocessing is a crucial phase in sentiment analysis and its relevant applications. The application of preprocessing techniques leads to improved accuracy. Data preprocessing is a significant stage in sentiment analysis, as selecting appropriate preprocessing strategies is vital. With the increasing digitization of the world, reading the entirety of social media content has become an arduous task. Hence, sentiment analysis and opinion mining (OM) systems, such as NB and SVM, are employed. NB and KNN, two machine learning algorithms, are utilized for feature selection and classification during the data-cleaning process (Ferdousy, Islam, & Matin, 2013).

In this study, we propose a text preprocessing mechanism for sentiment analysis based on a dataset comprising 8000 tweets related to product reviews in Urdu and English languages. The proposed mechanism consists of seven steps: data extraction, noise reduction, text normalization of cleansed text, language selection (Urdu and English), language translation, sentiment analysis, and scoring module. The research also delves into the detailed aspects of preprocessing, which significantly impacts sentiment analysis. It is observed that the normalization process becomes more complex when comments or reviews on social media are not in a structured form. The effectiveness and accuracy of sentiment analysis heavily rely on the preprocessing phase. If the dataset is properly cleaned, the outcomes will be more accurate. Our research study consists of two phases. In the first phase, we conduct a comparative study of existing research related to language usage in analysis, normalization techniques, and methods employed for conducting opinion mining and sentiment analysis. In the second phase, based on the comparative analysis of existing studies, we present a text preprocessing mechanism. The study demonstrates that preprocessing structured tweets enhances classification accuracy.

The remaining sections of this article are as follows: Segment 2 presents the Literature Review, Segment 3 discusses the Methodology, Segment 4 presents the Results and Discussion, and Segment 5 concludes with the Conclusion and Future Work**.**

2. Related works

An overview of the relevant work is provided in this section related to text pre-processing of sentiment analysis on the basis of a comparative study.
SA is the advanced rearrangement of individuals' perspectives and hypothesis communicated sentiments. It assumes an essential part in navigation. sentiment analysis is likewise vital in arrangement making and reshaping the business system. Albeit much work has been finished in various fields like wellbeing, training, café, legislative issues and item surveys and so on yet there is a desperate need to further develop sentiment analysis appropriately pertinent to preprocessing procedures and dataset groundwork for various dialects

2.1 Approaches for SA and OM

There exist 3 approaches for Sentiment analysis and Opinion Mining which are discussed below;

2.1.1 Machine learning approach (MLA)

The MLA is the main methodology utilized in opinion examination. The computer-based intelligence approach relies upon the famous ML estimations to handle the SA as an ordinary message portrayal issue that uses syntactic as well as phonetic components.

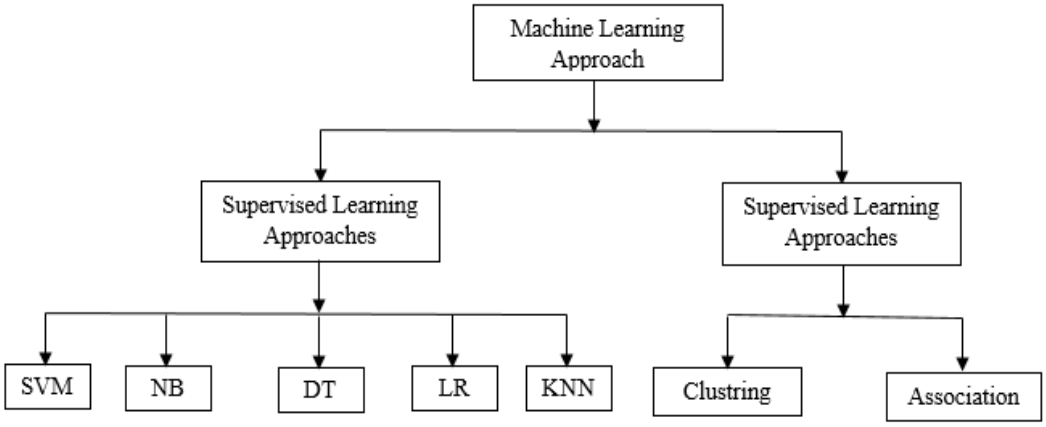

Fig.2.1: Machine Learning Approaches

2.1.2 Rule-based approach

In this methodology, a few circumstances are set and as per these circumstances, activities are performed for feeling examination and assessment mining. The rule-based approach, otherwise called the master framework approach, is a sort of man-made brainpower that utilizes a bunch of in the event that principles to decide or tackle issues.

