A Comparative Study of Machine Learning Models for Sentiment Analysis: Customer Reviews of E-commerce Platforms

Davoodi, Laleh; Mezei, Jozsef

Published in:
35th Bled eConference: Digital Restructuring and Human (Re)Action

DOI:
10.18690/um.fov.4.2022

Published: 23/06/2022

Document Version
Accepted author manuscript

Document License
CC BY

A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR SENTIMENT ANALYSIS: CUSTOMER REVIEWS OF E-COMMERCE PLATFORMS

LALEH DAVOODI[1], JÓZSEF MEZEI[2]

[1] Laleh Davoodi, Åbo Akademi University, Faculty of Social Sciences, Business and Economics, Turku, Finland; e-mail: laleh.davoodi@abo.fi
[2] József Mezei, Åbo Akademi University, Faculty of Social Sciences, Business and Economics, Turku, Finland; e-mail: jozsef.mezei@abo.fi

Abstract Understanding customers' preferences can be vital for companies to improve customer satisfaction. Reviews of products and services written by customers and published on various online platforms offer tremendous potential to gain important insights about customers' opinions. Sentiment classification with various machine learning models has been of great interest to academia and practice for a while, however, the emergence of language transformer models brings forth new avenues of research. In this article, we compare the performance of traditional machine learning models and recently introduced transformer-based techniques on a dataset of customer reviews published on the Trustpilot platform. We found that transformer-based models outperform traditional models, and one can achieve over 98% accuracy. The best performing model shows the same excellent performance independently of the store considered. We also illustrate why it can be sometimes more reliable to use the sentiment polarity assigned by the machine learning model, rather than a numeric rating that is provided by the customer.

Keywords:
customer reviews, sentiment analysis, RoBERTa, machine learning

1 Introduction

With the rapid advancement of Internet technology in recent years, online shopping has become a popular means for people to buy and consume goods. In particular, e-commerce sectors, which have thrived amid the COVID-19 issue, have experienced extraordinary and unexpected development. These trends increase the already high importance for organizations to understand and optimize the way customers interact with online e-commerce platforms in order to increase customer satisfaction. An increasingly utilized source of information that organizations can make use of is online reviews. While customer reviews can provide valuable information related to products and services, the sheer quantity of these reviews makes it infeasible and impractical for manual inspection. Companies are required to increasingly make use of advanced natural language processing (NLP) tools and machine learning techniques to understand their customers better and stay competitive in the market (Jagdale et al. 2019).

The most valuable piece of information in reviews, additionally to a numeric rating, is contained in the free-form comments on the product or service. Consumer evaluations typically include useful information regarding product quality as well as helpful recommendations. However, it is not a straightforward task to extract the relevant information. Various tools of NLP are now increasingly used in understanding customer satisfaction, such as sentiment analysis (Sun et al., 2019), topic modeling (Piris & Gay, 2021), text summarization (Tsai et al., 2020), and automated translation (Gangual & Mamidi, 2018).

By understanding the emotional polarity of messages with sentiment analysis, companies can gain a detailed understanding of customers, and identify what products and services are perceived negatively or positively by customers and why. The company needs to be able to constructively evaluate good and negative feedback and make better judgments based on the needs of customers. In this article, our main focus is to *compare the classification performance of traditional machine learning techniques and a more recent invention, the RoBERTA model* in sentiment evaluation of online reviews (Liu et al., 2019). By identifying the best performing model, we can aid decision makers in understanding customer's prerences better, and in turn improve the offered services. To address this research problem, we collected 3500 user reviews from the online platform Trustpilot, which hosts reviews to help consumers in online shopping. The messages were randomly selected and manually annotated by the authors of this article using the polarities 'positive', 'negative', and 'mixed'. After data preprocessing, several machine

learning models were tested and the best performing models were identified. The models utilized include Support Vector Machines, Naïve Bayes, BERT, and RoBERTa.

