SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Sentiment Polarities Detection from Large Heterogeneous Textual Datasets Using Natural Language Processing

Ganesh Kumar[1,*], Shuib Basri[1], Abdullahi Abubakar Imam[2], Sunder Ali Khuwaja[3], Abdullateef Oluwagbemiga Balogun[1], Hussaini Mamman[1]

[1] School of Engineering and Technology, Sunway University, No. 5, Jalan Universiti, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia
[2] School of Digital Sciences, Universiti Brunei Darussalam, BE1410, Brunei Darussalam
[3] Department of Telecommunication Engineering, Faculty of Engineering and Technology, University of Sindh, Jamshoro 76090, Pakistan

ABSTRACT

Various activities based on text data performed every day and in return it is producing a huge volume of textual data in every second. As data grows very fast and handling such type of voluminous data creates many challenges for data analysts and stakeholders. Information retrieval from user actions, emotions, and sentiments helps in the prediction of future growth and decisions. Extraction of Information from large textual datasets containing sentiment polarities, expression, and concern is found critical. To solve these issues, many approaches have been implemented by practitioners such as processing of textual data, findings the trends based on surveys, and interviews from the potential users. Most common preprocessing method adopted for textual dataset is use of natural language processing (NLP). It comprises of multiple steps such as tokenization, lemmatization, stemming, parts of speech removal. The performance of information retrieval techniques plays an important role in big data analytics and must be utilized properly at the time of implementation. In this paper, we used two heterogeneous textual datasets to detect the polarities (positive and negative) from daily emotional dialogs, daily actions and unique words emotions of ACE2020 and Sarcasm datasets. Experiments were conducted on python notebook for preprocessing and polarities detection. Results show that both positive and negative emotions have a great effect on decision making. The performance of NLP on detection of polarities from larger the datasets is promising with better precision, recall, and accuracy score.

Keywords:
Textual datasets; big data; heterogeneous data

1. Introduction

Big data is dealing huge and volumetric data which is produced by industries in unstructured format. For fetching information and retrieving useful knowledge data need to be preprocessed with NLP techniques and big data tools. Data generated by domains such as social media, entertainment, sports, tourism, hospitality, and hoteling in textual format and in the application

* *Corresponding author.*
E-mail address: ganesh_17005106@utp.edu.my

https://doi.org/10.37934/araset.60.1.137149

areas such as sentiment analysis, emotions, actions expressions and behaviors [1]. In addition, data is mainly processed to fetch information from product, process, materialistic thing, or context [2].

The like and dislike polarities are classified in [3], such as positive, negative, smiley, laugh, love, wint, frown, cry, elongated word like yummy, wow, which helps in prediction, decision making, challenges and promotions. Information can be retrieved only be achieved by preprocessing the textual data using NLP techniques such as tokenization, POS tagging, lemmatization, stemming and stop words removal [4].

With the advancement in Artificial Intelligence, Machine Learning and Text Mining, the systems are tending towards automatic and digitization without involvement of humans [5]. NLP algorithms for large textual dataset was developed to measure the performance, complexity, and credibility of computational power [6]. Many researchers proposed ideas related to emotion and sentiment but there is no such comparison proposed for large heterogeneous datasets. In this paper, the performance of NLP techniques on multi-class heterogeneous datasets is proposed to identify the polarities and their effects on precision, recall and accuracy for large heterogeneous textual datasets as well as on decision making.

For further understanding about polarities extraction from heterogeneous textual datasets this paper is organized as follows: Section discusses on literature review on big data, information retrieval, text preprocessing and performance measurement techniques for large textual datasets; section 3 describe the experiments conducted on heterogeneous datasets, section 4 discusses the results and performance measurement and in the last section the conclusion and future work are briefly described.

2. Literature Review

2.1 Big Data

"Big Data" refers to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Various industries with heterogeneous data are facing problems related to storing, managing, and analyzing of large amount of data. Big Data plays an important role in retrieving useful information from the large datasets with the help of advanced tools and algorithms [7, 8]. Nowadays, data produced in formats such as structured, semi-structured and unstructured data from a multidimensional nature of resources and applications that cannot be processed through simple tools. According to the IDC report, by 2020 the size of data will reach about 44 zettabytes and unstructured data account for 95% of global data with an estimate of the compound annual growth rate of 65% [9].

Many definitions have been proposed for big data, in [10], "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and analysis". In [11], Jacobs provided a meta- definition of big data which refers to "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time". The authors in [12, 13] also adopted a generic definition of big data which refers to "data that's too big, too fast or too hard for existing tools to process". Wu *et al.,* [14] proposed the HACE Theorem which defines big data as "large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data".

In general, Big Data can be explained according to three V's: Volume, Velocity and Variety [7] Also, the characteristics of Big Data described in [10] are volume, variety, velocity, veracity, valence, and value. Later on, in [15] 10V's volume, variety, velocity, veracity, variability, viscosity, volatility, viability, validity, and value their issues and challenges are highlighted in Table 1.

Table 1
Big Data characteristics, issues and challenges [15]

BD characteristics	Issues and challenges
Volume	Data scale
Value	Data usefulness in decisions
Variety	Data heterogeneity, SSU
Velocity	Data preprocessing
Veracity	Data quality and accuracy
Viscosity	Data complexity
Variability	Data flow inconsistency
Volatility	Data durability
Viability	Data activeness
Validity	Data properly understandable

Heterogeneous types of data formed in three types Structured, Semistructured and Unstructured (SSU) [16, 17]. Structured data is organized data in a predefined format and stored in tabular form whereas semi-structured data is a form of data which cannot be queried as it does not have a proper structure which confers to any data model and unstructured data is heterogeneous and variable in nature such as text, audio, video, and images. Due to heterogeneous data, it cannot be processed with simple tools and techniques which creates the problem heterogeneity [18, 19] in result, decision maker cannot make decision based on scattered data.