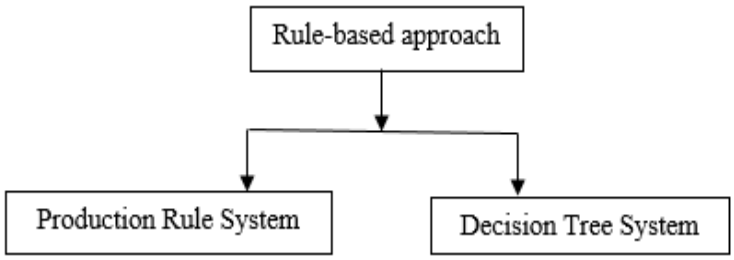

Fig.2.2: Rule-based approach

2.1.3 Lexicon-based Approach

The lexicon-based approach is a sort of NLP that utilizes a pre-characterized set of words or expressions, known as a lexicon or dictionary, to examine text and concentrate importance from it. The lexicon-based approach is based with respect to the possibility that the importance of a word still up in the air by setting and the words encompass it. Lexicon based approach can be classified into two types namely Dictionary based approach and Corpus based approach.

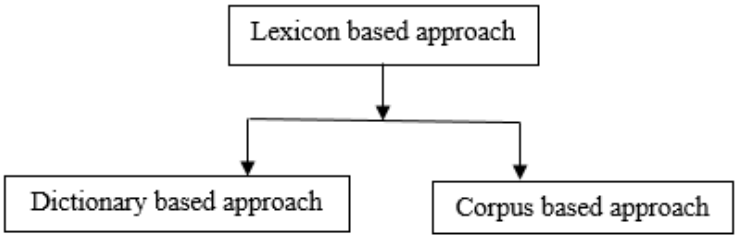

Fig 2.3: Lexicon-based approach

2.2 Comparative study of sentiment analysis

Nowadays information exchanges on Twitter, Facebook and WordPress etc**.** (Albaldawi & Almuttairi, 2020) performed sentiment analysis on the basis of comparative study of corona tweets in Apache spark for huge data analysis. They found that Random-forest out performed in comparison of Support vector machine, Logistic regression and multilayers perceptron. People point of view on different social media websites give useful information about everything. The people's suggestion on social media plays an important role in sentiment analysis. The main problem of SA is sentiment polarity. (NASSR , SAEL, & BENABBOU, 2019) presented a SA on different approaches on social media. They used pre-processing techniques for the success of sentiment analysis specially for those post and comment which was written in un-standards language. It is obvious to achieve high quality output and the input should be in high quality. Sentiment analysis of the majority in terms of product reviews, sentiment analysis of this massively generated data is really helpful.

2.3 Text pre-processing of sentiment analysis

Text preprocessing is a strategy to clean noise in the text information and prepare it to take care of information in the model. Text information contains clamor in different structures like feelings, accentuation, and text in an alternate case. (Sridevi & Velmurugan, 2022) classified Covid dataset of twitter using two machine learning algorithms BaysNet and LIBlinear for two types of data standard and non-standard form of data. (Babanejad, Agrawal, An, & Papagelis, 2020) examined the pre-processing techniques in the emotional examination based on word vector models for exhaustive analysis. They proposed three tasks which were SA, SD and EC of their pre-processing frameworks. (Alzahrani & Jololian, 2021) examined the effect of pre-processing techniques on the gender profile of the author by utilizing a transfer learning model.

2.4 Sentiment Analysis for diverse languages

Nowadays a lot of people by using social media from different cities or countries. It is very complicated to understand their opinions and classify those opinions correctly. (Baly, et al., 2017) described their efforts to develop the first MD-ArSen TD. The dataset was a collection of different tweets, gathered from 12 different Arabic countries. They also performed a comparative study on Egypt and United Arab Emirates tweets by various sentiment method. (Al-Harbi, 2019) proposed many feature selection methods. They used Information Gained (IG), Support Vector Machine (SVM), correlation, chi-square and Gini Index (GI). (BAL & GUNAL, 2022) expected to dissect the effect of normal used features and Pre-processing techniques on the presentation of automatic text abstraction, especially in the Turkish language. (Nisha, et al., 2022) inspected different MLAs on Twitter information. Results showed that the SVM gave higher exactness when contrasted with different classifiers. (Hassan, et al., 2020) proposed the early impact of tweet sentiments connected with research articles by applying LR. (Javed, et al., 2020) proposed a framework to handle Native language roman Urdu text along with English language for efficient Sentiment analysis.

