Sentiment Analysis Performance and Reliability Evaluation Using an XLNet-based Deep Learning Approach

Dhaval Bhoi, Dr. Amit Thakkar

•

CSPIT, Charotar University of Science and Technology, Changa-388421, Gujarat, India
dhavalbhoi.ce@charusat.ac.in

Abstract

Online reviews are now a global form of communication between consumers and E-commerce companies. When it comes to making day-to-day decisions, customers rely heavily on the availability of internet reviews, as well as their trustworthiness and performance. Due to the unique qualities of user reviews, customers are finding it increasingly difficult to define and examining the authenticity and reliability of sentiment evaluations. These sentiment classifications for user reviews can aid in understanding user feelings, review dependability, and customer perceptions of movie items. Deep Learning is a strong technique for learning several layers of data representations or features. When compared to traditional machine learning approaches, deep learning techniques yield better results. To assess, analyze, and weight the usefulness of each review comment, we employed the XLNet Deep Learning Model Approach on balanced movie review dataset.Experimental result demonstrates that the proposed deep learning model achieves higher performance evaluation than those of other classifiers.

Keywords: Sentiment Analysis, Machine Learning, Deep Learning, XLNet

1. Introduction

In recent years, the E-commerce industry has grown at a breakneck pace [1]. When a wide variety of items or products appear in customers' online shopping sites, however, determining their authenticity and trustworthiness becomes more complex, making it impossible to identify genuine goods from imitation or replica goods. Customers typically evaluate similar items based on quality information and pricing before making purchasing decisions.

According to studies, consumers who acquire information from online available posts or ratings are more interested in buying the product than those who just gather information from the manufacturer or producer. It implies that online remarks or sentiments left by previous consumers play a significant effect in the selection of online goods. According to a recent study, the number of online reviews are proportional to users' buy intent. Customers are more eager to buy if there are more online reviews [2]. Sentiment Analysis is a linguistic technique that involves extracting emotions from raw texts [3] [4]. It can be performed at Document Level, Sentence Level or Aspect Level shown in below Figure 1.

This is commonly used on social media posts and customer reviews to automatically determine whether some users are happy or unpleasant, as well as reasons. The original place of sentiment analysis or opinion mining is shown in Figure 2. The major purpose of this research is to demonstrate how deep learning may be used to perform sentiment analysis. It's a means of evaluating a document's, sentence's, or word's polarity. It is utilized in a wide range of industries, including marketing, medical diagnosis, education sector, film industries, and others, to aid businesses and customers. Based on the type of input, it can be broadly classified as Document Level sentiment Analysis, Sentence (or Line) Level Sentiment Analysis and Aspect (or Feature)