The rest of the article is structured as follows. In Section 2, a brief literature review is presented to discuss different approaches to sentiment analysis, and specifically how they have been applied to analyze customer reviews. The research methodology, data collection, and processing are presented in Section 3. Section 4 contains the main results: we interpret the findings and compare them to previous literature. Finally, we present some concluding remarks, limitations, and future research directions in Section 5.

2 Sentiment analysis of customer reviews

The goal of sentiment analysis is to determine how sentiments are represented in texts and whether the expressions suggest positive or negative attitudes about the subject (Nasukawa & Jeonghee, 2003). The most fundamental application of sentiment analysis is to gather people's opinions, in particular in the form of customer reviews. Many business decisions are influenced by such viewpoints (Rajput, 2020). Formally, as presented by Zeng et al. 2019, sentiment analysis can be defined as "the process of automatically analyzing the subjective commentary text with the customer's emotional color and deriving the customer's emotional tendency". In the following, we summarize the different approaches focusing on the different families of techniques used in sentiment analysis.

Based on the technique used to perform sentiment analysis, one can typically make a distinction between (i) lexicon-based and (ii) machine learning approaches. In lexicon-based approaches, an existing dictionary is utilized, which contains a sentiment polarity/intensity for words and expressions; then the sentiment is determined by the majority of polarities in the text. Additionally, to general dictionaries such as the Harvard GI, there exist some domain-specific dictionaries available, e.g. the Financial Polarity Lexicon in finance (Malo et al., 2014). While we can identify approaches to construct a domain-specific dictionary for user reviews, such as the one presented by Han et al. (2018), in this domain, generic dictionaries are mainly used to evaluate sentiments, with the most frequently used one being the SentiWordNet (Baccianella et al., 2010). Han et al. (2018a) present a good example of utilizing lexicon-based sentiment analysis using the text of reviews from the online commerce site Amazon. By accounting for the possibility of bias, the authors complement the use of lexicons with a

weighted processing strategy and found the results to be superior. In the model, optimized weights combining positive and negative score of a review are included as "many methods present more positive values than negative values, especially the lexicon-based method".

Lexicon-based approaches can offer a good baseline solution, as in most cases typically 70% sentiment classification accuracy is achievable. In recent times, a different stream of sentiment analysis techniques has become dominant: machine-learning based approaches. Developments from the last decade are numerous in this domain, with the first important being the new approaches to improve on traditional text representation methodologies (such as term frequency-inverse document frequency) with word, and later sentence encoders, including most importantly Word2Vec and Doc2Vec. These word/sentence encoders assign comparable vectors to words/sentences that appear in the same context because they are semantically similar. As an example, Shuai et al. (2018) applied Doc2vec and various machine learning algorithms such as Support Vector Machine, Logistic Regression, and Naïve Bayes on 11600 hotel reviews. They found that the best performance (80% precision and 88% recall) can be achieved by combining Doc2Vec with Support Vector Machines.

In order to further improve word and sentence embeddings, language transformer models were introduced (Vaswani et al., 2017) with the idea of creating different embeddings for a word instead of a fixed one, to incorporate information on different contexts. Making use of encoder and decoder, an extensive, unsupervised pre-training and fine-tuning on labeled data, transformers can be used to solve a variety of text classification problems, including sentiment classification. One of the first transformer models, and arguably the most influential one, is BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2019). The BERT approach has two stages: pre-training and fine-tuning. The pre-training stage includes two unsupervised tasks: a masked language model and sentence prediction. Through supervised learning, BERT's pre-trained language model can be fine-tuned for specific purposes. In the last few years, several extensions of BERT have been introduced, both general, such as RoBERTa, and domain-specific, such as FinBERT. These, and other transformer models (e.g. XLNet), have been proven to outperform traditional NLP models, in particular in sentiment analysis problems.