Big Data management was a challenging process, particularly when diverse data sources are employed to collect information for strategic planning and decision-making was experienced by [20]. Around 75% of enterprises use at least one kind of Big Data. In terms of data fusion complexity, data storage, analytic tools, and governance deficiency, the Big Data management domain posed new obstacles.

Moreover, according to authors in [21], the data are not only vast in size, but also in variety. Data generated by operational, transactional, sales, and marketing departments. In addition, big data includes a range of data kinds, such as texts, sounds, video, and images. This unorganized data is growing faster than huge amounts of data and comprises 90 percent of all data. Therefore, new types of processing skills are necessary for gaining data-driven insights that lead to improved decision making. In addition, "Big Data" refers to the technique necessary to manage vast volumes of data with the appropriate tools and techniques. Big Data is most accurately defined as datasets whose properties exceed the computational resources of frequently used software and hardware in a reasonable amount of time.

Big Data was concerned with how this data might be stored, processed, and analyzed so that they could be used to predict the future course of action with a high degree of precision and acceptable latency. The focus of marketers, insurance companies, and healthcare providers is on delivering high-quality, cost-effective treatment to patients, respectively [19].

Despite significant advances in data processing, gathering, assessment, and algorithms related to forecasting people's behavior, it was essential to understand the underpinning driving and controlling aspects that can aid in the design of robust models capable of handling large amounts of tweets and making accurate predictions from tweets [22].

Lee [23] also highlighted that utilizing Big Data Analytics along with its infrastructure for parallel computing and blocking-like strategies, one may achieve high levels of productivity. Successful data analytics needs access to semantically dense data that connects all relevant information for a particular analytical task.

The large dimensions and size of big data provide obstacles to its display. The primary objective of data visualization is to effectively communicate knowledge via the use of diagrams; to facilitate the flow of information to the user, hidden knowledge in complicated and large-scale datasets was made apparent. Nevertheless, due to the vast quantity and high dimensions of big data, data visualization in big data applications may be challenging to handle [24]. The studies demonstrate the following issues that sectors face: predictive analysis, social media analytics, content-based analytics, text analytics, audio analytics, and video analytics [25].

Textual data is produced by news, healthcare, education, sports, agriculture, Internet of Things (IoT) industries in massive amount. To process and retrieve information from it is a challenging task and to solve this issue latest tools and techniques are needed. Information can be in form of user feedback, sentiment, emotions, dialogs, and raw data.

2.2 Information Retrieval

Textual data comprised of unstructured data and information retrieval is important procedure to follow for fetching data from huge datasets. Information can be in many forms and varies according to situation and type. Mainly, the domain which contains information are sentiments, user feedback, news, dialogs, emotions, action, and expression. Sentiment analysis related information can be fetched from social media where sentiment polarities are positive and negative as well as type of data are textual, and image based [26].

Emotional recognition of location based on social networking where public wants to visit. Decision made based on reviews or recommendation posted on social networks. The technique was developed with NLP techniques, prediction algorithms for emotions and sentiments [27]. Fake news detection is also an example of sentiment analysis and information retrieval when a fake news was publishes in print and paper media about any popular personality. The emotion of public can be predicted using proposed system. Normally news related to any person can be viral through social media such as Twitter, Facebook, and Instagram but we get retweets of authorized person if it fake or misinformation went viral [28]. Twitter is widely used social media network and trusted by public, celebrity, politicians, and sport person. It contains very meaningful information especially emotions, sentiments, or expressions, and the structure of tweet data contains some unwanted data which need to removal before going to implement for information retrieval.

Contextual model developed for positive, negative, happy, anger, sad and neutral polarities in multimodal formulation of data [29]. Affective computing is used in textual data to modify and make system intelligent with respect to feelings, inferring, human expressions, actions, and emotions. Sentiments and emotions are processed through different tools and techniques for information retrieval [30]. Affective computing also helps textual data to retrieve information from domains such as product reviews, tweets, opinion mining, mobile product reviews, and online hotel reviews.

In addition to that, few domains were highlighted such as automated depression diagnosis, suicide emotions, customer sentiment, irony, and sarcasm, hate speech, lie detection, stress detection, cognitive assistants, affective intelligent tutoring systems as current domains where information retrieval plays a vital role [31]. Sentiment analysis of twitter was predicted for checking positive negative polarities [32]. Textual data-based emotions and similarity of movie content detected. By using information retrieval from movie content and similarity of content can predicted using multimodal approach [33].

2.3 Text Preprocessing

Natural language (NL) is a way to communicate information among the users using the protocols set. For text-based implementation of algorithms and information retrieval process, Natural Language processing (NLP) is widely used in many applications and domains. In UML class diagram, data extracted with the help of NLP techniques [34]. NLP is wide process and implemented when large textual data contains unwanted data to make it in useful form. For data processing with the help of NLP techniques such as stop words removal, stemming, lemmatization, vectorization, parts of speech removal were implemented to make data in a form so that tweets be classified into categorial format [35].

Lexical analysis of twitter data was compared in proposed model. Comparison between sent wordnet and wordNet were taken place after the preprocessing techniques used for feature extraction [4]. The twitter data has some standard format which contains unwanted data which needs to be preprocessed before retrieving information. In data acquisition, data is preprocessed with unigrams, parts of speech removal, negation, count of emoticon features such as smiley, laugh, love, wint, frown, cry, count of elongated word yummy. Also contains length of capital letters, ensemble learning and tokenization [3].

Health care data contains more useful data which is needed during the emergency as it deals with matter of precious health. Healthcare practitioners needed useful information and to retrieve this, it must process with latest tools and techniques of data processing. Basic techniques used here were discussed. Also, it will help in prediction of the values for decision making, pattern discovery and trend of genetic history [36].