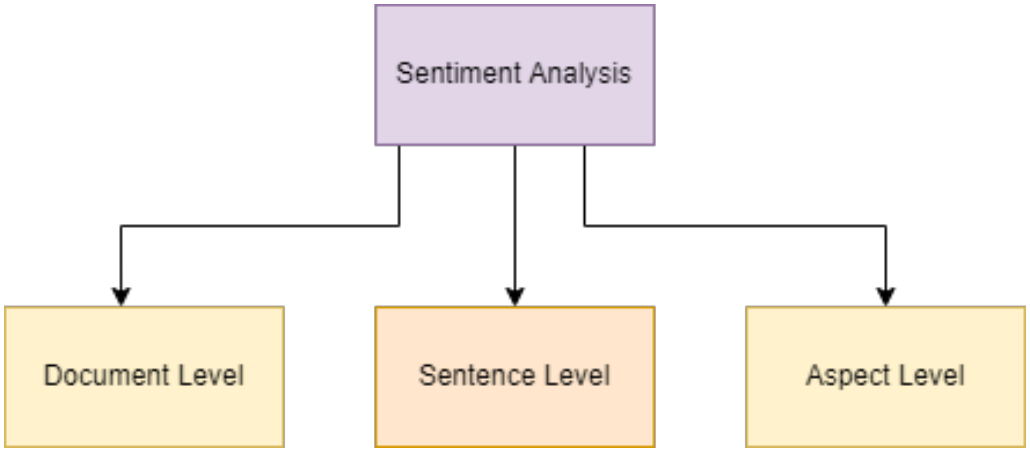

Figure 1: *Levels of Sentiment Analysis*

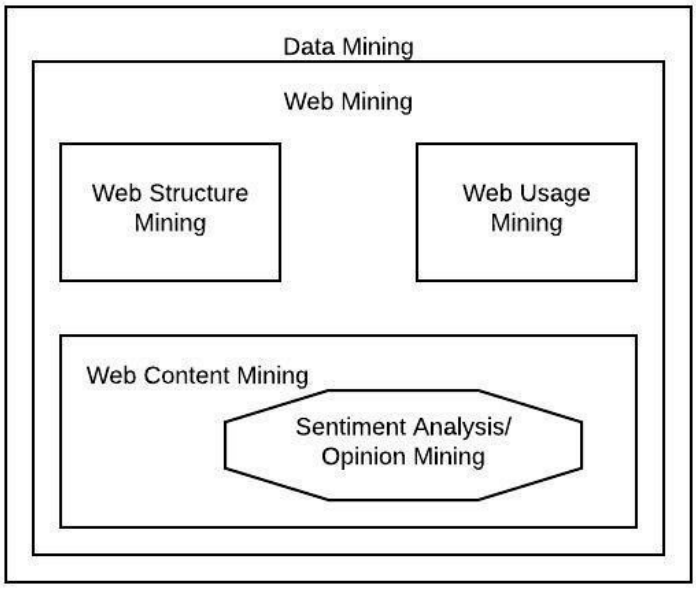

Figure 2: *Place of Sentiment Analysis in Data Mining*

level Sentiment Analysis [5]. The document's polarity is calculated by counting the number of times a positive or negative appears in a document. If there are more positive terms in a paper than negative words, it is determined that the document is positive. A sentence level check and a word level check can be performed in the same way as a document check [6]. The alternative method is termed Aspect Based, Entity Based, or Feature Based Sentiment Analysis, and it focuses on numerous characteristics of the situation [7]. Comment-based opinion mining or sentiment analysis is crucial for evaluating and examining the reliability of e-commerce items or products when human variables are involved. The e-commerce dependability evaluation results can be used in the iteration of product reliability design process, and offer product life cycle references management.

2. Related Work

Different researchers had used earlier lexicon based approach, machine learning approach and deep learning based approach to perform classification of sentiment [8] [9].

The authors of [10] have used BOW [Bag of Words] feature extraction technique and applied Navie Bayes and Support Vector Machine classifier produce improved classification result.

In this study, a stacked residual LSTM [Long Short Term Memory] model was utilised to estimate sentiment intensity, which improved prediction accuracy [11].

In recent work, authors of [12] used an Amazon review dataset to test the baseline deep learning models for LSTM, GRU, Bi-LSTM, and Bi-GRU. These tactics ignore the importance of word order and the many distinct meanings that words can convey.

Recently, the authors used a supervised machine learning-based technique to analyse sentiment in product reviews. They improved the result by combining two separate word embedding techniques, word2vec and FastText Word Embedding, with a CNN Model. They improved their performance by using FastText as a word embedding technique and CNN as a deep learning model [13].

3. Proposed System

To conduct this experiment, we employed a movie review dataset and associated binary sentiment polarity labels [14]. The primary rationale for selecting this dataset is that it is well-balanced. It comprises an equal number of positive and negative review data samples. We used an XLNet-based deep learning approach in the proposed method. The XLNetTokenizer is used to extract features from review text, and the XLNet deep learning model is used to train it. Figure 3 depicts the proposed approach's overall flow. The current reliability indices [15] mainly have time measurement and probability measurement for evaluation. Reliability is denoted as R. Function of Failure (F), Probability Density Function (f), and Failure Rate(λ) are the most commonly used probability measurements, whereas Mean Time To Failure (MTTF) and Mean Time Between Failures (MTBF) are the most common time measurements.

