ReviewCraft : A Word2Vec Driven System Enhancing User-Written Reviews

Gaurav Sawant, Pradnya Bhagat and **Jyoti Pawar**
Department of Computer Science and Technology, Goa University
gauravrsawant1313@gmail.com
pradnyabhagat91@gmail.com
jdp@unigoa.ac.in

Abstract

The significance of online product reviews has become indispensable for customers in making informed buying decisions, while e-commerce platforms use them to fine tune their recommender systems. However, since review writing is purely a voluntary process without any incentives, most customers opt out from writing reviews or write poor-quality ones. This lack of engagement poses credibility issues as fake or biased reviews can mislead buyers who rely on them for informed decision-making. To address this issue, this paper introduces a system that suggests product features and appropriate sentiment words to help users write informative product reviews in a structured manner. The system is based on Word2Vec model and Chi square test. The evaluation results demonstrates that the reviews with recommendations showed a 2 fold improvement both, in the quality of the features covered and correct usage of sentiment words, as well as a 19% improvement in overall usefulness compared to reviews without recommendations.

Keywords: Word2Vec, Chi-square, Sentiment words, Product Aspect/Feature.

1 Introduction

In the current digital era, customers view product reviews over the internet as a vital resource in their buying process (Salehan and Kim, 2016). The increase in the number of online products reviews makes it difficult for consumers to search for genuine and valuable information concerning a particular product. The companies are also held accountable by publicly shared reviews, which in turn prompt them to offer better products and services to maintain a good brand image (Kim and Srivastava, 2007). But there are many poor quality review that do not provide useful information, are biased and bad written therefore potential buyer cannot use these review to make informed decision to buy. This issue has led to a growing need for an advanced review system that can motivate more people to write good quality product reviews.

One of the primary challenges in review writing is the difficulty in articulating specific product features and their associated sentiment. For instance a customer may wish to comment on how bad the camera of a mobile phone is but they might not be aware of the technical words and features associated with camera specifications. For example, when expressing dissatisfaction with the camera, a customer may simply state that "The camera of this model is bad". Although the sentiment is conveyed here, this sentences is not of much use to other users since it is not really conveying the issues that make the camera bad. Most of the reviews are filled with such generic sentences. However, if a customer is able to pinpoint the specific problems that make the camera bad, for example, "The camera is of low megapixels" or " The night photography is bad", these details in the reviews can make the reviews submitted by the customers much more usable than the generic review, but most laymen users are not aware of such features or may not remember the intricacies when writing reviews, especially since review writing is often done as a casual activity. Besides this, the sentiment words may not convey similar meaning when they are used with different features, thus leading to ambiguity. for example, the word “slow” in "slow battery drain" might signify that the phone’s battery is long lasting but it will be a negative sentiment if it is referring to the "processing speed".

To address these challenges, this paper proposes a novel approach that is based on Word2Vec model (Mikolov et al., 2013) and Chi-square test (Greenwood and Nikulin, 1996), to assist users in writing more effective and valuable reviews. This paper is based on (Bhagat and Pawar, 2022) where authors propose a review writing recommender system based on LDA model to extract features, our approach extends beyond this method. Unlike

the LDA model, which falls short in finding similar words or examining words at a more detailed level, we utilize the Word2Vec model to extract relevant product features. The identification of feature-specific sentiment words in this paper is based on (Bhagat et al., 2023). These words are subsequently presented to users in a structured manner, ensuring an enhancement in both the quantity and quality of information within the reviews. Thus it provides potential consumers with an extra basis for choosing rightly. The rest of the paper is organized as follows: Section 2 provides a literature review of existing work on product reviews and recommender systems. Section 3 describes the proposed system and its components in detail. Section 4 provides the details of the experiment and the dataset used. Section 5 presents the results obtained. Finally, section 6 concludes the paper and highlights future direction.