In preprocessing, special characters and punctuations, stop words, and stemming were performed. NLP techniques implemented for text Analytics [37]. In roman Urdu sentiment analysis, the polarities containing dataset were preprocessed to get context of roman Urdu words communicated between peoples. The preprocessing process was based on removal of noise, punctuation, URL, space, newline, and spelling consistency [38].

Educational data contains information of student academic record, teacher's assessment, and research articles. Preprocessing performed to make data available in a form so that score related information can be used [39].

In Bengali language preprocessing techniques were used to create forms from large textual datasets [40]. Also, amazon product reviews were preprocessed with NLP techniques before presented to managers to make decision based on the reviews given by the customer [41].

2.4 Performance Measurement of Heterogeneous Datasets

Performance assessment is used to check and validate the performance of tools, techniques, algorithm, and model which have been went through different rounds of implementation, training, and testing. It helps the managers and data analyst to check the credibility of inputs given an output produced by assessing the computational power of it, the growth of advanced tools is calculated and widely used for many applications and domain.

Performance evaluation of techniques and tools used for large textual file emphasize on optimization level of computation power. These large datasets are treated in many ways and featured with respect to performance. Normally the performance measuring techniques used in [42-46] are precision, recall and accuracy. To analyze, the text document clustering the state of the

art algorithm was used and the performance was measured through precision, recall and accuracy [42].

Twitter data is used for emotion and sentiment analysis from user's comments in text format. For information retrieval data preprocessed through NLP techniques and supervised features extraction tools. The output of extraction and time was evaluated using precision recall and accuracy [43]. In large files of scene detection from textual data, there is combination of many stages from data generation, preprocessing and implementation [44]. Also, many scenes lost from the e-commerce if proper techniques are not implemented. To overcome this issue, a technique is proposed to measure the performance of model. The outcome shows that the precision, recall, and F-score were more than 78%. Accuracy, precision, recall and accuracy for online commerce data were above 91% for Naïve base and 83% for SVM which is based on customer reviews [45].

Electronically medical record is composed of many modules and sub modules, for data processing and text mining on huge volume of data needs time to fetch and present data. The performance measure techniques help in training and test of scienceIE dataset which in result shows variable values of precision recall and accuracy [46]. Also, the researchers in [48-52] produces a quality result in the domain of machine learning and data science.

3. Methodology

In this section the methodology of polarities detection from large heterogeneous textual datasets are described. Firstly, the detailed information about selected heterogeneous datasets are discussed and then the detailed procedure of preprocessing and extraction process of polarities are described.

3.1 Datasets

Two heterogeneous datasets in textual format used as input datasets, among them the ACE 2020 dataset is structured, Sarcasm headline dataset is semi-structured format. The ACE 2020 dataset is in XLS format which contains the information about the news from channels it was produced and recorded and in that the purpose of the news is labelled as text-target. In text samples label, the detailed information and content of news is stored. This data set comprises of 621 news of different categories as shown in Figure 1.

	text_sample
2	... junk messages. Company officials disagreed that AOL's market share was keeping out competitors. AOL executives cited a recent study by Media Metrix indicating that the messaging s growing in the United States. Why all the fuss over a free product that anyone, even those who don't subscribe to AOL, c
3	... (AP) _ Tornadoes destroyed homes and overturned cars in several areas of Alabama on Saturday and more than two dozen people were reported injured. At least one person was spokeswoman for the Tuscaloosa County Emergency Management Agency. Twisters struck just before noon in the state's rural southeaste
4	... Bandar Seri Begawan 11-15 (AFP) - American President Bill Clinton attempted today Wednesday to reassure money markets abou
5	... COSTA MESA, Calif. (AP) _ Joseph Conrad Parkhurst, who founded the motorcycle magazine Cycle World in 1962
6	... Philippines president, Joseph Estrada, was impeached last week, while Chen Shui-bian, the president of Taiwan, is trying to fend off a recall vote. Even in more stable countries, like Mal concerned by the haphazard pace of economic reforms. In many countries, the banks and corporations remain financially sha
7	... DENVER (AP) _ The Denver Broncos and the NFL want a commercial real estate agent to take down a sign advertisi American ...
8	... southward from the storm system, spreading scattered show showers and freezing rain into western sections of Kentucky and Tennessee. Farther south, scattered rain showers extende eastern Texas. Those showers were expected to move eastward into Alabama, Georgia, and northwestern Florida during the nigh
9	... deer from trees. The Michigan season, one of the busiest, continues through Nov. 30. Firearm hunting from elevated platforms has long been legal in other states with heavy hunting, li only in its third season in Michigan, and officials say it appears that hunters still have a lot to learn about safety. Lt. Suzanne
10	... bamboozled his Time Warner counterpart, Gerald Levin, into taking AOL's shares at the worst possible time. ``Short term, anyone objective would say probably Steve got the better p president at State Street Research in Boston and a longtime media investor. But Haverty, like many other Time Warner shareholo
11	... some key economic reports come out this week. kitty pilgrim has more in today's edition of &quot ahead of the curve. &quot wall street will be on fed watch this week. on tuesday, the final time this year to discuss interest rates. recent economic reports indicate the economy is slowing, leading analysts to believe
12	... officials say. The death toll from the raging floodwaters and fearsome mudslides reached 37 combined for the Alpine region of northern Italy and southern Switzerland. Eighteen of those the church of San Maurizio in Fenis, a mountain hamlet near the city of Aosta, villagers paid their final respects to five local vi

Fig. 1. ACE2020 dataset (sample)

Sarcasm headline dataset is in JSON format which contains the information about the news headline in semistructured format. This dataset comprises of 26709 lines and 75KB file size, as shown in Figure 2.