At the same time, we can only identify a handful of articles applying language transformers to sentiment classification. Xie et al. (2020) performed sentiment

analysis of Chinese e-commerce reviews. They perform aspect-based analysis using automated sequence annotation. The authors extend the basic version of RoBERTa and find that it can achieve up to 90% accuracy. Li et al. (2021) develop a novel sentiment analysis model for Chinese stock reviews based on BERT. The authors find that the best performance (92% accuracy) is obtained by adding a linear layer to the BERT outputs and using a fully connected layer for prediction. In this article, motivated by the findings of the presented literature review, we aim to present a comparison of various traditional and deep learning-based machine learning models for sentiment classification, by making use of a newly annotated dataset in the context of e-commerce. Liao et al. (2021) propose a multi-task aspect-category sentiment analysis model based on RoBERTa. The authors use the RoBERTa based on deep bidirectional Transformer to extract features from both text and aspect tokens, and the cross-attention method to instruct the model to focus on the features most relevant to the given aspect category, treating each aspect category as a subtask.

3 Methodology

In this section, we will present the NLP-based methodology of sentiment analysis. Our main goal is to identify the best-performing models of sentiment classification based on customers' online reviews. As shown in Figure 1, we have applied five steps to achieve our research objective. In this section, we discuss data collection, annotation, preprocessing, and present the machine learning models used for sentiment classification.

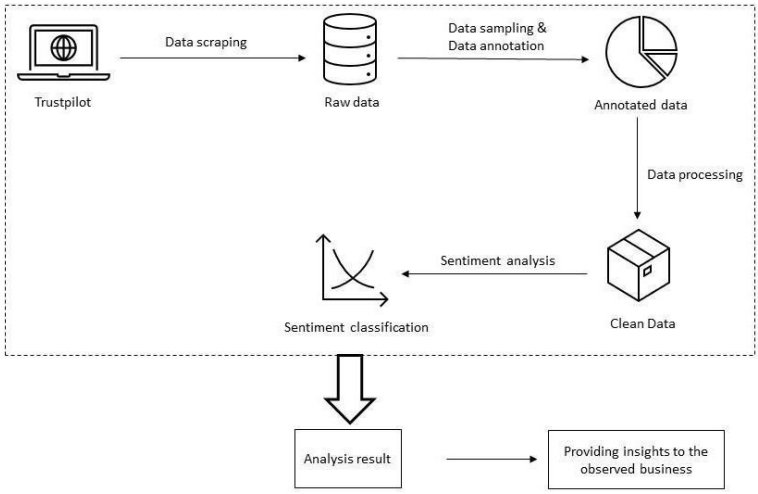

Figure 1 The stages of the research process

3.1 Data collection and processing

We aimed to collect data from one of the largest customer review sites, Trustpilot[1]. The data collection focused on reviews of online stores, written in English. After a preliminary check of available reviews, we opted to collect data from five e-commerce platforms, written in the time period between 2012 and 2021, with most of the comments from 2021. The total number of collected reviews is approx. 12,000. The stores included in the analysis are the following: Zalando (35% of data), Wish (24%), Sheinside (19%), Boozt (17%), and Nelly (5%). To complement the text of the reviews, the following additional information was collected: header of the review, numeric rating (range 1-5), location of the user, date of the reviews, total number of reviews previously written by the user.

In order to perform sentiment analysis, a polarity value needs to be assigned to each review. As highlighted in the literature review, while there exist lexicons for customer reviews, the results utilizing them are not fully convincing yet. For this reason, we opted for a manual sentiment annotation process. We have selected 3500 reviews randomly for the manual annotation and labeled each message according to the emotions found in the review. The data was annotated by two annotators, making use of three possible labels: positive, negative, and mixed. After the two annotators individually assigned sentiment values, every single disagreement was discussed until a consensus was reached. After resolving all the disagreements, we removed all the reviews of insufficient quality (e.g., too short, not comprehensible, not in English). The final dataset includes 3159 data points, consisting of 59% positive, 33% negative, and 8% mixed sentiment values. After the annotation, we performed text cleaning: transforming the text into lowercase, removing all the non-alpha characters, cleaning out the contractions, removing the HTML tags and URLs, and finally, lemmatization. After text processing, the total number of words is 108253 and the total number of unique words is 4766.