2 Literature Survey

There has been a significant amount of research in the fields of social networking and e-commerce, specifically in the areas of Natural Language Processing (NLP) and Machine Learning (ML) algorithms. (Bridge and Healy, 2012) presents GhostWriter-2.0, a Case-Based reasoning system that suggests topics for new product reviews based on existing reviews. (Dong et al., 2012b) introduced a tool called Reviewer's Assistant which focuses on the development of the browser plugin as a tool to assist users in writing better quality reviews. The study suggests a unique methodology exploiting association rule mining for extracting relevant product attributes based on the content being typed by the user. (Dong et al., 2012a) describes practical implementation and assessment for a browser plugin. A user study was then carried out by the researchers to assess how useful the plugin could be in improving the quality of the reviews generated by users. (Dong et al., 2012c) proposes a system which uses unsupervised topic extraction with the LDA algorithm to suggest relevant topics for users writing product reviews. Another study (Bhagat and Pawar, 2018) conducted a comparative analysis of various methods used in literature for extracting features from user reviews. (Bhagat and Pawar, 2020) presents a method to extract features using Latent Dirichlet Algorithm (LDA), this study can be useful for grouping related words into topics and extracting features, it may not be as effective for finding similar words or performing word-level analysis. (Sharma and Bhattacharyya, 2013) applies the Chi-square test statistical measure to determine the polarity of sentiment words by examining their occurrence in positive and negative reviews. While this work focuses on polarity differences at the domain level, it does not explore potential changes in polarity at the feature level.

3 Methodology

The proposed work aims to extract aspects and aspect-based polarity of sentiment words in user reviews. The methodology involves two key steps: aspect extraction using Word2Vec model and finding the aspect-based polarity of the sentiment word using the Chi-square test statistical measure.

The Word2Vec is a neural network-based model that generates word embeddings, representing words as vectors in a high-dimensional space. This allows for mathematical operations like addition and subtraction, enabling the system to capture the semantic meaning of words and their relationships. For example, if we wanted to find a word that is similar to "iPhone" we could perform the operation such as

Smartphone - Android + IOS = iPhone

The word smartphone gets a certain vector representation based on its context and so do Android and IOS. By subtracting the vector for "Android" and then adding the vector for "iOS", the resulting vector could theoretically be closest to the word embedding for "iPhone".

In our approach, we utilize Word2Vec to extract aspects from user reviews. We first preprocess the data by removing stop words, punctuation and converting all text to lowercase. Next, we train the model on a large corpus of preprocessed product reviews, where it learns the underlying relationships between different product features based on their co-occurrence in the corpus. The model generates word embeddings, where (word) vectors are close together if the corresponding words have similar meanings or appear in similar contexts.

To find similar words, the cosine similarity between the vectors of each word is calculated. The words with the highest cosine similarity scores are considered the most similar in meaning to the target word. Given a mentioned product feature in a review, the Word2Vec model helps in suggesting other related features, empowering users to include

comprehensive aspects in their reviews. The users are free to accept or reject the suggestions made by the system.

Once we have extracted the aspects, we need to find the associated sentiment words and the aspect-based polarity of those sentiment words using the Chi-square test. Our focus is on identifying adjectives that occur alongside feature nouns, as these adjectives are likely to convey sentiment towards the feature. However, not all nouns in a review contribute to product features. To address this, we employ Part-of-Speech (POS) tagging and focus specifically on nouns that appear in close proximity to adjectives. This association suggests a strong likelihood that the noun is a feature noun.

Our assumption is based on the observation that when users share their experiences with a specific product or its features, they often express their sentiments using adjectives. For instance: "This is a really good smartphone with an excellent camera and a stunning display." in this sentence the nouns smartphone, camera, and display have some adjectives associated with them.

Next, we categorize reviews as positive or negative based on their star ratings. Negative reviews are those with a rating of 1 star, while positive reviews have a rating of 5 stars. Then we prioritize nouns accompanied by adjectives as feature nouns and extract sentiment-feature pairs for further analysis. We formulate a null hypothesis that sentiment words are neutral for aspects. To validate this hypothesis, we examine the frequency of each sentiment-feature pair in both positive and negative reviews. If a pair exhibits significantly higher occurrence in one category, we reject the null hypothesis and assign polarity accordingly. Otherwise, the word is considered neutral for that aspect. For instance, the word "heavy" can have different sentiments depending on the feature it is associated with. "This smartphone feels heavy in my hand," this may be interpreted as negative sentence since such device may be unpleasant in hands for long time. But if we say "the new rugged smartphone is built with heavy-duty materials," makes sense in a positive way. We then use chi-square as our statistical measure of whether the adjective tends to lean toward one of the sentiment categories (i.e., positive or negative) with a particular attribute. Also, the chi-square test can help determine whether the presence of a particular sentiment term and a feature in a particular category of reviews is significant beyond chance. The threshold value for the Chi-square value taken as a measure of significance of a particular sentiment-feature pair is 1.76. Thereafter, we compute the Chi-squared value for every sentiment-feature pair by comparing the actual observed count to the expected count for sentiment words with regard to a feature in both the category (positive and negative). A chi-squared value greater than the threshold would indicate the occurrence of the sentiment word in one of the categories of reviews is not chance but the effect of domain specific polarity with regard to a feature.