Table 2.1 Sentiment Analysis performed for different languages with various approaches

Studies	Problem addressed	Dataset	Language	Approach	Application	Limitations	Future work
(Hiroshi, Tetsuya, & Hideo, 2004)	SA for machine translation	WWW	English, Japanese	Lexicon and corpus-based.	Used in machine translation.	Develop at low cost.	Work in enhancement of SA.
(Mihalcea, Banea, & Wiebe, 2007)	SA of Multi-word expressions	Romanian newspaper	English, Romanian	Rule-based classifier	Classifying social media reviews.	Limited sources.	Work on large source.
(Denecke, 2008)	Multilingual Sentiment Analysis.	Amazon	German, English	Logistic regression	Determining polarity	This method is only for on language.	Applied in other domains.
(Wan, 2008)	Techniques for Chinese SA.	Chinese IT product web.	English, Chinese.	Lexicon-based.	Checking Chinese's review polarity	Limited resources.	Large dataset of English used.
(Rosell & Kann, 2010)	Construction of Swedish dictionary	People's Dictionary	English, Swedish	Lexicon-based.	Used in people judgment.	polarity classification in Swedish.	Large seed words.

3. Method and Design

Social media are intelligent advancements that work with the creation and sharing of data, thoughts, interests, and different types of articulation through virtual networks and organizations. Social media elevates clients to impart content to other people and show content to improve a specific brand or item. Social media permits individuals to be inventive and share intriguing thoughts with their supporters or fans. Certain virtual entertainment applications like Twitter, Facebook, and Instagram are spots where clients share explicit political or sports content. Numerous columnists and writers produce updates and data on sports and political news. It can really give clients appropriate and important data to keep awake to date on applicable reports and themes. Because of a lot of information, text cleaning is required. Text cleaning here alludes to the most common way of eliminating or changing specific pieces of the text so the text turns out to be all the more effectively reasonable for NLP models that are learning the text. This frequently empowers NLP models to perform better by decreasing commotion in text information.

3.1 Proposed Mechanism for Preprocessing

Based on existing examinations, we have likewise proposed another mechanism of text pre-processing for effective and exact feeling investigation of wanted spaces which comprised of the following steps Data extraction, Noise reduction, Text normalization of cleaned text, Desired language (Urdu) and English, Language translation, Sentiment analysis and Scoring module.

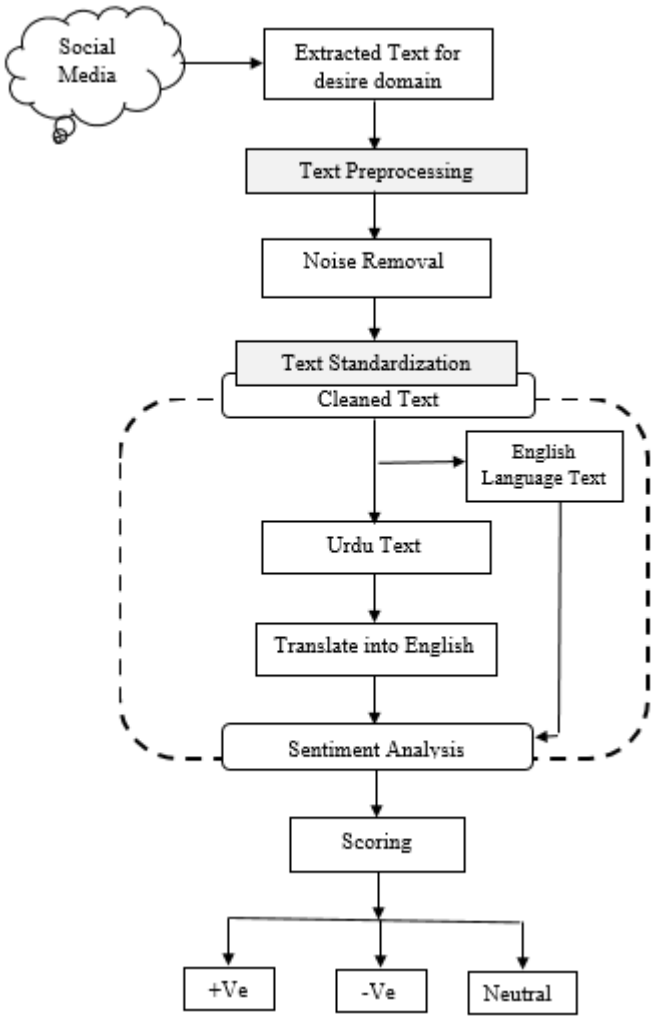

Fig 3.1: Mechanism for Text Preprocessing for Sentiment Analysis

3.1.1 Data extraction

In this phase data is extracted from different source of specific domain for experimental process. We have extracted dataset of product review for analysis through Twitter APIs and 80lexcrawller by using English and Urdu keywords.

