A MACHINE LEARNING AND NLP BASED FRAMEWORK FOR EFFICIENT WEB MINING FOR SENTIMENT ANALYSIS

[1]Dr. Kaja Masthan, [2]Dr.SrikanthLakumarapu, [3]M.Varaprasad Rao, [4]Dr.Banoth Samya

[1]Assistant Professor, Department of Computer Science & Engineering Sphoorthy Engineering College,Nadargul
drkhaja.cse@gmail.com

[2,3,4]Associate Professor,Department of Computer Science & Engineering, CVR COLLEGE OF ENGINEERING
[2]dr.srikanthcse@gmail.com , [3]varam78@gmail.com, [4]samyabanoth@gmail.com

Abstract:-Web mining plays crucial role in modern technological world. It helps in automating different mining procedures in order to gain useful insights. Since web mining involves textual contents most of the time, NLP plays vital role in in understanding the data. It does mean that NLP and Machine Learning (ML) go hand in hand in processing such data. Many existing methods that are used for web mining through ML and NLP suffer from mediocre performance. To overcome this problem, in this paper, we proposed a framework known as Web Mining for Sentiment Analysis Framework (WMSAF).It has strong pre-processing methodology that exploits multiple procedures including NLP and ML based approach for machine learning towards useful analysis of data. We considered e-commerce case study dealing with automatic data collection from Amazon product reviews web URLs and perform pre-processing, NLP and machine learning towards automatic sentiment analysis. Our methodology is based on the BERT model which is found efficient when compared with other models. We proposed an algorithm named Web Mining and Sentiment Analysis of Amazon Product Reviews (WMSA-APR) to realize the proposed framework. From the results it is observed that the proposed BERT based method showed highest accuracy with 96% when compared with existing models such as UniLM, Reformer and XLNet.

Keywords –Web Mining, E-Commerce, NLP, Sentiment Analysis, Machine Learning

DOI: 10.48047/ecb/2023.12.si12.160

1. INTRODUCTION

Learning based approaches using AI became popular in the real world to solve many problems. Web applications are largely driving many businesses. In such scenario, there have been attempts to explore analysis of web application usage, its performance and different other performance statistics. Machine learning has become an important research area applicable to different domains [1]. ML and web mining are widely used to solve many real world problems. Many social media related data analytics and BI derivation were made using web mining and ML techniques [2], [3], [17]. Web mining is used for discovering trends in web usage and user behaviour besides data analytics [12].

Sentiment analysis is one of the research areas linked to web mining. It is meant for automatic discovery of data from given web applications and discover user or customer sentiments towards ascertaining facts. In [2] and [6] Bayesian network and intelligent learning approaches respectively are used to analyse sentiments. E-commerce is one of the important domains where sentiments play vital role in understanding customer feedback and make corrective steps. Amazon is one of the largest e-Commerce portal where it is possible to view product reviews and analyse sentiments. In this paper, we proposed a methodology through web mining, ML, NLP and sentiment analysis to acquire Amazon product reviews automatically and perform sentiment analysis. From the literature review, it is observed that many existing methods that are used for web mining through ML and NLP suffer from mediocre performance. Our contributions are as follows.

1. We proposed a framework known as Web Mining for Sentiment Analysis Framework (WMSAF).
2. We proposed an algorithm named Web Mining and Sentiment Analysis of Amazon Product Reviews (WMSA-APR) to realize the proposed framework.
3. We built an application to evaluate the framework which is based on BERT and compared with other state of the art models such as UniLM, Reformer and XLNet.

The remainder of the paper is organized as Section 2 to review literature, Section 3 to present proposed framework, Section 4 to analyse results and Section 5 to conclude our work.