4 Implementation Details and Dataset Used

The implementation of the experiment is done in Python programming language (Sanner et al., 1999). For text processing tasks, we utilize the Natural Language Toolkit (NLTK) library (Loper and Bird, 2002). Part-of-speech tagging is performed using the spaCy library (Srinivasa-Desikan, 2018). The Gensim library is employed to train a Word2Vec model on pre-processed review text. The dataset employed in our experiment consists of reviews sourced from Amazon.com (Ni et al., 2019). Specifically, we conduct our tests on the dataset pertaining to Cell Phones and its related accessories

We conducted an experiment on a group of 29 students. These students were requested to write reviews about the mobile phones they currently use, on a website we created, sharing their levels of satisfaction or dissatisfaction. The students wrote reviews in the first phase without a recommendation system. In the second stage, the students again wrote another review, although they were provided with an automated recommendation system that provided guidance about the features and sentiment words that they could incorporate in the review as shown in Figure 1. Two human judges evaluated the quality of every student's review. The judges were provided with three parameters: 1) The quality of features covered, 2) The correct usage of sentiment words to describe the features and 3) The overall usefulness of the review. The reviewers used a three-point scale: i.e. 1 – poor, 2 – average and 3 – good to rate the reviews. The Fleiss's kappa inter-rater reliability measure was used to monitor how much agreement existed between the judges. The Fleiss kappa was interpreted using the Landis and Koch criteria with kappa value less then 0 indicating poor agreement, 0.01–0.20 slight agree-

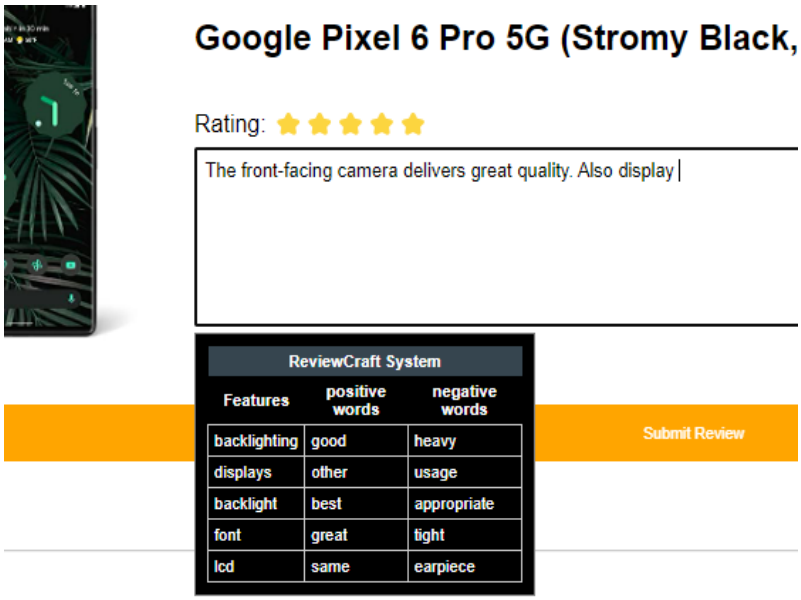

Figure 1: ReviewCraft system demonstration

Features	Similarity
oled	0.7752
pixels	0.7331
lcd	0.7171
ips	0.7126
resolution	0.7073
pixel	0.7016
amoled	0.6977
retina	0.6892
backlit	0.6838

Table 1: Top 10 similar aspects obtained using Word2Vec model for aspect “display”.