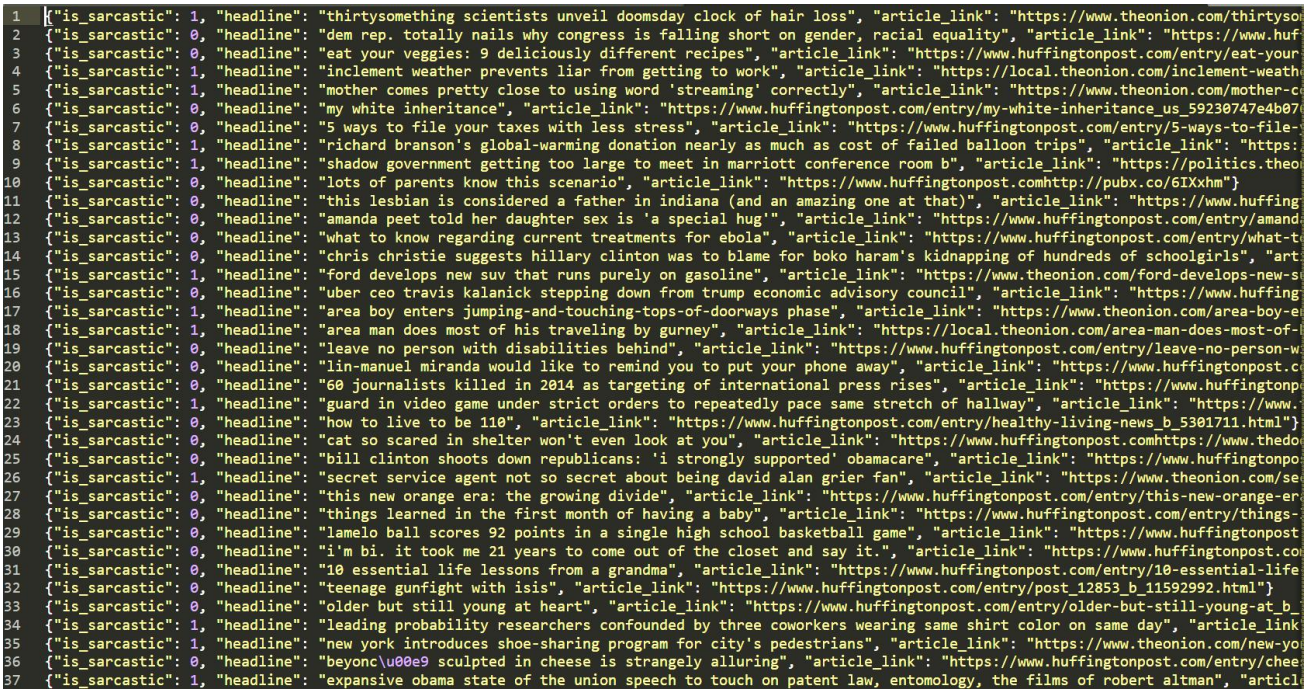

Fig. 2. Sarcasm dataset (sample)

3.2. Experiments

Experiments were conducted on the input datasets i.e., ACE2020 and Sarcasm news. Firstly, the textual data samples from the input datasets are pre-processed by adopting the NLP methods i.e., tokenization, stop words removal, stemming, lemmatization, parts of speech removal to make data in a structured form as shown in Figure 3.

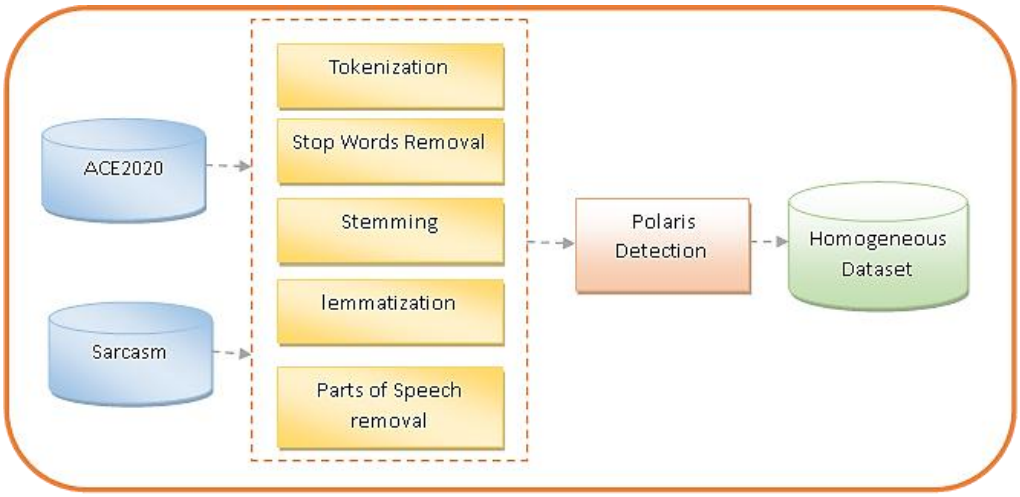

Fig. 3. Architectural diagram for polarities detection technique

For pre-processing of heterogeneous datasets different NLP steps are carried out with the help of efficient methods. After converting large textual datasets into rows and in form of vectors, the next steps were carried out to produce tokens for that punktsentenceTokenizer was used.

Afterwards, the stops words using are removed using NLTK WordNet to prepare datasets for identifying the stems and lemma of the emotions. For identifying stem, PorterStemmer method was used. In addition to that, WordNetLemmatizer is used to extract the lemma of the words identified in input datasets. Lastly, parts of speech tagging were performed.

After pre-processing the heterogeneous datasets, the positive and negative polarities are extracted from the input datasets. It's worth noting here that the Sarcasm dataset is in semi-structured format, and it contains keys (headlines) and values (sarcastic or not). Positive polarity is classified as not sarcastic headlines and negative polarity classified as sarcastic headlines. Experiments are conducted to detect the polarities from daily dialogue action, daily dialogue emotion and unique words classes of input datasets. The results of the input dataset's classes are presented in the following section.

4. Results

Information retrieval techniques had been used by many researchers for diverse problems and data but for textual datasets having large number of sentiments, emotions, routine dialogs, and unique words identification are left behind. Also, the performance measure of polarities included in textual datasets needed to be measured for evaluation of computational cost.