2. RELATED WORK

This section reviews literature on existing methods for web mining for sentiment analysis. Learning based framework is proposed in [1] for decision making process. In [2] mining process is carried out based on Twitter data by exploiting a Bayesian-based analysis framework. A deep learning approach for analysing social media data is explored in [3] towards disaster assessment. An ensemble approach is exploited in [4] for automatic disease diagnosis and providing classification results in healthcare domain. An e-learning system is automatically analysed in [5] using ML approaches towards generating recommendations. In [6] an opinion mining model is built to achieve intelligent learning towards helping governments in making decisions. A smart city use case is investigated in [7] for discovering energy efficient mechanisms through machinelearning. Analysis of large volumes of data in IoT healthcare use case is made in [8] for healthcare data analytics. LSTM and enhanced NLP models are used for sarcasm identification in [9] towards realizing a framework.

In [10] data analytics framework is proposed to find safety dynamics of protection equipment. In [11] ML models are exploited towards automatic analysis of X-ray docs and diagnose Covid-19 cases. In [12] web mining based approach is analysed towards mapping an innovation ecosystem. In [13] a manufacturing related operations are analysed using ML techniques. Other contributions found in the literature include object detection [14], price prediction [15], blockchain based ML [16], blockchain adoption prediction [17], human safety prediction [18], power usage pattern discovery [19] and spatio-temporal mining [20]. From the literature review, it is observed that many existing methods that are used for web mining through ML and NLP suffer from mediocre performance.

3. PROPOSED FRAMEWORK

We proposed a framework known as Web Mining for Sentiment Analysis Framework (WMSAF). It has strong pre-processing methodology that exploits multiple procedures including NLP and ML based approach for machine learning towards useful analysis of data. We considered e-commerce case study dealing with automatic data collection from Amazon product reviews web URLs and perform pre-processing, NLP and machine learning towards automatic sentiment analysis. Our methodology is based on the BERT model which is found efficient when compared with other models.