Noun	Adjective	Sentiment
reset	hard	Negative
phones	smart	Neutral
phones	worst	Negative
phones	good	Positive
phones	best	Positive
phones	bad	Negative
phones	better	Neutral

Table 2: Results obtained using Chi-Square test.

ment, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81–1.00 indicating almost perfect agreement (Landis and Koch, 1977) (Driessen et al., 2021).

5 Results

Table 1 displays the results obtained from training the Word2Vec model on a dataset of 200,000 reviews. It showcases the top similar words, ranked by cosine similarity of their corresponding vectors

Table 2 presents the outcomes of the Chi-Square test containing sentiment words categorized as positive and negative, indicating their significance. Words labeled as neutral are considered insignificant as they do not convey explicit positive or negative sentiments about the feature. For instance, the term "smartphone" is classified as neutral as it does not express any specific positive or negative sentiment related to the feature.

Table 3 shows the results of our inter-rater reliability analysis using Fleiss’ Kappa for three key parameters related to the quality assessment of reviews. For all the parameters we observe a fair level of agreement between the judges for reviews without recommendations. For reviews with recommendations, we see moderate level of agreement between the judges. Table 4 displays sample reviews written by students, with and without recommendations.

Figure 2 indicates the average of the ratings by Judge 1 and Judge 2. The average of the ratings given by both the judges shows that out of 29, 6 reviews were reported to be of quality in terms of quality of features mentioned. We can see from the graph that after the use of the system, the number increased to 11. Similarly, for correct use of sentiment words, without the use of any help, only in 4

	Reviews without Recommendations	Reviews with Recommendations
Quality of features covered	0.4423	0.5250
Correct usage of sentiment words to describe the feature	0.3614	0.5816
Overall usefulness of the review	0.3088	0.4276

Table 3: Agreement between the judges calculated by Fleiss Kappa

Reviews without Recommendations	Reviews with Recommendations
The phone is good for many apps, but sometimes the camera gets stuck. Battery is not good. But it has a good zoom to take close pictures.	The phone is optimized for handling multiple apps, but the camera is laggy, it also has nice zoom feature for closeups. Charging speed is decent but drain quickly barely lasts a day. the audio quality for calls is clear and loud.
This is a great phone I have been using it for 4 years. The camera quality is also great. The only issue is the video lags a lot.Overall for such a cheap rate the phone is pretty good.	It is great phone which has been used for very long time. So now there is some issue with the phone while gaming streaming as the phone starts heating up which is pretty frustrating. It is pretty heavy so it will be pretty good if you guys make lighter phones.
Its the best purchase I have done. It has been there with me for more Two years now and still working as I have brought new. just you might have the problem with the battery . its overall good phone to purchase. i would recommend to buy its as you get most of the feature in cheap price.	camera megapixels is excellent. the back camera is great the display is decent. the speakers are durable. the phone might you feel heavy and bulky.the overall features and capabilities are good at this price .
The phone is very slim and handy. The fingerprint response to unlock the phone is fast takes about 2-3 secs. If the phone has lot of apps installe on it the battery won't last for long. It has a very user friendly UI	I had this phone for the last 4 years its a g8 budget phone, both the front and rear cameras are wonderful, sound quality is quite good, display quality is very good but the saturation is on the higer side, battery is decent with 4000mah, looks very sleek and is also very light gives a premium feel, but the box doesnt contain earphone, and also the sim tray will either support one sim and a sd-card or two sims.