3.1.2 Noise reduction

Noise reduction is responsible for dispensing with uproar from the text. Noise reduction incorporates the ejection of undesired and unnecessary pictures and names like URLs, retweet pictures, and exceptional characters. Python apparatus stash is used in the departure of bothersome text. The text isolated from online diversion objections is overflowing with clamor. This text contains URLs, Symbols, undesired highlights and some remarkable correspondence characters for instance @, RT and < >, etc. These symbols and marks have no part in

assessment gathering endeavors so all such kind of complements ought to be shed preceding mining and assessment. In this time of pre-handling, we have utilized HTML parser for contents cleaning.

3.1.3 Desired language (Urdu) and English

In this step of research methodology is to choose the ideal dialects for the opinion investigation task. The two dialects picked for this study are Urdu and English. Urdu is a broadly communicated in language, basically in Pakistan and portions of India, with a huge presence on different virtual entertainment stages and online discussions. English, then again, is a worldwide language and is widely utilized in computerized correspondence around the world. By choosing these two dialects, this examination plans to give bits of knowledge into opinion investigation across various etymological settings.

3.1.4 Language translation

The translation is the method involved with adjusting text from one language into one more to keep up with the original text and correspondence. Translation of one language into one more language is known as code-exchanging. For code-exchanging, we utilize various procedures like machine Translation (MT), or the production of lexical assets like word references and dictionaries. To lead a comparative study of text preprocessing procedures for opinion examination in Urdu and English, language translation is a fundamental part of the methodology. The course of language interpretation includes changing the text data starting with one language then onto the next while protecting the hidden opinion and importance. This examination will give experiences into the viability of text preprocessing methods for opinion investigation in various dialects and distinguish likely difficulties and open doors in cross-lingual sentiment analysis.

3.1.5 Sentiment analysis

Sentiment analysis is the most common way of dissecting advanced message to decide whether the close to home tone of the message is positive, negative, or impartial. Today, organizations have huge volumes of text information like messages, client assistance talk records, web-based entertainment remarks, and audits. the aim of this phase is to analyze the sentiments of text. This work concerns just two classes; they are positive and negative.

3.1.6 Scoring module

In this phase of proposed mechanism where each opinionative token gets weight as indicated by its faculties. At the point when every one of the tokens are scored by their faculties then the entire tweet is marked as positive, negative or Impartial as per the entire score of a tweet at sentence level utilizing lexicon-based approach. On the off chance that the score of a tweet is positive, the extremity of the tweet is Positive and on the off chance that the entire score is negative, the extremity of the tweet is Negative in any case the extremity of the Tweet is Unbiased.

4. Results and Discussion

The sentiment analysis is the purpose in portraying general conclusions as great or critical on the underpinnings of the computational score. Opinion examination or assessment Mining is the field where the framework of public sentiments and notions is made with the ultimate objective of evaluation. The evaluation results propose whether a specific substance is delighted in or despised. In this review, we proposed a framework for better and more viable evaluation of public sentiments and ends conveyed in English and non-English text. This part

covers the results and appraisals made through the proposed framework. The rest of the part elaborate Fragment 4.1 sentiment analysis as investigation district Region 4.2 presents evaluation organizations and execution of proposed design and this section likewise show the near examination of various existing exploration studies.
To evaluate the sufficiency of the proposed instrument broad preliminaries are performed on twitter data of item audits. The data is gathered about Huawei smartphone from available people's reviews by using Twitter APIs. Manual comment is performed to dole out polar classes to each separated tweet so a dataset of 8000 tweets in which 4000 positive tweets and 4000 negative tweets.