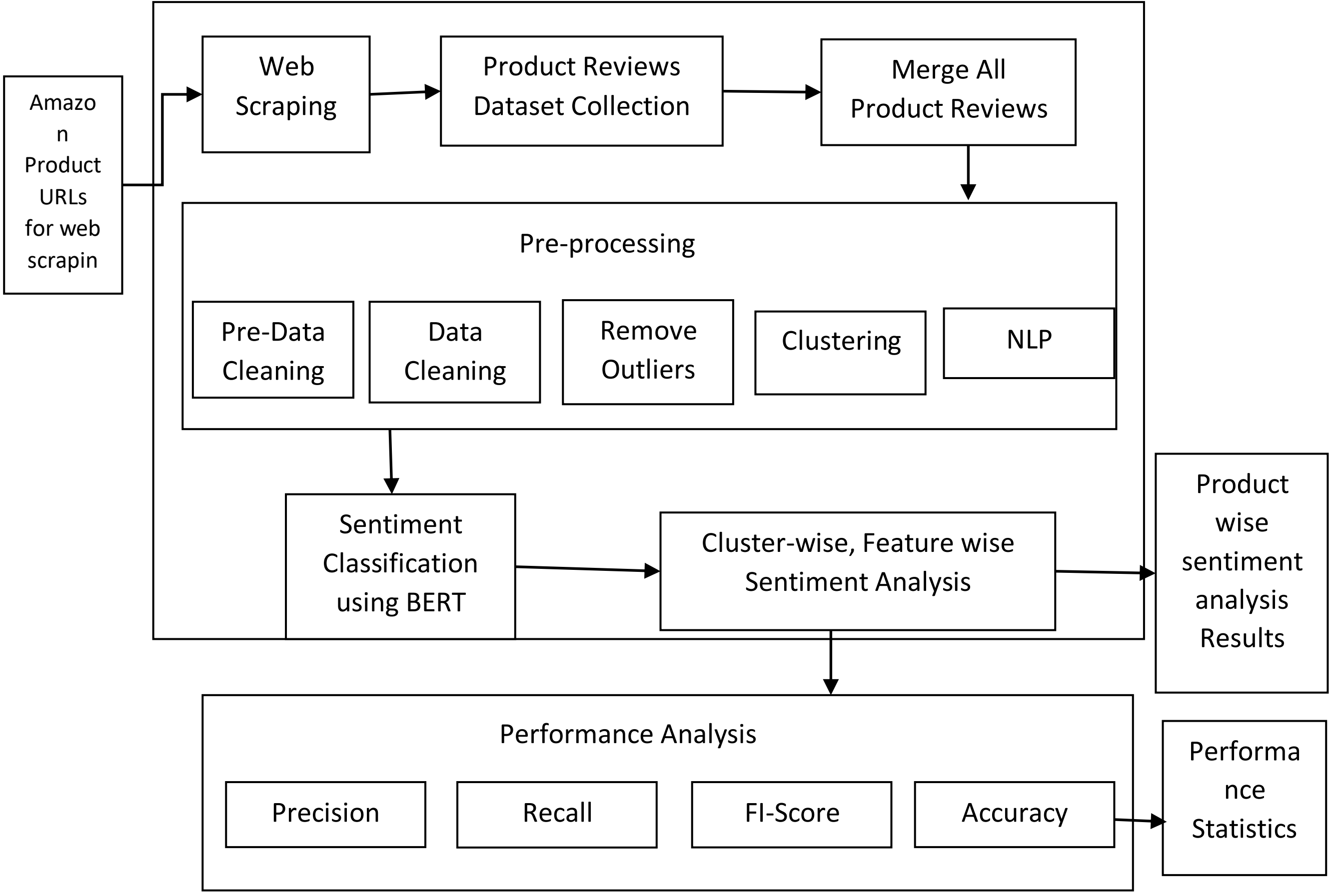

Figure 1: Web Mining for Sentiment Analysis Framework (WMSAF)

The given input URLs are given to web scrapper tool. The tool collects a set of many product reviews. Then all the product reviews are merged to form a single dataset containing review of e-Commerceproducts associated with Amazon. Afterwards, the product reviews are subjected to pre-processing that includes data cleaning, removal of outliers, clustering for each of further processing and NLP techniques. Then BERT model is used to perform sentiment analysis. BERT model is from Google AI. It is a pre-trained model for processing text based contents. It has achieved great success in ML applications. It has underlying NLP capabilities that makes it superior to many such models. It has many mathematical operations in its underlying functionality. An attention mask matrix is computed as in Eq. 1.

$$M_{(i,j)} = \begin{cases} 0 \text{ if } \langle n_i, n_j \rangle \notin E \; and \; \langle n_j, n_i \rangle \notin E \; and \; i \neq j. \\ 1 \text{ otherwise.} \end{cases} \qquad (1)$$

BERT has an important aspect pertaining to language modelling. It is used to predict original tokens in spite of masking. This functionality is expressed as in Eq. 2.

$$Loss_{MLM} = \sum_{x_i \in T_{mask} \cup C_{mask}} -log\, p(x_i) \qquad (2)$$

Where the prediction probability is denoted as p(x_i). BERT exploits latent semantic relationships between context and mathematical formulae. In order to predict the association of input context with given text, it is expressed as in Eq. 3 and Eq. 4.

$$Loss_{CCP} = -\delta \log p - (1-\delta)\; log(1-p) \qquad (3)$$

$$\delta = \begin{cases} 1 \; if \; C = C'. \\ 0 \text{ otherwise.} \end{cases} \qquad (4)$$

Where the probability of C=C' is denoted by p. BERT has support for prediction of masked substructure. Structural composition of an operation is to predict different probabilities. It is expressed as in Eq. 5 and Eq. 6.

$$Loss_{MSP} = \sum_{n_i \in N_{mask}} \sum_{n_i \in N} \big(-\delta \log p(n_i, n_j) - (1-\delta)\log(1 - p(n_i, n_j)))\big) \qquad (5)$$

$$\delta = \begin{cases} 1 \; if \; e_{i,j} \in E \; or \; e_{j,i} \in E. \\ 0 \qquad \text{otherwise.} \end{cases} \qquad (6)$$

The above equations are utilized in order to compute total loss function which plays important role while predicting sentiments. The derived loss function is as expressed in Eq. 7.

$$Loss_{total} = Loss_{MLM} + Loss_{CCP} + Loss_{MSP} \qquad (7)$$

Since BERT is pre-trained model, it is incrementally used to train with additional data available. Then it is used to predict sentiments from the data collected from e-commerce application. Amazon product reviews are analysed and sentiments are classified.