The possibility that a product or item will execute particular function for a certain amount of time duration under specified conditions without failure is known as reliability. To put it another way, if T is a product's time to failure, and the reliability function at time t is as shown in 1

$$R(t) = P(T > t) \tag{1}$$

The average time that it takes for an unfixable product or item to perform properly under given specific conditions until it fails is called MTTF. When the number of samples given is N and the life of sample I is t_i, the MTTF can be calculated as in 2

$$MTTF = \frac{1}{N}\sum_{1}^{N} t_i \tag{2}$$

If a product is repairable; the average continuous time between product or item failures during the operation or testing is called Mean Time Between Failure (MTBF), and it is calculated as shown in 3

$$MTBF = \int_0^\infty t f(t) dt \tag{3}$$

We have introduced the notion of Reliability for Sentiment (R_s), which is considered a weighted average of Sentimental Analysis Value (S) of products or items obtained from consumers' opinions, sentiments or evaluations about the product they've bought or used, weighted by the Importance or Usefulness (U) of a specific given review or comment. To quantify something, we use the following function shown in equation 4.

$$R_s = \frac{\sum S_i * Ui}{\sum U_i} \tag{4}$$

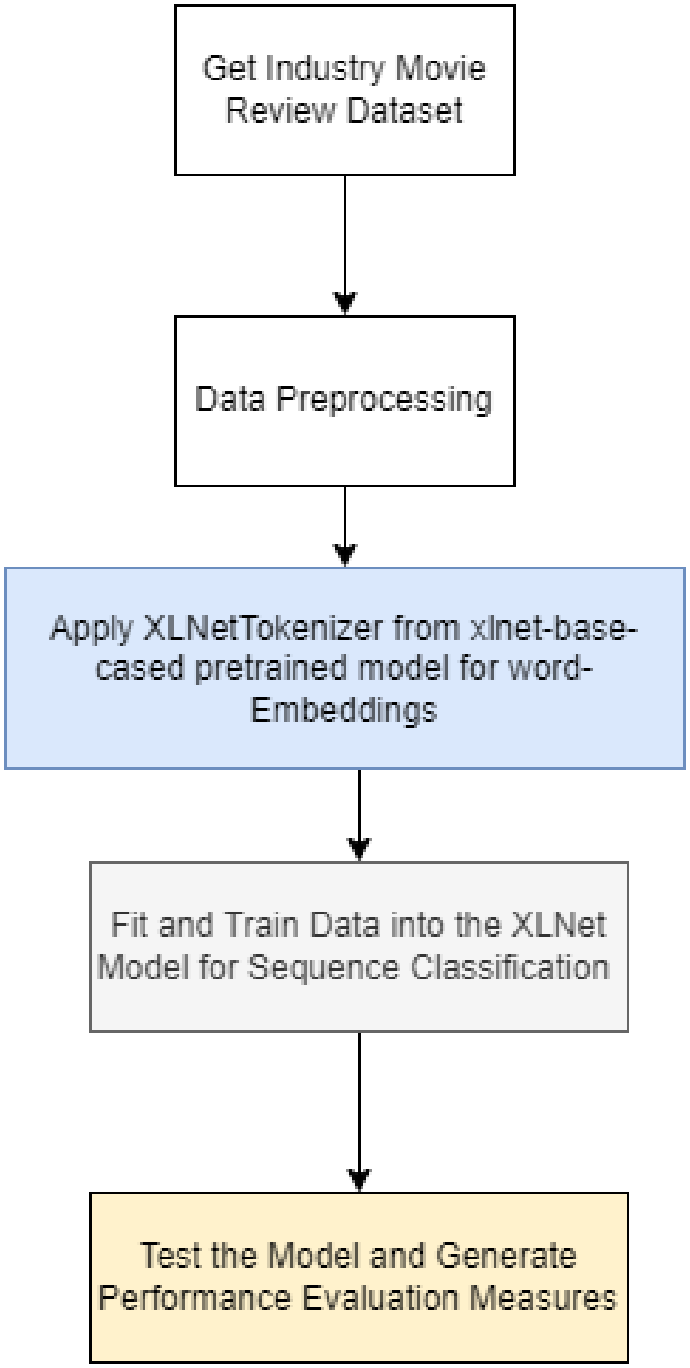

Figure 3: *Proposed Approach*

Sentiment analysis based on this metric of reliability can benefit not just consumers, but also businesses and organisations looking to enhance their operations and strategy.