Table 4.1: Insights of Tweets Characterization for Product

Domains	**English**		**Non-English**		**Total**
	Positive	**Negative**	**Positive**	**Negative**	
Product	3000	3000	1000	1000	8000

We have surveyed the display of the proposed structure through the confusion table. Table 4.3 shows the disarray network. A line of this table addressed the Veritable Class (Human Remarked on) and sections address Expected Class (Machine explained).

Table 4.3: Product Confusion Matrix

Names	**Confusion Matrix for Product**			
		Machine Anticipated Values		
Human anticipated Value	Classes	Positive	Negative	**Total**
	Positive	3834 (TP)	166 (FN)	4000
	Negative	358(FP)	3642(TN)	4000
	Total	4192	3808	**8000**

We have used four execution estimates Accuracy, Recall, F-measure, precision to evaluate the sufficiency of our proposed research.

4.1 Precision

precision estimates the Accuracy of a classifier. A higher precision implies less misleading up-sides, while a lower precision implies all the more bogus up-sides. numerically precision can be represented as in Eq.4.1.

$$p = \frac{TP}{TP+FP} \qquad \textbf{Eq.}\ 4.1$$

Precision of Product for English Text

$$p\ (\text{Positive}) = \frac{TP}{TP+FP} = \frac{3024}{3834+358} = \mathbf{72.13\%}$$

$$p\ (\text{Negative}) = \frac{TN}{TN+FN} = \frac{3105}{3624+166} = \mathbf{81.53\%}$$

Precision of Product for non-English text

$$p\ (\text{Positive}) = \frac{TP}{TP+FP} = \frac{810}{3834+358} = \mathbf{19.32\%}$$

$$p\ (\text{Negative}) = \frac{TN}{TN+FN} = \frac{537}{3624+166} = \mathbf{14.10\%}$$

Precision of Product for both English and non-English text

$$p\ (\text{Positive}) = \frac{TP}{TP+FP} = \frac{3024+810}{3834+358} = \mathbf{91.45\%}$$

$$p\ (\text{Negative}) = \frac{TN}{TN+FN} = \frac{3105+537}{3642+166} = \mathbf{95.64\%}$$

4.2 Recall

Recall estimates the fulfillment, or responsiveness, of a classifier. Higher recall implies fewer misleading negatives, while lower recall implies all the more bogus negatives. Mathematically recall can be represented as in Eq. 4.2.

$$r = \frac{TP}{TP+FN} \qquad \textbf{Eq.}\ 4.2$$

Recall of Product for English Text

$$r\ (Positive) = \frac{TP}{TP+FN} = \frac{3024}{3834+166} = \mathbf{75.6\%}$$

$$r\ (Negative) = \frac{TN}{TN+FP} = \frac{3105}{3642+358} = \mathbf{77.62\%}$$

Recall of Product for non-English Text

$$r\ (Positive) = \frac{TP}{TP+FN} = \frac{810}{3834+166} = \mathbf{20.25\%}$$

$$r\ (Negative) = \frac{TN}{TN+FP} = \frac{537}{3642+358} = \mathbf{13.42\%}$$

Recall of Product for both English and non-English text

$$r\ (Positive) = \frac{TP}{TP+FN} = \frac{3024+810}{3834+166} = \mathbf{95.85\%}$$

$$r\ (Negative) = \frac{TN}{TN+FP} = \frac{3105+537}{3642+358} = \mathbf{91.05\%}$$

4.3 F1-Measure

The F-Measure is the symphonious mean of positive insightful characteristics (precision) and responsiveness (recall). F-Measure is the weighted symphonious mean of accuracy and Review that surveys the P/R tradeoff. Numerically F1-Measure can be represented as in Eq. 4.3.

$$F1 - Measure = \frac{2PR}{P+R} \qquad \textbf{Eq}.\ 4.3$$

F1-Measure of Product for English text

$$F1 - Measure(P) = \frac{2PR}{P+R} = \frac{2(72.13 * 75.6)}{72.13+75.6} = \mathbf{73.82\%}$$

$$F1 - Measure(N) = \frac{2PR}{P+R} = \frac{2(81.53 * 77.62)}{81.53+77.62} = \mathbf{79.52\%}$$

F1-Measure of Product for non-English text

$$F1 - Measure(P) = \frac{2PR}{P+R} = \frac{2(19.32 * 20.25)}{19.32+20.25} = \mathbf{19.77\%}$$

$$F1 - Measure(N) = \frac{2PR}{P+R} = \frac{2(14.10 * 13.42)}{14.10+13.42} = \mathbf{13.72\%}$$