Algorithm: Web Mining and Sentiment Analysis of Amazon Product Reviews (WMSA-APR)

Inputs: Amazon product reviews URLs D

Output: Product reviews P, sentiment analysis results R, performance statistics S

```
Begin
rawDataCollection←WebScriping(D)
For each collection c from rawDataCollection
reviewsData←CollectReviews(c)
finalProductReviews←finalProuctReviews+reviewsData
```

```
End For
D←finalProductReviews
Display  D
D'←PreProcess(D)
R←SentimentAnalysis(D',Pre-Trained BERT Model)
S←PerformanceEvaluation(R)
Display R
Display S
End
```

Algorithm 1: Web Mining and Sentiment Analysis of Amazon Product Reviews (WMSA-APR)

As presented in Algorithm 1, it takes Amazon product reviews URLs D as input and produces output in the form of product reviews P, sentiment analysis results R, performance statistics S. Web scraping tool is used to collect product reviews linked to Amazon e-Commerce portal. The reviews collection is merged to form a single dataset of product reviews. Then the final product reviews are subjected to pre-processing. Afterwards, the pre-processed data is given to BERT based model in order to analyse sentiments. The resultant sentiments are used to verify with ground truth to arrive at performance statistics.

4. EXPERIMENTAL RESULTS

This section presents experimental3 results of our research that includes web mining using Amazon e-commerce product reviews URLs using web scraping, data collection, NLP and sentiment analysis.

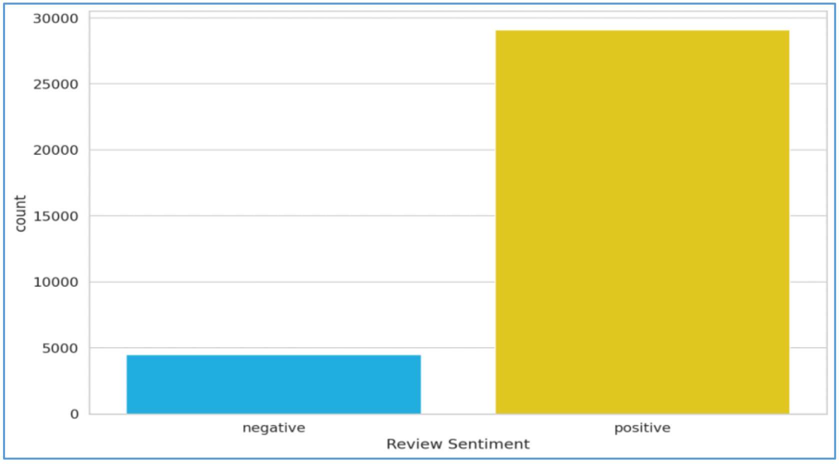

Figure 2: Discovered sentiments for Amazon product reviews collected through web mining

As presented in Figure 2, after pre-processing and NLP, the resultant product reviews are subjected to sentiment analysis. The positive and negative sentiment distribution found in the collected data are provided. More than 4000 negative sentiments are found while positive sentiments are greater than 28000.