We determined Accuracy, Precision, Recall, and F1-score [16] for performance evaluation, as given in the equations below 5, 6, 7 and 8.Precision and recall are balanced by the F-measure or F1-Score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{8}$$

Where TP, TN, FP and FN are number of true positive, true negative, false positive and false negative samples in a review dataset.

4. Results and Discussion

Based on the findings shown in table 1, we can conclude that our proposed deep learning model surpasses all existing machine learning methods, such as Logistic Regression, Naive Bayes classifier, and Support Vector Machine due to proper feature representation using XLNetTokenizer followed by XLNet Deep Learning Model. XLNet is the most recent and most advanced model to come from the burgeoning field of Natural Language Processing (NLP). XLNet is an auto-regressive language model that uses a transformer architecture with recurrence to output the joint probability of a sequence of tokens hence it takes more time for training.

Table 1: *Comparative Performance Result Analysis*

Type of Model	Model Applied	Accuracy	Precision	Recall	F1-Score
Machine Learning Models	**Naive Bayes**	81.02	78.61	73.98	75.66
	Random Forest	84.76	83.39	77.84	80.52
	SVM	86.13	82.87	82.99	82.88
Deep Learning Model	**XLNet**	96.00	95.50	96.00	96.00

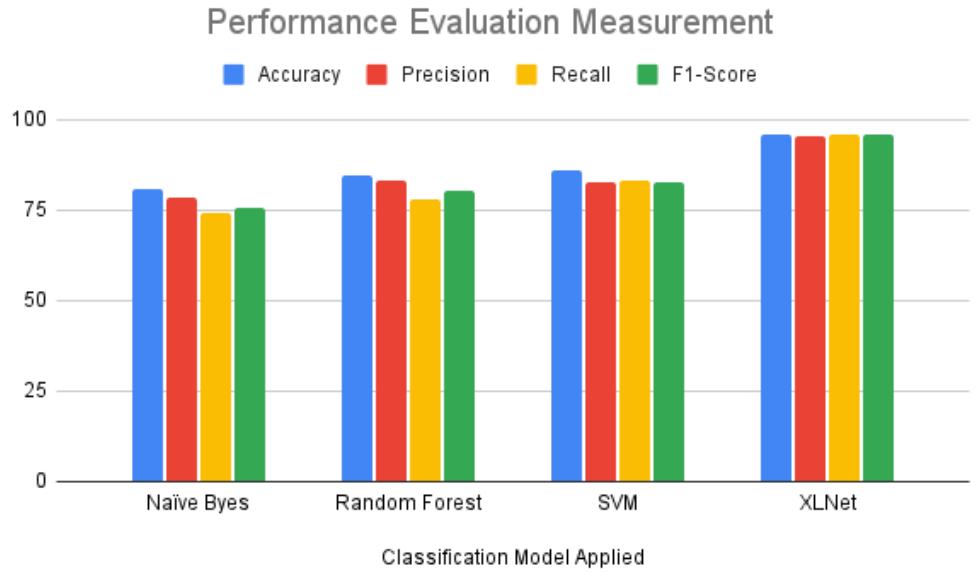

Figure 4: *Performance Analysis Evaluation*

In terms of F1-score, the suggested XLNet deep learning model performs 13.12 percent better than the top performing SVM machine learning model, as shown in Figure 4. We have enhanced accuracy, precision, and recall by 9.87 percent, 12.53 percent, and 13.02 percent, respectively.