To perform sentiment classification on the reviews, the text of the reviews needs to be transformed into a set of features that is understandable for a classifier. The performance of the machine learning algorithms depends intensely on this feature extraction process, i.e. creating a feature vector representation (Zainuddin

[1] https://www.trustpilot.com/ The dataset is available from the authors by request.

et al., 2014). In this article, we make use of two feature extraction methods: Bag of Words (BOW) and Term Frequency Inverse Document Frequency (TF-IDF). BOW is a feature extraction technique that takes text data as input and produces an unordered collection of the unique vocabulary of the documents to be used in further classification process (Barry, 2017). When using TF-IDF, 2 components are assigned to each word and combined to obtain a final representation (Dang et al., 2020): (i) term frequency outlines how often a specific word occurs within a document, while (ii) inverse document frequency estimates the informativeness of a word.

3.2 Machine learning models for sentiment classification

In order to build and compare different sentiment classification models for customer reviews, we have chosen the most frequently used models in sentiment analysis literature: Naïve Bayes, Support Vector Machines, and BERT. Additionally, we have selected RoBERTa, as one of the recently introduced methods, gaining interest and showing good performance in various applications.

The Naïve Bayes classification algorithm is based on Bayes' theorem, and uses the idea of identifying the most probable class using the conditional frequency of occurrences in terms of the feature vectors. Naïve Bayes' has been applied successfully in many domains, such as text classification, medical diagnosis, and systems performance management (Rish, I., 2001; Dey et al., 2016).

Support-vector Machines (SVM) were originally introduced by Vapnik et al., (1997). One of the advantages of SVM models, namely their ability to handle large numbers of features, makes SVMs a widely used choice in text classification problems (Zainuddin et al., 2014). SVMs have shown excellent performance in prior sentiment analysis studies (Mullen et al., 2004).

Additionally to the traditional machine learning techniques, we selected two recently introduced language transformer models: BERT and RoBERTa. RoBERTa (Liu et al., 2019) is a modified version of BERT; although transformers (Vaswani et al., 2017) are used in both models as the main architecture, RoBERTa is trained differently compared to BERT. The major modifications in training the RoBERTa model are: 1) the model is trained longer, with larger batches and data, 2) the next sentence prediction objective is

removed, 3) using entire sentences as input, 4) text encoding, and 5) dynamic masking.

3.3. Model building

In order to identify the best-performing model, we tested different combinations of text feature extraction methods and machine learning algorithms. In the case of Naïve Bayes and SVM, we tested BOW and TF-IDF to transform the text to the numeric features, while for BERT and RoBERTa, we used a transformer tokenizer. Before model building, we divided the dataset into the train (80% of the data), validation (12%), and test (8%) sets. These sets were fixed for all the models, and we performed the analysis without any further resampling, such as cross-validation. As the dataset is imbalanced in terms of the number of labels, we used stratified random sampling. The validation set was used to identify the optimal cost parameters for the traditional machine learning models: (i) by changing the n-gram vectorizer parameters and regularization parameters for the linear SVM model, and (ii) n-gram vectorizer parameters and the α value for Multinomial Naïve Bayes. The same steps were performed for SVM and Naïve Bayes with BOW and TF-IDF features to compare the performance, and the best performing model in each case is selected as the one minimizing the number of misclassifications.

For text transformation in transformer-based models, we utilized tokenizers (AutoTokenizer for BERT, and RobertaTokenizer for RoBERTa). For BERT, we used the main model as the baseline, and added a dropout layer to control overfitting, and a dense layer for the classification task. When training RoBERTa, in addition to a dropout of 0.3 and a Linear layer, our network model includes the main model with 12 layers and 768 hidden dimensions. For both BERT and RoBERTa, the optimizer, loss function, and performance metric were specified as Adam, categorical cross-entropy, and accuracy, respectively.