F1-Measure of Product for both English and non-English text

$$F1 - Measure(P) = \frac{2PR}{P+R} = \frac{2(91.45 * 95.85)}{91.45+95.85} = \mathbf{93.59\%}$$

$$F1 - Measure(N) = \frac{2PR}{P+R} = \frac{2(95.64 * 91.05)}{95.64+91.05} = \mathbf{93.28\%}$$

4.4 Accuracy

In a quantifiable assessment, precision can be described as the quality or exactness of target data or models in term of certifiable worth. The precision is the degree of rightness. Mathematically accuracy can be shown as in Eq. 4.4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \textbf{Eq.}4.4$$

Accuracy of Product for English Text

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3024+3105}{3834+3642+358+166} = \mathbf{76.61\%}$$

Accuracy of Product for Non-English Text

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{810+537}{3834+3642+358+166} = \mathbf{16.83\%}$$

Accuracy of Product for both English and Non-English text

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3024+810+3105+537}{3834+3642+358+166} = \mathbf{93.45\%}$$

Graphically the achieved accuracy for English, Non-English and for Combined tweets is shown below in Fig. 4.1

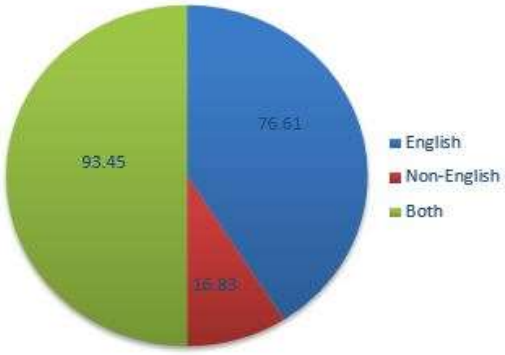

Fig.4.3 Accuracy for English., Non-English and Both tweets

Graphical representation for positive and Negative instances for English, Non-English and for both languages can be depicted as Fig. 4.2 and Fig. 4.3.

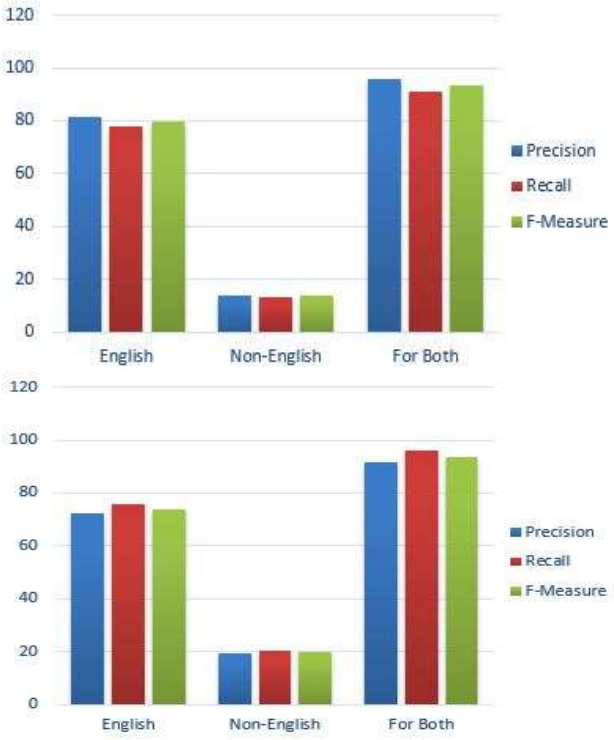

Fig.4.2: Negative instance

Fig.4.3: Negative instance

4.5 Comparison with Existing Studies

Keeping in view the goal issue and datasets in proposed study, our framework is separated in connection with existing assessment as shown underneath in Table 4.5 and it is seen that our

proposed framework outmaneuvered the others opinion examination system over English and non-English tweets for item space.