5. Conclusion and Future Work

All other machine learning models, including LR, NB, and SVM, are outperformed by our proposed deep learning methodology, XLNet. However, this method has the drawback of requiring more training time. We have applied our proposed approach on a dataset from the movie business; however, we can apply this model to other industry domains to determine how effective it is. As the proposed approach yields better results, it clearly tackles the reliability and performance issues based on sentiment analysis.

Acknowledgement

The authors would like to thank the Principal and Dean of the Faculty of Technology and Engineering, as well as the Head of the U P U. Patel Department of Computer Engineering at CSPIT, Charotar University of Science and Technology, Changa, for their continuous suggestions, encouragement, guidance, and support in completing this research work. We want to express our gratitude to Management in particular for their moral guidance and encouragement.

Declaration of Conflicting Interests

The Author(s) declare(s) that there is no conflict of interest.

References

[1] Lakshmi P, StalinDavid D, Kalaria I, Jayadatta S, Sharma A, Saravanan D. Research on Collaborative Innovation of E-Commerce Business Model for Commercial Transactions. Turkish Journal of Physiotherapy and Rehabilitation.;32(3):787-94.
[2] Zhang X, Xie G, Li D, Kang R. Reliability Evaluation Based on Sentiment Analysis of Online Comment. In2018 12th International Conference on Reliability, Maintainability, and Safety (ICRMS) 2018 Oct 17 (pp. 88-91). IEEE.
[3] Devika, M.D. C, Sunitha Ganesh, Amal. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science. 87. 44-49. 10.1016/j.procs.2016.05.124.
[4] Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in twitter to improve information filtering. InProceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval 2010 Jul 19 (pp. 841-842).
[5] Farooq U, Mansoor H, Nongaillard A, Ouzrout Y, Qadir MA. Negation Handling in Sentiment Analysis at Sentence Level. J. Comput.. 2017 Sep 1;12(5):470-8.
[6] Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications. 2017 Apr 15;72:221-30.
[7] Vanaja S, Belwal M. Aspect-level sentiment analysis on e-commerce data. In2018 International Conference on Inventive Research in Computing Applications (ICIRCA) 2018 Jul 11 (pp. 1275-1279). IEEE.
[8] Taj S, Shaikh BB, Meghji AF. Sentiment analysis of news articles: A lexicon based approach. In2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) 2019 Jan 30 (pp. 1-5). IEEE.
[9] Long C, Ziyu G, Jinhong H, Jinye P. A survey on sentiment classification. Journal of Computer Research and Development. 2017 Jun 1;54(6):1150.
[10] Zou H, Tang X, Xie B, Liu B. Sentiment classification using machine learning techniques with syntax features. In2015 International Conference on Computational Science and Computational Intelligence (CSCI) 2015 Dec 7 (pp. 175-179). IEEE.
[11] Wang J, Peng B, Zhang X. Using a stacked residual LSTM model for sentiment intensity prediction. Neurocomputing. 2018 Dec 17;322:93-101.
[12] Sachin S, Tripathi A, Mahajan N, Aggarwal S, Nagrath P. Sentiment analysis using gated recurrent neural networks. SN Computer Science. 2020 Mar;1(2):1-3.

[13] Shah A. Sentiment Analysis Of Product Reviews Using Supervised Learning. Reliability: Theory Applications. 2021(SI 1 (60)).
[14] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)
[15] E. A. Elsayed, Reliability Engineering, 2nd ed., New Jersey: Wiley Sons, Inc., pp.3“5, 15“69, 2012.
[16] Phan HT, Tran VC, Nguyen NT, Hwang D. Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. IEEE Access. 2020 Jan 3;8:14630-41.