

Opinion Mining-Based Fake Review Detection Using Deep Learning Technique



Koustav Pal, Sayan Poddar, S. L. Jayalakshmi , Madhumita Choudhury, S. K. Saif Ahmed, and Soumyajit Halder

Abstract Recently, the field of text mining gained more attention due to its enormous opportunities and challenging problems in the exponential growth of unstructured textual data. People frequently express their thoughts in complicated ways, making automated labelling of textual data challenging. The usage of mislabelled data sets reduced the efficiency of automated labelling tasks. In this paper, we proposed an opinion mining-based fake review detection using deep learning technique to express word semantic sentiment. SentiWordNet and WordNet lemmas are used to retrieve word synsets and sentiment scores. The Amazon shoes review data set is used for implementation. The proposed model achieved 92% accuracy for the Amazon shoes review data set.

Keywords Opinion mining · Sentiment analysis · Attention mechanism · Amazon review analysis · Deep learning · Lemmatization · Long short-term memory (LSTM)

1 Introduction

Recent social media growth has made it possible for users to publish opinions about things, people, events, and subjects in a range of formal and informal ways. Reviews, forums, social media posts, blogs, and discussion boards are a few examples of these types of environments. The computational analytics involved with such text is referred to as the problem of opinion mining and sentiment analysis [1]. Enhancing customer happiness, boosting conversions, and improving revenue are the main goals of collecting customer reviews. Regularly updated customer reviews attract more people to the website. Online customer reviews are a crucial element of the purchasing process for e-commerce sites. A product's page will have a greater chance of getting

K. Pal · S. Poddar · S. L. Jayalakshmi (✉) · M. Choudhury · S. K. Saif Ahmed · S. Halder
Department of Computer Science, Pondicherry University, Puducherry, India
e-mail: sathishjayalakshmi02@pondiuni.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
A. Kumar et al. (eds.), *Proceedings of the 4th International Conference on Data Science, Machine Learning and Applications*, Lecture Notes in Electrical Engineering 1038,
https://doi.org/10.1007/978-981-99-2058-7_2

more traffic and a higher click-through rate on search engine result pages if it has an acceptable rating [2].

Many positive reviews on the platform increase the chances of encouraging clients to spend extra on services/products. Favourable reviews will increase client trust in the brand. Negative reviews may damage the company's reputation, credibility, and reliability. Numerous buyers are hesitant to buy from firms with no or many unfavourable evaluations [3]. Customers are 86% less likely to buy from companies with poor ratings. A bad review may turn off 22% of buyers, and the unfavourable reviews can turn off 59%. As a result, any poor e-commerce product reviews might harm the brand and cause a drop in sales [4]. It is time-consuming and inaccurate to classify positive and negative reviews manually; as a result, the development of automatic methods to identify the sentiment of thoughts has become essential. The majority of the opinion mining-based approach for false review detection has used conventional machine learning methods using the term frequency-inverse document frequency (TF-IDF) values, and subsequently, the discrimination of linguistic features is also estimated to enhance the accuracy of the system [5, 6]. Recently the attention-based deep model with convolutional neural network (CNN) is used for sentimental analysis [7].

This work uses both sentiment analysis and opinion mining terms interchangeably. This has been accomplished using an attention-based technique. In this work, we describe a deep learning model that automatically learns representations and features for the classification of reviews using an attention-based mining approach. We develop a pre-processing step that conducts fundamental feature extraction to decrease the number of model inputs. When compared with the conventional methods, the proposed model produced better results. This work is organized as follows: Sect. 2 reviews the related work on the classification of customer reviews. The proposed model is then described in Sect. 3. Experimental result and discussions are presented in Sect. 4. Finally, Sect. 5 presents conclusion.

2 Literature Review

Sedighi et al. [8] proposed a deep neural networks approach combined with an attention mechanism to perform classification by constructing a model to learn and capture the discriminating characteristics automatically. Their model's evaluation reveals that it performs noticeably better than conventional models and does not require any hand feature engineering. Kudakwashe et al. [9] proposed a concept called sentiment polarity, and it extracts the information about hotel services from reviews to automatically construct a sentiment data set for testing and training. A comparison analysis with Naive Bayes multinomial, minimal sequential optimisation, complement Naive Bayes, and Composite hypercubes on iterated random projections was performed to determine an effective machine learning method for the framework's classification component.