To evaluate the performance of the models, we made use of some traditionally employed measures, namely accuracy and F1. Various measures of classification performance with binary outcome (positive and negative classes) can be defined using the confusion matrix. It has four components: (i) true positives (TP, positive cases correctly classified as positive), (ii) true negatives (TN, negative cases correctly classified as negative), (iii) false positives (FP, negative cases incorrectly classified as positive), and (iv) false negatives (FN, positive cases

incorrectly classified as negative). Using these notions, accuracy can simply be defined as the percentage of correctly classified cases, i.e., (TP+TN)/(TP+TN+FP+FN). This is the most widel used measure, although it has several issues, in particular in problems with imbalanced cases. One alternative measure is F1, which is defined as TP/(TP + 0.5(FP + FN)).

4 Results

In this section, we will present the results of the experiments and show the performance of the various machine learning models in sentiment classification, and compare our results to previous academic research. Furthermore, we discuss some observations highly relevant in managing expectations regarding this performance and provide illustrations using the constructed models.

4.1 Sentiment classification performance

Altogether, six different models were constructed as discussed in the previous section; the results (the best performance for all models after extensive parameter selection) are presented in Table 1. In the table, we present the accuracy and F1 value for the test set (and as a comparison to the validation set), the number of misclassifications in the test set, and the execution time of building the models.

Method	**Validation set accuracy**	F1 on validation set	**Test set accuracy**	**F1 on test set**	**Number of misclassification(test set)**	**Execution time(in seconds)**
Naive Bayes(TFIDF)	0.868	0.889	0.901	0.889	25	0.720
Naive Bayes(BOW)	0.889	0.883	0.916	0.906	21	5.493
SVM(BOW)	0.862	0.854	0.873	0.866	32	10.581
SVM(TFIDF)	0.907	0.895	0.905	0.889	24	1.192
BERT	0.902	0.889	0.921	0.899	20	1132.78
RoBERTa	**0.976**	**0.984**	**0.988**	**0.992**	**3**	378

Table 1 Sentiment classification performance

The results reveal some interesting insights. First of all, it is clear from the performance evaluation that language transformer-based models are superior to traditional machine learning models. Furthermore, RoBERTa outperforms even BERT significantly, with accuracy on the test set above 98%. Second, considering only the traditional ML models, the best accuracy can be achieved by Naive Bayes in combination with Bag of Words. Interestingly, SVMs perform better when used with features extracted with TF-IDF. Finally, regarding execution time, as it can be expected, transformer-based models take more time to train, which is the cost for the improved performance.

Regarding the specific misclassifications generated by the RoBERTa model, the mistakes are coming from identifying the Mixed polarity. It is not unexpected, as the number of reviews with Mixed polarity is fewer than the other 2 polarities. As an example misclassification, the following review was misclassified as Positive instead of Mixed:

I previously wrote a very angry review, as I paid for faster shipping and it hasn't been delivered any faster. This being said, SheInside saw my review and offered to compensate for the extra shipping costs. It was more the principle than anything—I hate feeling scammed. SheInside fixed those feelings and contacted me to solve my issue. Thank you!

4.2 Discussion

In the following, we will discuss the performance presented above from both a technical perspective and also the relevance for companies. First, regarding our main results, we have identified contributions from the literature that presented sentiment classification on user reviews (Fang et al., 2015; Nanda et al., 2018; Dey et al., 2016; Colón-Ruiz et al., 2020; Basani et al., 2019; Munna et al., 2020; Pipalia et al., 2020). SVM (e.g. Colón-Ruiz et al., 2020), Naive Bayes (e.g. Basani et al., 2019), and Neural Network models (Munna et al., 2020) are some of the most widely used methods in the sentiment classification task, i.e. the same models we have tested in this article. Although performance results are not directly comparable to previous research as we focused on a new dataset, our results with regards to traditional models align with previous findings: accuracy in the range of 85-95% is achievable, depending on the domain and also the language of the reviews. This can also be seen as a promising performance considering the execution time and simplicity of building such models for business applications. In our experiments, the neural network-based transformer

models seem to offer the best performance, which is in line with the very limited amount of research available.