In this study we selected two heterogeneous textual datasets containing sentiment polarities from various sources of news to extract the polarities based on the classes (daily dialog emotion detection, daily dialog action and Unique words). The textual dataset contains huge number of unwanted data which includes parts of speech, prefix, suffix, URL, special words, and spaces which needs to be removed from these datasets. For analysis of performance of NLP methods, various measurement techniques such as precision, recall and accuracy are used.

Basically, two polarities positive and negative were selected to check the performance of polarities. After adopting NLP steps such as, tokenization, stop words, stemming, lemmatization, parts of speech removal the positive and negative polarities were detected. Information containing sentences, further evaluated with true positive, true negative, false positive and false negative with respect to threshold. The predicted and detection of sentences of both textual datasets presented below.

Overall results are divided with threshold to make a section between positive and negative polarities. The whole mechanism of selection of positive and negative value for TP, TN, FP, and FN helps in measurement of performance assessment. After all process precision, recall and accuracy are calculated among all values predicted as shown in Figures 4, 5, and 6.

To compare the existing techniques with proposed technique, a summary of performance of existing technique is presented here. In [22], the accuracy of the proposed technique is 55.27 % on Bengali tweets using LSTM. On the other hand, roman Urdu corpus was classified in [38] using KNN, DT, RF and SVM for six classification and results shows that the highest accuracy among all was SVM with 69%. Additionally, semantic, and linguistic based short answer technique was adopted in [39] using LSTM model and the quadratic weighted Kappa on ASAP dataset was 0.76.

Similarly, other counterpart in [40] used supervised ML algorithms to classify the Benali language and results shows that the accuracy of 99.5%. lastly, another study in [41] uses ML models such as Naïve Bayes, KNN, Logistic Regression, Decision Tree and Random Forest and results shows that the accuracy of random forest on Bag of words, TF-IDF and Word 2 Vec is better. But the existing techniques [22 ,38-41] have many limitations based on their scope. All five techniques are only focusing on homogeneous dataset only, the classes are very limited, and datasets contains very less data. To overcome these issues, a technique is proposed which accepts heterogeneous

datasets, classes of daily dialog emotion, action and unique words as well as focuses on positive and negative polarities. The sample of data in both datasets are more than 29000. Detailed performance of proposed technique is presented below.

The results show that the large textual data having more sentiment polarities perform well. Performance measurement of daily dialog emotion detection on ACE2020 dataset such as precision is about 70%, accuracy is 80% and recall 90% for positive polarity and for negative polarity accuracy is 70%, precision is 60% and recall is 80%. Performance measurement on sarcasm dataset such as precision is about 60%, accuracy is 70% and recall 80% for positive polarity and for negative polarity accuracy is 60%, precision is 50% and recall is 70% as shown in Figure 4.

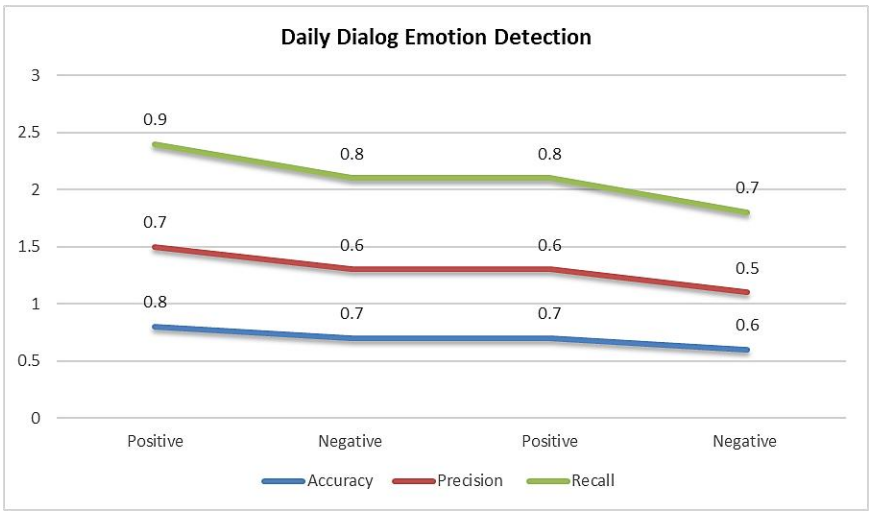

Fig. 4. Daily dialogue emotion detection

Performance measurement of daily dialog action detection on ACE2020 dataset such as precision is about 90%, accuracy is 80% and recall 85% for positive polarity and for negative polarity accuracy is 80%, precision is 70% and recall is 75%. Whereas performance measurement on sarcasm dataset such as precision is about 85%, accuracy is 75% and recall 80% for positive polarity and for negative polarity accuracy is 75%, precision is 65% and recall is 70% as shown in Figure 5.

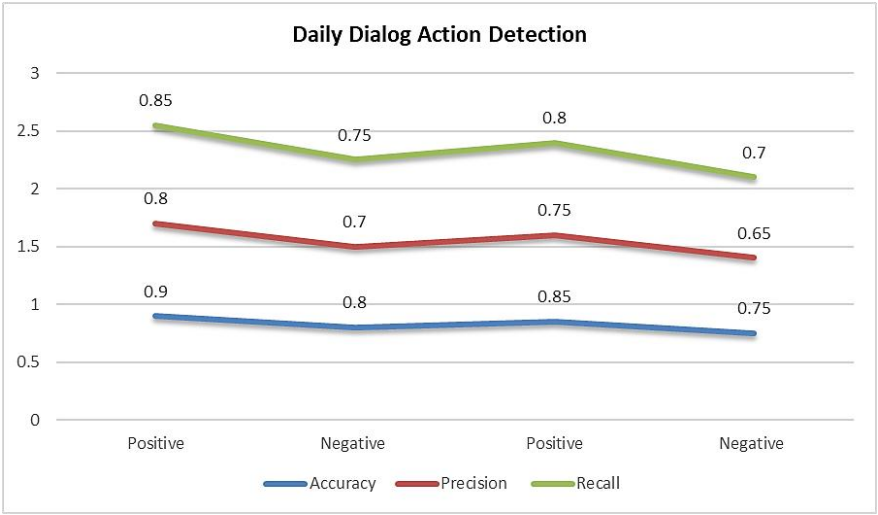

Fig. 5. Daily dialogue action detection

Performance measurement of unique word detection on ACE2020 dataset such as precision is about 85%, accuracy is 95% and recall 90% for positive polarity and for negative polarity accuracy is