Table.4.5: Comparative study of proposed mechanism

Research study	Precision		Recall		F1-Measure		Accuracy
	positive	negative	positive	negative	positive	negative	
(Wilson, Wiebe, & Hoffmann, 2005)	63.4%	64.7%	59.3%	83.9%	61.2%	73.1%	61.7%
(Baqapuri , 2015)	70%	61%	66%	59%	68%	60%	64%
(Javed & Kamal, 2018)	81.85 %	82.87 %	83.14 %	81.57	82.48 %	82.21 %	82.35%
(Banjar, Ahmed, Daud, Abbasi, & Dawood, 2021)	72.12 %	74%	75.3%	66%	73.67 %	69.83 %	86.5%
Proposed research	**91.45 %**	**95.64 %**	**95.85 %**	**91.05 %**	**93.59 %**	**93.24 %**	**93.45%**

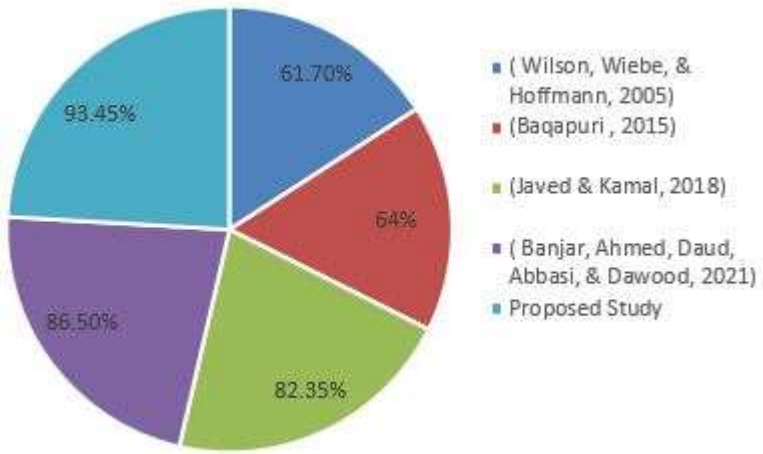

Fig.4.4 Comparative analysis of existing study with proposed research

5. Conclusion

This study intended to propose a mechanism of text preprocessing for SA through a comparative study. The goal was to assess different preprocessing strategies and their effect on SA execution. By leading a far-reaching investigation and correlation of different preprocessing strategies, we have acquired significant bits of knowledge into their viability and materialness. The consequences of our comparative study show that text preprocessing assumes an essential part in opinion examination errands. Through broad trial and error and assessment, we dissected the effect of every procedure on opinion examination precision, accuracy, recall, and F1-Measure. Our discoveries uncover that different preprocessing strategies show fluctuating levels of impact on feeling examination execution. The results of this comparative

study have down to earth suggestions for opinion examination assignments, empowering scientists and professionals to arrive at informed conclusions about the preprocessing procedures to utilize. By executing a fitting preprocessing system, opinion investigation models can be improved, prompting more precise and dependable feeling order results.

Notwithstanding the critical commitments of this review, there are a couple of limitations that ought to be recognized. To begin with, the similar review directed in this exploration zeroed in on a particular arrangement of preprocessing strategies ordinarily utilized in feeling examination. There might be other preprocessing techniques that were not thought of, which might actually yield various outcomes. Second, the near study was directed utilizing a particular dataset or datasets. Third, the assessment measurements utilized in this concentrate principally centered around feeling examination precision, accuracy, recall, and F1-score. While these measurements give significant experiences into the exhibition of the preprocessing strategies, they may not catch the subtleties and nuances of opinion investigation completely. Future exploration could consider consolidating extra assessment measurements, for example, opinion power investigation or viewpoint-based feeling examination, to give a more far-reaching assessment of the preprocessing methods.

SA isn't restricted to English text; it stretches out to different dialects also. Future exploration could examine the adequacy of various preprocessing methods for SA in multilingual situations, taking into account the exceptional etymological attributes of every language. There is test space for future exploration to dive further into text preprocessing for SA. By tending to the impediments of this review and investigating new headings, analysts can keep on working on the exactness and strength of feeling examination models, adding to progressions in the field of regular language handling.

References

Albaldawi, W. S., & Almuttairi, R. M. (2020). Comparative Study of Classification Algorithms to Analyze and Predict a Twitter Sentiment in Apache Spark. 2nd International Scientific Conference of Al-Ayen University (ISCAU-2020) (pp. 1-14). Hillah: IOP Publishing.

Al-Harbi, O. (2019). A Comparative Study of Feature Selection Methods for Dialectal Arabic Sentiment Classification Using Support Vector Machine. IJCSNS International Journal of Computer Science and Network Security, 19(1), 167-176.

Alzahrani, E., & Jololian, L. (2021). How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Author. 3rd International Conference on Machine Learning & Applications (pp. 01-08). Toronto: Computer Science & Information Technology (CS & IT).