Using CNN and two different bidirectional RNN networks, the authors of [10] created an attention-based sentiment analysis. We first employ a pre-processor to enhance the data quality by removing noisy data and correcting mistakes. In order to reduce feature dimensionality and retrieve contextual information, our model employs CNN with max-pooling. Thirdly, to capture long-term dependencies, two separate bidirectional RNNs, namely long short-term memory and gated recurrent unit, are used. Sagnika et al. [11] proposed a model that cleverly combines long short-term memory (LSTM) with a CNN model. Modern deep learning models like CNN and LSTM can effectively analyse textual material and spot innate relationships and patterns with different degrees of abstraction. The suggested approach creates an ensemble model that incorporates the advantages of the two models. The addition of an attention network improves the model's effectiveness.

3 Opinion Mining-Based Fake Review Detection Using Deep Learning Technique

In this paper, we introduced an opinion mining-based fake review detection using deep learning technique that concatenates group of lexicons, a model for Word2vec, and an attention mechanism (AM) to express sentiment observation. AM is used to create sentiment data from SentiWordNet and the Liu's lexicon. However, the context of the text is first determined using the Word2vec model before analysis. These two tasks work together to handle the linguistic problems using the deep learning model. Before generating the text's sentiment features from the context, it attempts to comprehend the words in terms of their sentiment conflict. This process simulates how a person could approach this problem. As shown in Fig. 1, the system is divided into three primary components, and the following sections provide descriptions of each component.

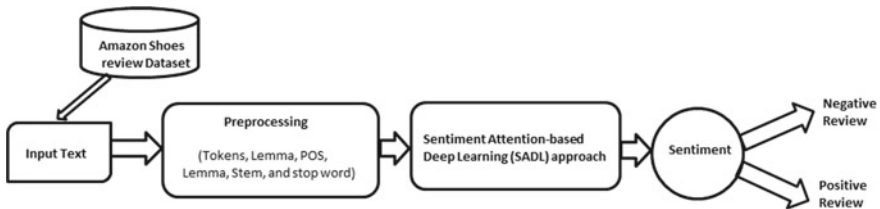


Fig. 1 Architecture of the proposed system

3.1 Text Pre-processing

In sentiment analysis, text pre-processing is one of the important step for analysing the information from social networks, this is an important task in sentiment analysis. Our pre-processing method is composed of four steps: (i) Text extraction; (ii) Text cleaning; (iii) Lemmatizing; and (iv) Negation marking. Language grammar and structure are used in this NLP-based method. By word lemmatizing with WordNet, downstream analysis tasks that involve matching terms from the text with those from lexicons are made simple.

Text extraction

In the text extraction, few sentences in the reviews are first divided into clauses based on the punctuation that is available at the clause level.

Text cleaning

This will change the capital letters to lowercase and eliminate any special characters.

Lemmatization

By using this technique, the text is changed into its stemmed form. Recent methods in the literature frequently apply an intelligent approach to collapse various word templates by seeking to remove affixes. Suffixes ending with -es (or) -s can make a noun singular or plural. For a verb to be in its present or past participle form, apply the suffixes -ing or -ed. The -est suffix allows adjectives to take comparative or superlative forms. SentiWordNet searches for synsets of a particular word are, however, impractical because of the erroneous stemmed forms produced by these algorithms. In this work, we adopted the methodology from [12], in which SentiWordNet and WordNet lemma were used to retrieve word synsets and sentiment scores.

3.2 Embedding of Attention Word

One of the most important developments in deep learning research is attention mechanism (AM). According to word sentiment scores, the proposed technique uses lexicons and AM to teach students how to concentrate on the key words in a phrase. The proposed method considers the word representations in the forms of either semantics or sentiment by using of both lexicons and Word2vec models. SentiWordNet and the Liu's lexicon are used to extract word sentiment information. To retrieve the sentiment information from the lexicons, the sentiment score calculation method is used.

3.3 Components of Opinion Mining-Based Fake Review Detection Using Deep Learning Technique Approach

Dropout

It was suggested to use dropout as a regularization strategy for neural network (NN) classifiers. When training, dropout method picks neurons at random that are not used. The activation of downstream neurons is not influenced by these neurons during the forward phase, and no updation of weight is made to the neurons during the back propagation step. The most common method of resolving overfitting is dropout. The proposed approach is also useful for handling low priority levels.