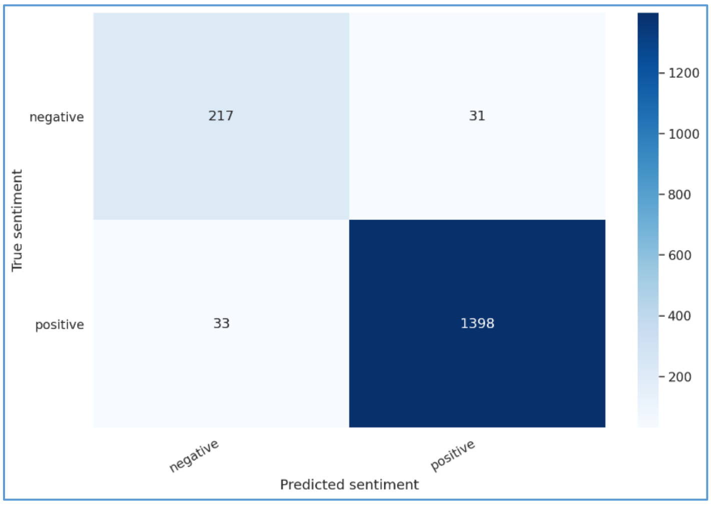

Figure 3: Confusion matrix pertaining sentiment classification

As presented in Figure 3, the confusion matrix shows true sentiments and predicted sentiments in order to support performance evaluation of the proposed BERT based method for sentiment analysis.

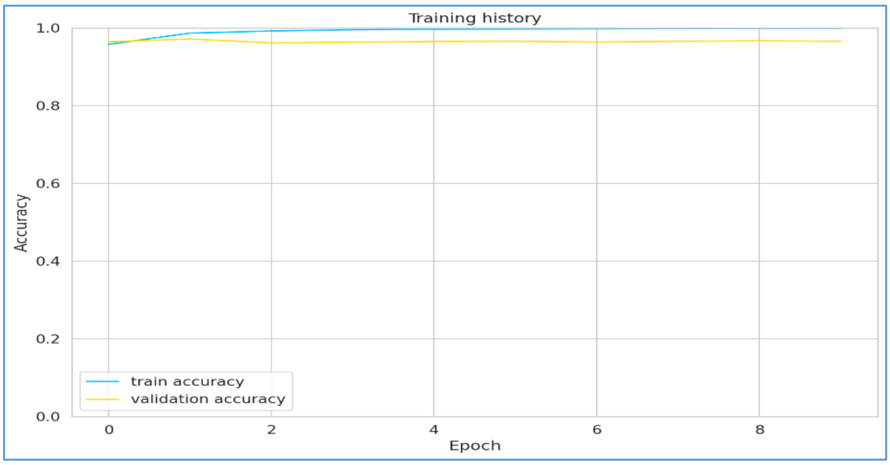

Figure 4: Shows accuracy details of the proposed BERT based approach to sentiment analysis

As presented in Figure 4, performance of BERT based method for sentiment analysis is provided against number of epochs.