Banjar, A., Ahmed, Z., Daud, A., Abbasi, R. A., & Dawood, H. (2021). Aspect-based sentiment analysis for polarity estimation of customer reviews on Twitter. Computers, Materials & Continua, 67(2), 2203-2225.

Babanejad, N., Agrawal, A., An, A., & Papagelis, M. (2020). A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5799–5810). Toronto:

BAL, S., & GUNAL, E. S. (2022). The Impact of Features and Preprocessing on Automatic Text Summarization. ROMANIAN JOURNAL OF INFORMATION, 25(2), 117–132.

Baqapuri , A. I. (2015). Twitter sentiment analysis. arXiv preprint arXiv, 1509.04219, 9-53.

Baly, R., Moukalled, R., Aoun, R., Hajj, H., Shaban, K. B., El-Hajj, W., & El-Khoury, G. (2017). Comparative evaluation of sentiment analysis methods across Arabic dialects. Procedia Computer Science. 117, pp. 266-273. Dubai: Elsevier B.V. Association for Computational Linguistics.

Denecke, K. (2008). Using SentiWordNet for Multilingual Sentiment Analysis. In 2008 IEEE 24th international conference on data engineering workshop (pp. 507-512). Cancun, Mexico: IEEE.

Ferdousy, E. Z., Islam, M. M., & Matin, M. A. (2013). Combination of Naïve Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models. Computer and Information Science, 6(3), 48-56.

Hardeniya, T., & Borikar, D. A. (2016). An Approach To Sentiment Analysis Using Lexicons With Comparative Analysis of Different Techniques. IOSR Journal of Computer Engineering (IOSR-JCE), 18(3), 53-57.

Hassan, S. -U., Aljohani, N. R., Idrees, N., Sarwa, R., Nawaz, R., MartínezCámara, E., . . . Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in Twitter. Knowledge-Based Systems, 192, 105383.

Hiroshi, K., Tetsuya, N., & Hideo, W. (2004). Deeper Sentiment Analysis Using Machine Translation Technology. . In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics (pp. 494-500). Tokyo: Association for Computational Linguistics.

Javed, M., & Kamal, S. (2018). Normalization of unstructured and informal text in sentiment analysis. International Journal of Advanced Computer Science and Applications, 9(10), 78-85.

Javed, M., Ziauddin, Kamal, S., Nasir, J. A., Raza, A. A., & Habib, A. (2020). Socio monitoring framework (SMF): Efficient sentiment analysis through informal and native terms. International Journal of Advanced and Applied Sciences, 7(12), 113-126.

K, B. B., Rodrigues, A. P., & Chiplunkar, D. N. (2017). Comparative Study of Machine Learning Techniques in Sentimental Analysis. In 2017 International conference on inventive communication and computational technologies (ICICCT) (pp. 216-221). Coimbatore, India: IEEE.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 976–983). Prague, Czech Republic: Association for Computational Linguistics.

Narendra, M. B., Uday Sai, M. K., Rajesh, M. G., Hemanth, M. K., Chaitanya Teja, M. M., & Deva Kumar, M. K. (2016). Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies. International Journal of Intelligent Systems and Applications, 8(8), 66-70.

NASSR , Z., SAEL, N., & BENABBOU, F. (2019). A comparative study of sentiment analysis approaches. In Proceedings of the 4th International Conference on Smart City Applications (pp. 1-9). Morocco: Association for Computing Machinery.

Nisha, K. A., Kulsum, U., Rahman, S. u., Hossain, M. F., Chakraborty, P., & Choudhury, T. (2022). A comparative analysis of machine learning approaches in personality prediction using MBTI. In Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021 (pp. 13-23). Singapore: Springer.

Rosell, M., & Kann, V. (2010). Constructing a Swedish General Purpose Polarity Lexicon Random Walks in the People's Dictionary of Synonyms. In Proceedings of Swedish language technology conference (pp. 19-20). Stockholm: Association for Computational Linguistics.

Sridevi, P. C., & Velmurugan, T. (2022). Impact of Preprocessing on Twitter Based Covid-19 Vaccination Text Data by Classification Techniques. Proceedings of the International Conference on Applied Artificial Intelligence and Computing ICAAIC 2022 (pp. 1130-1136). Chennai: IEEE.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of human language technology conference and conference on empirical methods in natural language processing (pp. 347-354). Canada: Association for Computational Linguistics.

Wan, X. (2008). Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 553-561). Honolulu: Association for Computational Linguistics.