Bi-directional LSTM

The LSTM, a particular gated RNN, was introduced to handle the issue of handling long-term reliance. Fundamental building blocks of an LSTM network architecture are input gates, output gates, forget gates, and memory cells. The memory cell travels immediately through the entire chain to store the information for either long or short periods of time. What data should be removed from the cell is determined by the forget gate. Which fresh information will be added to the cell's memory is decided by the input gate. The output gate controls how much data the LSTM produced. The proposed system includes a BiLSTM to combine more data in the form of sentiment and semantics.

Dense layer

In the proposed system, the output is categorized into positive and negative opinions using a fully connected dense layer with a SoftMax function.

4 Experimental Studies

4.1 Data set Used

An Amazon shoe review [13] data set served as the data source for our project, and it is shown in Fig. 2. It reflects representative collections of consumer assessments, opinions, and variations for how a product is viewed in various geographic regions and the presence of promotional bias or intent in reviews. It contains the following columns: marketplace, which is the Web site's two-letter ISO 3166 country code. More than 100,000 consumer reviews are available as tab-separated values (TSV) files in the amazonreviews-PDS S3 bucket in the AWS US East Region. A specific review is indicated in the data file using with tab-delimited, with no quotes, and with escape characters.

A seemingly random number is allocated to a single author is called a customer ID. Each review is identified by using an unique review ID. The product ID identifies

Unnamed: 0	customerid	reviewid	productid	productparent	producttitle	rating	verifiedpurchase	reviewtitle	Review body	
804898	2361047	25929976	R3TFM0HVOOS20D	B000L0O09G	281605411	Merrell Men's Moab Ventilator Hiking Shoe	5	Y	Great Hiking Shoe	I bought these shoes for a trailing hiking vac...
765165	1825603	45076746	R213KXZ4P5TE3	B003UHU5T6	866537507	New Balance Men's MIR860 Running Shoe	1	Y	Returned Shoe	New Balance shoe was advertised as 13B. Shoe ...
743357	1170569	6639856	R2XBVYDVERME00D	B00FDXRMRI	245930226	Jessica Simpson Women's Wintee Dress Pump	4	Y	Four Stars	Great shoe 2014-12-28
105749	898438	18849452	R1JBHTVGQ9YOG	B008G3BC7I	398112824	RYKA Women's Influence Cross Training Shoe	5	Y	Five Stars	Perfect fit! Getting much use in Zumba class!...
610266	1481949	43263299	R1C50G9Q895L6W	B000G615UC	234909654	Bates Men's Ultra-lites 8 Inches Tactical Sport...	5	Y	Five Stars	Great warm comfortable boots

Fig. 2 Description of Amazon shoe review data set

the review given for the particular product. Similarly, reviews for the same product written in various languages can be grouped using the same product ID in multilingual data set. The name of the item is the product title. Product category is a big product category used to group reviews. Rating is having the values from 1 (lowest) to 5 (highest). The number of constructive votes indicates the total number of votes the review has gotten. This review was inspired by the Vine programme. Next, verified purchase tells that a review is given after a real purchase. Review heading contains the title of the review, review body shows the review's text, and the review date records the date the review was written.

4.2 Performance Analysis

Figure 3 shows the various performance measures of the proposed system. The '0' represents the negative reviews, and '1' illustrates the positive reviews. It can be observed from Fig. 4 that positive reviews are clearly identified from negative (fake) reviews with the help of the proposed deep learning-based approach. The proposed model achieved the $F1$ -score as 95% and accuracy as 92%. The graph showed increased accuracy for the epoch when the model was run on training and validation data sets. The accuracy on the Y -axis is given on a scale of 0–1. Figure 5 shows the accuracy with which the model classified the reviews positively and negatively on the training and validation data set. The accuracy also increased with the increasing number of the epoch. It can be observed from Fig. 5 that the presence of some anomalies in the validation data set, increased the value of loss function. In future work, the data set can be further refined, and some useless attributes from the data set can be eliminated.

	precision	recall	f1-score	support
0	0.63	0.43	0.51	4096
1	0.94	0.97	0.95	36534
accuracy			0.92	40630
macro avg	0.79	0.70	0.73	40630
weighted avg	0.91	0.92	0.91	40630