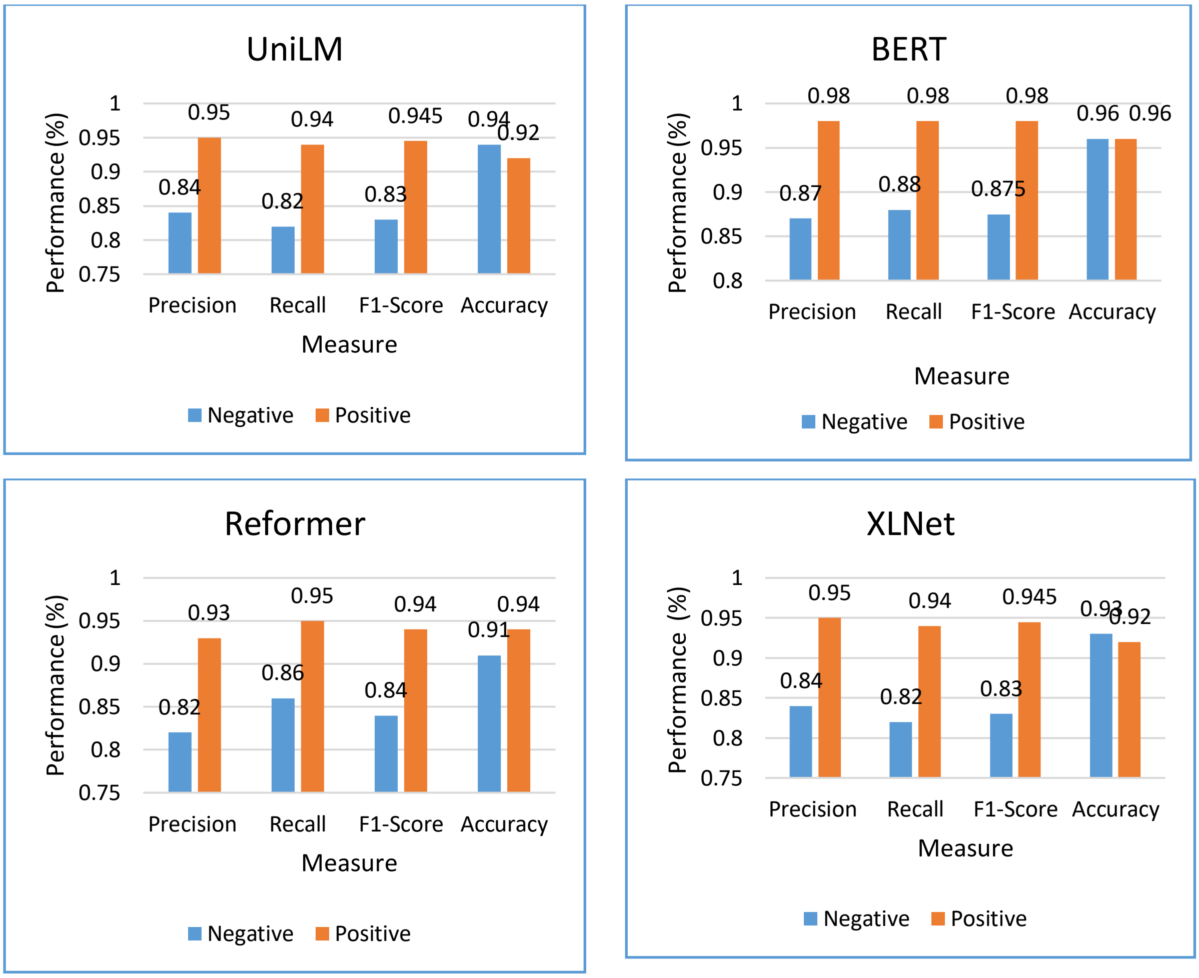

Figure 5: Performance of different models used for sentiment analysis

As presented in Figure 5, the proposed BERT based method is compared against many existing methods such as UniLM, Reformer and XLNet. UniLM achieved 84% and 95% precision, 82% and 94% recall, 83% and 94.50% F1-score and 94% and 92% accuracy for identification of negative and positive sentiments respectively. Reformer achieved 82% and 93% precision, 86% and 95% recall, 84% and 94% F1-score and 91% and 94% accuracy for identification of negative and positive sentiments respectively. XLNet achieved 84% and 95% precision, 82% and 94% recall, 83% and 94.50% F1-score and 93% and 92% accuracy for identification of negative and positive sentiments respectively. The proposed BERT based model achieved 87% and 98% precision, 88% and 98% recall, 87.50% and 98% F1-score and 96% and 96% accuracy for identification of negative and positive sentiments respectively.