85%, precision is 75% and recall is 80%. Performance measurement of unique word detection on sarcasm dataset such as precision is about 80%, accuracy is 90% and recall 85% for positive polarity and for negative polarity accuracy is 80%, precision is 70% and recall is 75% as shown in Figure 6.

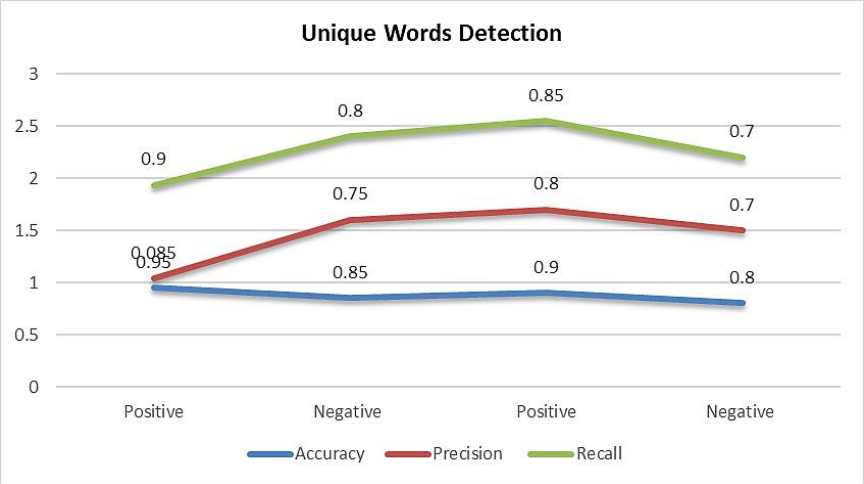

Fig. 6. Unique words detection

The comparison between the proposed technique and existing techniques [21,37-40] presents a better understanding of NLP on different datasets and in different domains. The results indicate that the proposed technique performs better than [21, 37-38, 40, 46]. The overall performance shows data having many polarities will be more productive of using techniques of sentiments.

5. Conclusions

Large textual data contain meaningful information which helps in decision making, predictions, reviews, and recommendation systems. Information retrievals have large number of categories such as sentiment analysis, emotions, action, and dialogs. In this study sentiment polarities such as positive and negative from the different classes such as daily emotional dialog, daily emotional action and unique word detection were detected. The NLP preprocessing techniques such as tokenization, stemming, and lemmatization were applied on large textual datasets i.e., ACE20202 and Sarcasm news for producing promising results. The promising results are produced for the detection of positive and negative polarities from the input datasets. In future, this work will be evaluated on other polarities and types of emotional and sentimental datasets.

References

[1] Kumar, Ganesh, Shuib Basri, Abdullahi Abubakar Imam, Sunder Ali Khowaja, Luiz Fernando Capretz, and Abdullateef Oluwagbemiga Balogun. "Data harmonization for heterogeneous datasets: A systematic literature review." *Applied Sciences* 11, no. 17 (2021): 8275. https://doi.org/10.3390/app11178275

[2] Cui, Jingfeng, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. "Survey on sentiment analysis: Evolution of research methods and topics." *Artificial Intelligence Review* 56, no. 8 (2023): 8469-8510. https://doi.org/10.1007/s10462-022-10386-z

[3] Kumar, Akshi, and Geetanjali Garg. "Sentiment analysis of multimodal twitter data." *Multimedia Tools and Applications* 78 (2019): 24103-24119. https://doi.org/10.1007/s11042-019-7390-1

[4] Gull, Karuna, Sudip Padhye, and Dr Subodh Jain. "A comparative analysis of lexical/NLP method with WEKA's bayes classifier." *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)* 5, no. 2 (2017): 221-227.

[5] Rathore, Bharati. "Digital transformation 4.0: integration of artificial intelligence & metaverse in marketing." *Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal* 12, no. 1 (2023): 42-48. https://doi.org/10.56614/eiprmj.v12i1y23.248

[6] Kumar, Ganesh, Shuib Basri, Abdullahi Abubakar Imam, and Abdullateef Oluwagbemiga Balogun. "Data harmonization for heterogeneous datasets in Big Data-A conceptual model." In *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4*, pp. 723-734. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-63322-6_61

[7] Avci, Cigdem, Bedir Tekinerdogan, and Ioannis N. Athanasiadis. "Software architectures for big data: A systematic literature review." *Big Data Analytics* 5, no. 1 (2020): 5. https://doi.org/10.1186/s41044-020-00045-1

[8] Basri, Shuib Bin, Ganesh Kumar, Fatin Fakhira Fahrurazi, Putri Emieldza Balqis Azmi, Abdullateef O. Balogun, and Hussaini Mamman. "Current trend of software requirement engineering process in IT small and medium enterprises (SMEs)-A systematic literature review." In *2023 13th International Conference on Information Technology in Asia (CITA)*, pp. 82-87. IEEE, 2023. https://doi.org/10.1109/CITA58204.2023.10262498

[9] Anagnostopoulos, Ioannis, Sherali Zeadally, and Ernesto Exposito. "Handling big data: Research challenges and future directions." *The Journal of Supercomputing* 72 (2016): 1494-1516. https://doi.org/10.1007/s11227-016-1677-z