Providing the input for classification models, data annotation, i.e., assigning sentiment polarity to the reviews, plays a crucial role. In the literature, machine labeling (e.g. Fang et al., 2015) and manual annotation (e.g. Munna et al, 2020) are the most used annotation methods. Automated annotations are less expensive in terms of time and cost in comparison to manual annotation, however in general they are less accurate. In this research, we chose manual annotation to acquire an understanding of the content of the reviews as related to the target companies. A third alternative approach to assign a sentiment to reviews would be to use the rating provided by the users, as it is done for example by Pipalia et al. (2020). Our manual inspection during the annotation process has shown that this approach could be problematic, as there are a large number of reviews when there is a disagreement between the rating and the sentiment of the review. For example, the following review was accompanied with the rating of 5 (on a 1-5 scale, with 5 being the highest value), although it is clear that the associated sentiment should be Negative:

"Some items are too expensive on the shipment and is too long for the wait"

The following is a similar example with the rating of 5 but clearly not a Positive sentiment:

I ordered a big clock for my wall it came in so very small I want the big one please

As these and numerous similar examples show, using the rating can be deteriorating for the performance of sentiment classification. Companies should be careful when assessing their popularity and the sentiment level of customers based on ratings. As our results show, manually annotating a subset of reviews can result in models that can offer close to perfect sentiment polarity classification. In fact, the following example from our test set was correctly classified as Negative, while the original user rating was 4:

Deliveries take too long. Postage added is a turn off

The following example review was also classified as Negative, although the user rating was 4:

My order says shipped but has no information on its whereabouts. It only said "processing" for about a day before it said my order was shipped. I'm confused

As a final point relevant to businesses, we can note that the best sentiment classification model, RoBERTa, has performed equally well across the different shops present in the data. The accuracy was observed as follows: 97% for Boozt, 98% for Nelly, 100% for Sheinside, 100% for Wish, and 98% for Zalando. These results illustrate the great potential of transformer-based models in sentiment classification.

5 Conclusions

With the wide availability of unstructured big data, in particular textual information in the form of customer reviews, we observed a rapid development in data processing and analysis techniques to make sense of all this information. Sentiment analysis has been discussed and evaluated numerous times in the literature, however, since there is a plethora of new models appearing, it is always an important task to critically assess and compare new and established models. In particular for e-commerce platforms, sentiment analysis of online reviews can be the main source of information to understand the opinions of customers. In this article, we have compared the performance of traditional machine learning models with language transformer models, and found that the neural network-based models offer much higher accuracy in sentiment classification tasks. We illustrated the value of manual annotation through several examples of problems with user ratings that can be avoided by the sentiment classification models built using manually assigned sentiment values.

Regarding future research, the most important continuation of the work will focus on performing aspect-based sentiment classification. While we have shown that very high performance can be achieved using transformer-based models, this is sufficient to understand the reasons for the polarity of sentiments. In the case of the analyzed reviews of e-commerce stores, instead of assigning a general sentiment, one could assess the sentiment polarity with respect to, e.g., shipping, received item quality, payment process. By constructing models that can identify the core aspects in the reviews and the associated sentiment, companies will know where to focus their efforts to improve customer satisfaction. Additionally, some limitations of the study have to be acknowledged. First, we cannot say that the sample of reviews is representative in any way, so models need to be tested on larger datasets and different e-commerce stores. Second, while the annotation was performed and then cross-checked by two researchers, there is still a possibility of incorrect sentiment assignments, which in turn may impact the constructed models and performance. Third, while we used the models most frequently utilized in the sentiment classification literature, there are numerous machine learning models available that could be tested for performance.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Basani, Y., Sibuea, H. V., Sianipar, S. I. P., & Samosir, J. P. (2019, March). Application of sentiment analysis on product review e-commerce. In Journal of Physics: Conference Series (Vol. 1175, No. 1, p. 012103). IOP Publishing.