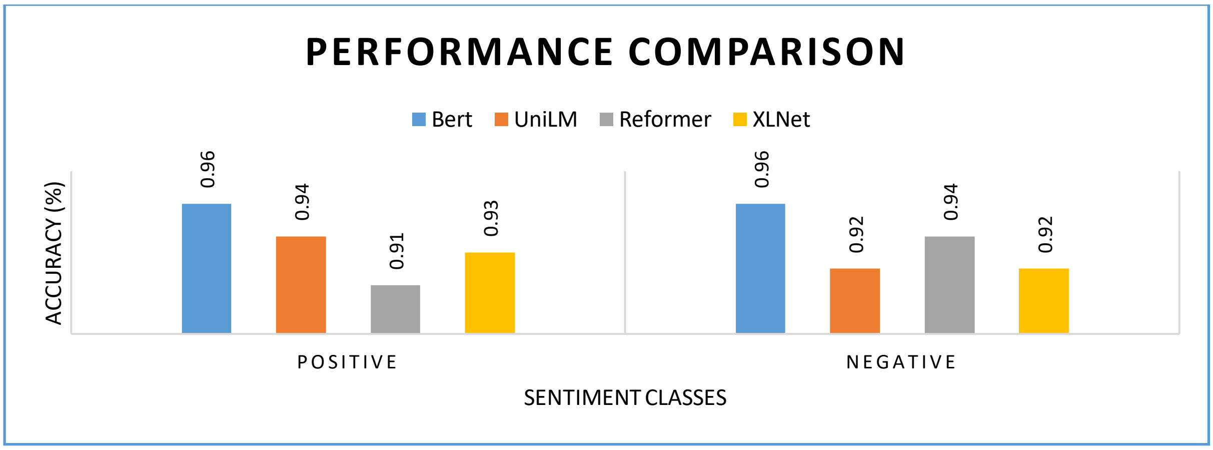

Figure 6: Accuracy comparison among different models used for sentiment analysis

As presented in Figure 6, accuracy of all models is compared for two sentiment classes. For identification of positive sentiments, Reformer achieved least accuracy with 91%. The accuracy of XLNet is 93%, UniLM is 94% while BERT is 96%. Similarly, for identification of negative sentiments, UniLM and XLNet achieved least accuracy with 92%. The accuracy of Reformer is 94% while BERT is 96%. From the results it is observed that the proposed BERT based method showed highest accuracy with 96%.

5. CONCLUSION AND FUTURE WORK

We proposed a framework known as Web Mining for Sentiment Analysis Framework (WMSAF). It has strong pre-processing methodology that exploits multiple procedures including NLP and ML based approach for machine learning towards useful analysis of data. We considered e-commerce case study dealing with automatic data collection from Amazon product reviews web URLs and perform pre-processing, NLP and machine learning towards automatic sentiment analysis. Our methodology is based on the BERT model which is found efficient when compared with other models. We proposed an algorithm named Web Mining and Sentiment Analysis of Amazon Product Reviews (WMSA-APR) to realize the proposed framework. From the results it is observed that the proposed BERT based method showed highest accuracy with 96% when compared with existing models such as UniLM, Reformer and XLNet. In future, we intend to improve our framework with deep learning techniques to leverage prediction performance. Another direction is to use our framework with data of different domains.

References

[1] GABRIELA CZIBULA, GEORGE CIUBOTARIU, MARIANA-IOANA MAIER, AND HANNELORE LISEI. (2022). IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining. *IEEE*. 10, pp.80651-80666.

[2] Stefanos Zervoudakis;EmmanouilMarakakis;HaridimosKondylakis and Stefanos Goumas; (2021). OpinionMine: A Bayesian-based framework for opinion mining using Twitter Data . Machine Learning with Applications. http://doi:10.1016/j.mlwa.2020.100018.

[3] Bhoi, Ashutosh; Pujari, Sthita Pragyan; Balabantaray and Rakesh Chandra (2020). A deep learning-based social media text analysis framework for disaster resource management. Social Network Analysis and Mining, 10(1), 78–. http://doi:10.1007/s13278-020-00692-1.

[4] Tuli, Shreshth; Basumatary, Nipam; Gill, Sukhpal Singh; Kahani, Mohsen; Arya, Rajesh Chand; Wander, Gurpreet Singh ansBuyya, Rajkumar (2020). HealthFog: An ensemble deep learning based Smart Healthcare System for Automatic Diagnosis of Heart Diseases in integrated IoT and fog computing environments. Future Generation Computer Systems, 104, 187–200. http://doi:10.1016/j.future.2019.10.043.