[10] Adnan, Kiran, and Rehan Akbar. "Limitations of information extraction methods and techniques for heterogeneous unstructured big data." *International Journal of Engineering Business Management* 11 (2019): 1847979019890771. https://doi.org/10.1177/1847979019890771

[11] Saggi, Mandeep Kaur, and Sushma Jain. "A survey towards an integration of big data analytics to big insights for value-creation." *Information Processing & Management* 54, no. 5 (2018): 758-790. https://doi.org/10.1016/j.ipm.2018.01.010

[12] Jacobs, Adam. "The pathologies of big data." *Communications of the ACM* 52, no. 8 (2009): 36-44. http://doi.acm.org/10.1145/1536616.1536632

[13] Madden, Sam. "From databases to big data." *IEEE Internet Computing* 16, no. 3 (2012): 4-6. https://doi.org/10.1109/MIC.2012.50

[14] Wu, Xindong, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. "Data mining with big data." *IEEE Transactions on Knowledge and Data Engineering* 26, no. 1 (2013): 97-107. https://doi.org/10.1109/TKDE.2013.109

[15] Khan, Nawsher, Mohammed Alsaqer, Habib Shah, Gran Badsha, Aftab Ahmad Abbasi, and Soulmaz Salehian. "The 10 Vs, issues and challenges of big data." In *Proceedings of the 2018 international conference on big data and education*, pp. 52-56. 2018. https://doi.org/10.1145/3206157.3206166

[16] Arora, Yojna, and Dinesh Goyal. "Big data: A review of analytics methods & techniques." In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 225-230. IEEE, 2016. https://doi.org/10.1109/IC3I.2016.7917965

[17] Maheshwari, Himani, Luxmi Verma, and Umesh Chandra. "Overview of Big Data and its issues." *IJRECE* 7 (2019): 256.

[18] Sindhu, C., and Nagaratna P. Hegde. "Handling complex heterogeneous healthcare big data." *International Journal of Computational Intelligence Research* 13, no. 5 (2017): 1201-1227.

[19] Bhadani, Abhay Kumar, and Dhanya Jothimani. "Big data: Challenges, opportunities, and realities." *Effective big data management and opportunities for implementation* (2016): 1-24. https://doi.org/10.4018/978-1-5225-0182-4.ch001

[20] Gheisari, Mehdi, Guojun Wang, and Md Zakirul Alam Bhuiyan. "A survey on deep learning in big data." In *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)* 2 (2017): 173-180. https://doi.org/10.1109/CSE-EUC.2017.215

[21] Mehmood, Hassan, Ekaterina Gilman, Marta Cortes, Panos Kostakos, Andrew Byrne, Katerina Valta, Stavros Tekes, and Jukka Riekki. "Implementing big data lake for heterogeneous data sources." In *2019 ieee 35th international conference on data engineering workshops (icdew)*, pp. 37-44. IEEE, 2019. https://doi.org/10.1109/ICDEW.2019.00-37

[22] Sarkar, Kamal. "Sentiment polarity detection in Bengali tweets using LSTM recurrent neural networks." In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pp. 1-6. IEEE, 2019. https://doi.org/10.1109/ICACCP.2019.8883010

[23] Lee, In. "Big data: Dimensions, evolution, impacts, and challenges." *Business horizons* 60, no. 3 (2017): 293-303. https://doi.org/10.1016/j.bushor.2017.01.004

[24] Arici, Hasan Evrim, Nagihan Cakmakoglu Arıcı, and Levent Altinay. "The use of big data analytics to discover customers' perceptions of and satisfaction with green hotel service quality." *Current Issues in Tourism* 26, no. 2 (2023): 270-288. https://doi.org/10.1080/13683500.2022.2029832

[25] Al-Sai, Zaher Ali, and Rosni Abdullah. "Big data impacts and challenges: A review." In *2019 IEEE Jordan international Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pp. 150-155. IEEE, 2019. https://doi.org/10.1109/JEEIT.2019.8717484

[26] Xu, Jie, Feiran Huang, Xiaoming Zhang, Senzhang Wang, Chaozhuo Li, Zhoujun Li, and Yueying He. "Sentiment analysis of social images via hierarchical deep fusion of content and links." *Applied Soft Computing* 80 (2019): 387-399. https://doi.org/10.1016/j.asoc.2019.04.010

[27] Nie, Weizhi, Hai Ding, Dan Song, and Xingjian Long. "Location emotion recognition for travel recommendation based on social network." *Signal, Image and Video Processing* 13 (2019): 1259-1266. https://doi.org/10.1007/s11760-019-01457-w

[28] Singhal, Shivangi, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. "Spotfake: A multi-modal framework for fake news detection." In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp. 39-47. IEEE, 2019. https://doi.org/10.1109/BigMM.2019.00-44

[29] Majumder, Navonil, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. "Multimodal sentiment analysis using hierarchical fusion with context modeling." *Knowledge-Based Systems* 161 (2018): 124-133. https://doi.org/10.1016/j.knosys.2018.07.041

[30] Poria, Soujanya, Erik Cambria, Rajiv Bajpai, and Amir Hussain. "A review of affective computing: From unimodal analysis to multimodal fusion." *Information fusion* 37 (2017): 98-125. https://doi.org/10.1016/j.inffus.2017.02.003

[31] Shoumy, Nusrat J., Li-Minn Ang, Kah Phooi Seng, DM Motiur Rahaman, and Tanveer Zia. "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals." *Journal of Network and Computer Applications* 149 (2020): 102447. https://doi.org/10.1016/j.jnca.2019.102447

[32] Shirzad, Amirhossein, Hadi Zare, and Mehdi Teimouri. "Deep Learning approach for text, image, and GIF multimodal sentiment analysis." In *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 419-424. IEEE, 2020. https://doi.org/10.1109/ICCKE50421.2020.9303676