Fig. 3 Performance measure

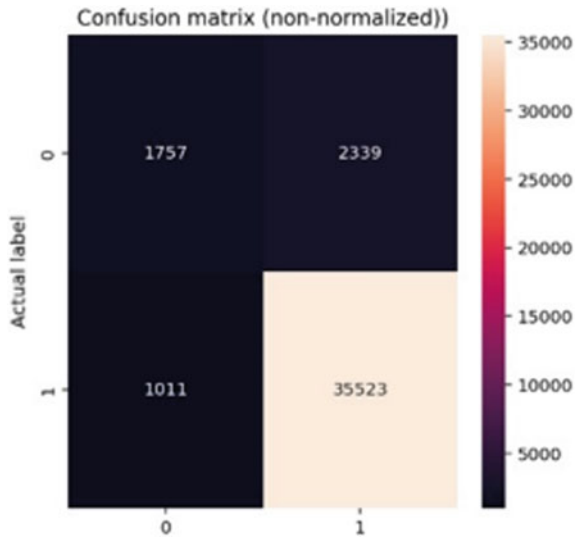


Fig. 4 Confusion matrix

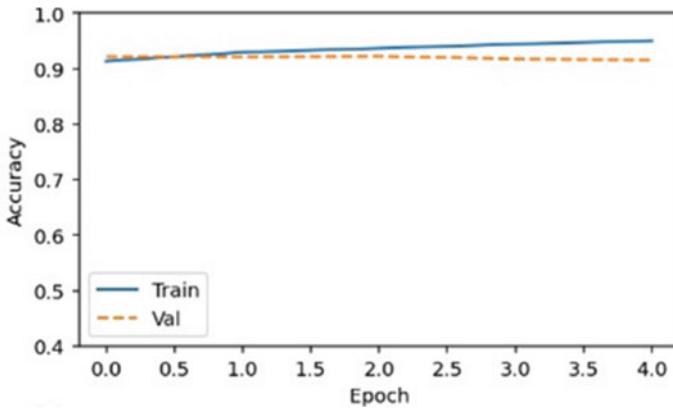


Fig. 5 Accuracy Vs Epoch

5 Conclusion

Recently, opinion mining-based approach gained more attention in the literature. In this paper, we proposed an opinion mining-based fake review detection using deep learning technique approach for detecting the fake customer reviews in an online Web site. For the same Amazon shoe review data set was used. The proposed approach used SentiWordNet and WordNet lemma to retrieve word synsets and sentiment scores. The implementation result shows that the proposed model achieved 92% accuracy. In future, the presence of anomalies in the validation data set can be further refined to improve the accuracy.

References

1. Aggarwal CC (2022) Opinion mining and sentiment analysis. Machine learning for text. Springer, Cham, pp 491–514
2. Anas SM, Kumari S (2021) Opinion mining based fake product review monitoring and removal system. In 2021 6th international conference on inventive computation technologies (ICICT-IEEE), pp 985–988
3. Sun S, Luo C, Chen J (2017) A review of natural language processing techniques for opinion mining systems. *Inf Fusion* 36:10–25
4. <https://www.powerreviews.com/research/how-fake-reviews-destroy-consumer-trust/>
5. Wankhade M, Rao ACS, Kulkarni C (2022) A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* 55:5731–5780
6. Arif SA, Hossain TB (2021) Opinion mining of customer reviews using supervised learning algorithms. In: 2021 5th international conference on electrical information and communication technology (EICT), pp 1–6
7. Kamyab M, Liu G, Rasool A, Adjeisah M (2022) ACR-SA: attention-based deep model through two-channel CNN and Bi-RNN for sentiment analysis. *Peer J Comput Sci*
8. Sedighi Z, Ebrahimipour-Komleh H, Bagheri A, Kosseim L (2019) Opinion spam detection with attention-based neural networks. In: The thirty-second international flairs conference, 2019
9. Zvarevashe K, Olugbara O (2018) A framework for sentiment analysis with opinion mining of hotel reviews. In: 2018 conference on information communications technology and society (ICTAS)
10. Kamyab M, Liu G, Adjeisah M (2021) Attention-based CNN and Bi-LSTM model based on TF-IDF and glove word embedding for sentiment analysis. *Appl Sci* 11(23):11255
11. Sagnika S, Mishra BSP, Meher SK (2021) An attention-based CNN-LSTM model for subjectivity detection in opinion-mining. *Neural Comput Appl* 33(24):17425–17438
12. Husnain M, Missen MMS, Akhtar N, Coustaty M, Mumtaz S, Prasath VB (2021) A systematic study on the role of SentiWordNet in opinion mining. *Front Comput Sci* 15(4):1–19
13. <https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazonreviewsusShoesv100.tsv.gz>