[5] Khanal, Shristi Shakya; Prasad, P.W.C.; Alsadoon, Abeer and Maag, Angelika (2019). A systematic review: machine learning based recommendation systems for e-learning. Education and Information Technologies. http://doi:10.1007/s10639-019-10063-9.

[6] Sharma, Abhilasha and Shekhar, Himanshu (2020). Intelligent Learning based Opinion Mining Model for Governmental Decision Making. Procedia Computer Science, 173, 216–224. http://doi:10.1016/j.procs.2020.06.026.

[7] Zekić-Sušac, M., Mitrović, S., & Has, A. (2020). *Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. International Journal of Information Management, 102074.* http://doi:10.1016/j.ijinfomgt.2020.1020.

[8] (2021). A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System . Mobile Networks and Applications. http://doi:10.1007/s11036-020-01700-6.

[9] Aytug Onan and Mansur Alp Tocoglu; (2021). A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification . IEEE Access. http://doi:10.1109/access.2021.3049734.

[10] Nath, Nipun D.; Behzadan, Amir H. and Paal, Stephanie G. (2020). Deep learning for site safety: Real-time detection of personal protective equipment. Automation in Construction, 112, 103085–. http://doi:10.1016/j.autcon.2020.103085.

[11] Sara Hosseinzadeh Kassania;Peyman Hosseinzadeh Kassanib;Michal J. Wesolowskic;Kevin A. Schneidera and Ralph Detersa; (2021). Automatic Detection of Coronavirus Disease

(COVID-19) in X-ray and CT Images: A Machine Learning Based Approach . Biocybernetics and Biomedical Engineering. http://doi:10.1016/j.bbe.2021.05.013.

[12] Kinne, Jan and Axenbeck, Janna (2020). Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics. http://doi:10.1007/s11192-020-03726-9.

[13] Dogan, Alican and Birant, Derya (2020). Machine Learning and Data Mining in Manufacturing. Expert Systems with Applications, 114060–. http://doi:10.1016/j.eswa.2020.114060.

[14] Mittal, Payal; Sharma, Akashdeep and Singh, Raman (2020). Deep learning-based object detection in low-altitude UAV datasets: A survey. Image and Vision Computing, 104046–. http://doi:10.1016/j.imavis.2020.104046.

[15] Patel, Mohil Maheshkumar; Tanwar, Sudeep; Gupta, Rajesh and Kumar, Neeraj (2020). A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions. Journal of Information Security and Applications, 55, 102583–. http://doi:10.1016/j.jisa.2020.102583.

[16] Prabhat Kumar;RandhirKumar;GautamSrivastava;Govind P. Gupta;RakeshTripathi;Thippa Reddy Gadekallu and Neal N. Xiong; (2021). PPSF: A Privacy-Preserving and Secure Framework Using Blockchain-Based Machine-Learning for IoT-Driven Smart Cities . IEEE Transactions on Network Science and Engineering. http://doi:10.1109/TNSE.2021.3089435.

[17] Kamble, Sachin S.; Gunasekaran, Angappa; Kumar, Vikas; Belhadi, Amine and Foropon, Cyril (2020). A machine learning based approach for predicting blockchain adoption in supply Chain. Technological Forecasting and Social Change, 120465–. http://doi:10.1016/j.techfore.2020.120465.

[18] Bhuiyan, Md. Rafiuzzaman; Khushbu, Sharun Akter and Islam, Md. Sanzidul (2020). 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) - A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3, 1–5. http://doi:10.1109/ICCCNT49239.2020.9225384.

[19] Liu, Xue; Ding, Yong; Tang, Hao and Xiao, Feng (2020). A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. Energy and Buildings, 110601–. http://doi:10.1016/j.enbuild.2020.110601.

[20] Wang, Senzhang; Cao, Jiannong and Yu, Philip (2020). Deep Learning for Spatio-Temporal Data Mining: A Survey. IEEE Transactions on Knowledge and Data Engineering, 1–1. http://doi:10.1109/TKDE.2020.3025580.