[33] Bougiatiotis, Konstantinos, and Theodoros Giannakopoulos. "Enhanced movie content similarity based on textual, auditory and visual information." *Expert Systems with Applications* 96 (2018): 86-102. https://doi.org/10.1016/j.eswa.2017.11.050

[34] Elallaoui, Meryem, Khalid Nafil, and Raja Touahni. "Automatic transformation of user stories into UML use case diagrams using NLP techniques." *Procedia Computer Science* 130 (2018): 42-49. https://doi.org/10.1016/j.procs.2018.04.010

[35] Arora, Monika, and Vineet Kansal. "Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis." *Social Network Analysis and Mining* 9, no. 1 (2019): 12. https://doi.org/10.1007/s13278-019-0557-y

[36] Luque, Carmen, José M. Luna, Maria Luque, and Sebastian Ventura. "An advanced review on text mining in medicine." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, no. 3 (2019): e1302. https://doi.org/10.1002/widm.1302

[37] Umidjon, Odiljonov. "Unlocking the power of natural language processing (NLP) for text analysis." *World scientific research journal* 17, no. 1 (2023): 66-73.

[38] Majeed, Adil, Hasan Mujtaba, and Mirza Omer Beg. "Emotion detection in roman urdu text using machine learning." In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pp. 125-130. 2020. https://doi.org/10.1145/3417113.3423375

[39] Ramesh, D., and S. K. Sanampudi. "Semantic and linguistic based short answer scoring system." *International Journal of Intelligent Systems and Applications in Engineering* 11, no. 3 (2023): 246-251.

[40] Parves, Abdul Bari, Abdullah Al Imran, and Md Riazur Rahman. "Incorporating supervised learning algorithms with NLP techniques to classify Bengali language forms." In *Proceedings of the International Conference on Computing Advancements*, pp. 1-7. 2020. https://doi.org/10.1145/3377049.3377110

[41] Prabhavathi, C., N. Vishali, P. S. Reddy, and J. V. Chandramouli. "Machine Learning Model for Classifying L _ Text Using Nlp (Amazon Product Reviews)." *International Research Journal of Computer Science* 6, no. 4 (2019): 161-178. https://doi.org/10.26562/IRJCS.2019.APCS10088

[42] Abualigah, Laith Mohammad, Ahamad Tajudin Khader, and Essam Said Hanandeh. "A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis." *Engineering Applications of Artificial Intelligence* 73 (2018): 111-125. https://doi.org/10.1016/j.engappai.2018.05.003

[43] Rout, Jitendra Kumar, Kim-Kwang Raymond Choo, Amiya Kumar Dash, Sambit Bakshi, Sanjay Kumar Jena, and Karen L. Williams. "A model for sentiment and emotion analysis of unstructured social media text." *Electronic Commerce Research* 18 (2018): 181-199. https://doi.org/10.1007/s10660-017-9257-8

[44] Zhou, Xinyu, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. "East: An efficient and accurate scene text detector." In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5551-5560. 2017. https://doi.org/10.1109/CVPR.2017.283

[45] Vanaja, Satuluri, and Meena Belwal. "Aspect-level sentiment analysis on e-commerce data." In *2018 International conference on inventive research in computing applications (ICIRCA)*, p. 1275-1279. IEEE, 2018. https://doi.org/10.1109/ICIRCA.2018.8597286

[46] Sun, Wencheng, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang. "Data processing and text mining technologies on electronic medical records: A review." *Journal of healthcare engineering* 2018, no. 1 (2018): 4302425. https://doi.org/10.1155/2018/4302425

[47] Kumar, Ganesh, Shuib Basri, Abdullahi Abubakar Imam, Abdullateef Oluwaqbemiga Balogun, Hussaini Mamman, and Luiz Fernando Capretz. "A novel multidimensional reference model for heterogeneous textual datasets using context, semantic and syntactic clues." *International Journal of Advanced Science and Applications* 14, no. 10 (2023): 754-763. https://doi.org/10.48550/arXiv.2311.06183

[48] Abdullah, Nashimah, Wan Nur Zafirah Wan Razak, Nur Aliya Ezzaty Azali, Khairun Nasrin Azman, Khairunnisa Mohd Kharfizi, Siti Nur Syakinah Jansi, Nurru Anida Ibrahim, Salisa Abdul Rahman, and Siti Norbakyah Jabar. "Electric vehicle adoption: A comparative analysis in Malaysia and ASEAN countries." *Semarak International Journal of Electronic System Engineering* 1, no. 1 (2024): 60-68.

[49] Din, Roshidi, Nuramalina Mohammad Na'in, Sunariya Utama, Muhaimen Hadi, and Alaa Jabbar Qasim Almaliki. "Innovative machine learning applications in non-revenue water management: Challenges and Future Solution." *Semarak International Journal of Machine Learning* 1, no. 1 (2024): 1-10. https://doi.org/10.37934/sijml.1.1.110

[50] Agoeng, Chandra, Nurul Dini Faqriah Miza Azmi, Hakimah Mat Harun, Nurzulaikha Abdullah, Wan Azani Mustafa, and Fakhitah Ridzuan. "Leveraging correlation and clustering: An exploration of data scientist salaries." *Journal of Advanced Research in Computing and Applications* 35, no. 1 (2024): 10-20. https://doi.org/10.37934/arca.35.1.1020

[51] Haron, Nor Hafiza, Nor Hafiza Abd Samad, Anis Juanita Mohd Zainudin, Ramlan Mahmod, and Fatimah Bibi Hamzah. "Automatic detection system of open access predatory journals: A unique application." *Journal of Advanced Research in Computing and Applications* 33, no. 1 (2023): 1-6. https://doi.org/10.37934/arca.33.1.16

[52] Salam, M.S.H., Jou, and Ahmad, A.F. "3D object manipulation using speech and hand gesture." *Journal of Advanced Research in Computing and Applications*, 31, no. 1 (2024):1–12. https://doi.org/10.37934/arca.31.1.112