Barry, J. (2017). Sentiment Analysis of Online Reviews Using Bag-of-Words and LSTM Approaches. In AICS (pp. 272-274).

Bhatti, A., Akram, H., Basit, H. M., Khan, A. U., Raza, S. M., & Naqvi, M. B. (2020). E-commerce trends during COVID-19 Pandemic. International Journal of Future Generation Communication and Networking, 13(2), 1449-1452.

Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. Journal of Biomedical Informatics, 110, 103539.

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. Electronics, 9(3), 483.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifier. arXiv preprint arXiv:1610.09982.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 1-14.

Gangula, R. R. R., & Mamidi, R. (2018, March). Impact of Translation on Sentiment Analysis: A Case-Study on Telugu Reviews. In 19th International Conference on Computational Linguistics and Intelligent Text Processing.

Han, H., Zhang, J., Yang, J., Shen, Y., & Zhang, Y. (2018). Generate domain-specific sentiment lexicon for review sentiment analysis. Multimedia Tools and Applications, 77(16), 21265-21280.

Han, H., Zhang, Y., Zhang, J., Yang, J., & Zou, X. (2018a). Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias. PLoS One, 13(8), e0202523.

Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing (pp. 639-647). Springer, Singapore.

Li, M., Chen, L., Zhao, J., & Li, Q. (2021). Sentiment analysis of Chinese stock reviews based on BERT model. Applied Intelligence, 51(7), 5016-5024.

Liao, W., Zeng, B., Yin, X., & Wei, P. (2021). An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. Applied Intelligence, 51(6), 3522-3533.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.

Muflikhah, L., & Haryanto, D. J. (2018). High performance of polynomial kernel at SVM Algorithm for sentiment analysis. Journal of Information Technology and Computer Science, 3(2), 194-201.

Mullen, T., & Collier, N. (2004, July). Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 412-418).

Munna, M. H., Rifat, M. R. I., & Badrudduza, A. S. M. (2020, December). Sentiment analysis and product review classification in e-commerce platform. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

Nanda, C., Dua, M., & Nanda, G. (2018, April). Sentiment analysis of movie reviews in hindi language using machine learning. In 2018 International Conference on Communication and Signal Processing (ICCSP) (pp. 1069-1072). IEEE.

Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77).

Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

Pipalia, K., Bhadja, R., & Shukla, M. (2020, December). Comparative analysis of different transformer based architectures used in sentiment analysis. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 411-415). IEEE.

Piris, Y., & Gay, A. C. (2021). Customer satisfaction and natural language processing. Journal of Business Research, 124, 264-271.

Rajput, A. (2020). Natural language processing, sentiment analysis, and clinical analytics. In Innovation in Health Informatics (pp. 79-97). Academic Press.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Shuai, Q., Huang, Y., Jin, L., & Pang, L. (2018, October). Sentiment analysis on Chinese hotel reviews with Doc2Vec and classifiers. In 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 1171-1174). IEEE.

Sun, Q., Niu, J., Yao, Z., & Yan, H. (2019). Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level. Engineering Applications of Artificial Intelligence, 81, 68-78.

Tsai, C. F., Chen, K., Hu, Y. H., & Chen, W. K. (2020). Improving text summarization of online hotel reviews with review helpfulness and sentiment. Tourism Management, 80, 104122.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Xie, S., Cao, J., Wu, Z., Liu, K., Tao, X., & Xie, H. (2020, July). Sentiment Analysis of Chinese E-commerce Reviews Based on BERT. In 2020 IEEE 18th International Conference on Industrial Informatics (INDIN) (Vol. 1, pp. 713-718). IEEE.

Zainuddin, N., & Selamat, A. (2014, September). Sentiment analysis using support vector machine. In 2014 international conference on computer, communications, and control technology (I4CT) (pp. 333-337). IEEE.

Zeng, D., Dai, Y., Li, F., Wang, J., & Sangaiah, A. K. (2019). Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism. Journal of Intelligent & Fuzzy Systems, 36(5), 3971-3980.