Table 4: Example of reviews written by students

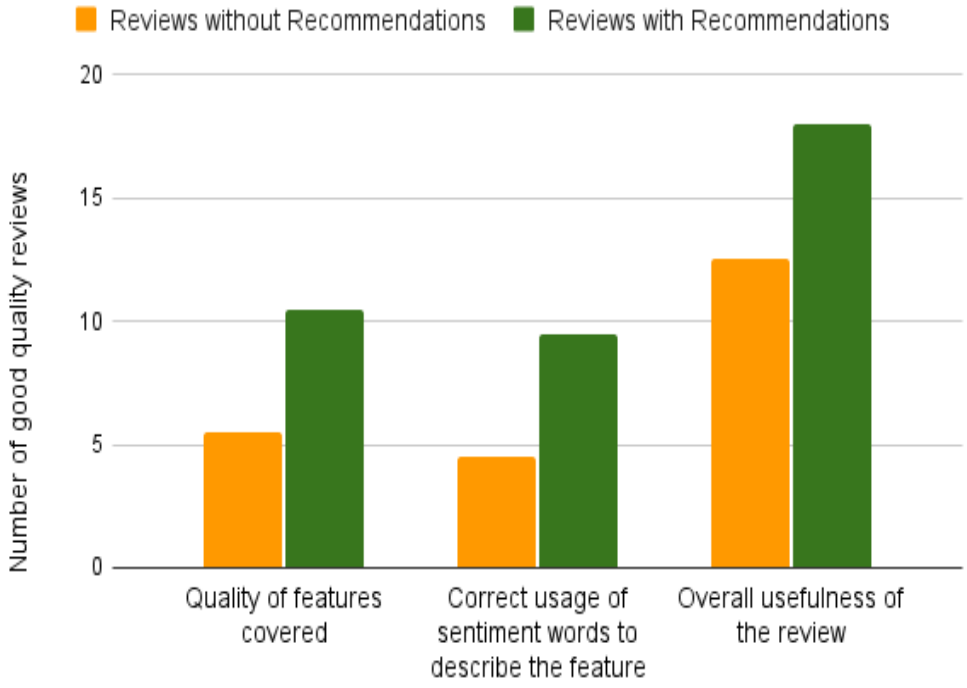

Figure 2: Graph showing number of good quality reviews (average of rating provided by both the judges) with and without using the ReviewCraft Recommendation System.

reviews, people used more informative sentiment words to describe the features; which increased to 9 after using the system. And in terms of the overall usefulness of the review the number of reviews increased from 12 to 18.

6 Conclusion

Good quality product reviews play a vital role in establishing a robust ecommerce system. They enable customers to express their opinions effectively and ensure that their voices are acknowledged and addressed. This research introduces a promising approach to help users write informative reviews confidently and efficiently, improving the overall quality of product reviews on e-commerce websites.

The word2vec model helps to suggest other related features by learning from the context of the user reviews. Use of the Chi-Squared test helps to filter out a lot of words with low sentiment intensity by assigning the sentiment orientations to words whose Chi Squared value is above a particular threshold. The evaluation results indicate a notable 19% improvement in the overall quality of the product reviews with the use of the recommendation system, compared to reviews written without it. As future work, we aim to simplify the review writing process, enhancing user-friendliness and accessibility, while also addressing language barriers.

References

Pradnya Bhagat, Pratik D Korkankar, and Jyoti D Pawar. 2023. Aspect-based sentiment words and their polarities using chi-square test. *Computación y Sistemas*, 27(2).

Pradnya Bhagat and Jyoti D Pawar. 2018. A comparative study of feature extraction methods from user reviews for recommender systems. In *Proceedings of the ACM India joint international conference on data science and management of data*, pages 325–328.

Pradnya Bhagat and Jyoti D Pawar. 2020. A two-phase approach using lda for effective domain-specific tweets conveying sentiments. In *Computational Intelligence and Machine Learning: Proceedings of the 7th International Conference on Advanced Computing, Networking, and Informatics (ICACNI 2019)*, pages 79–86. Springer.

Pradnya Bhagat and Jyoti D Pawar. 2022. A product review writing recommender system based on lda and tf-idf. *Computación y Sistemas*, 26(3):1107–1117.

Derek Bridge and Paul Healy. 2012. The ghostwriter-2.0 case-based reasoning system for making con-

tent suggestions to the authors of product reviews. *Knowledge-Based Systems*, 29:93–103.

Ruihai Dong, Kevin McCarthy, Michael O'Mahony, Markus Schaal, and Barry Smyth. 2012a. Towards an intelligent reviewer's assistant: recommending topics to help users to write better product reviews. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 159–168.

Ruihai Dong, Markus Schaal, Michael P O'Mahony, Kevin McCarthy, and Barry Smyth. 2012b. The reviewer's assistant: Recommending topics to writers by association rule mining and case-base reasoning. In *The 23rd Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2012), Dublin City University, Dublin, Ireland*, pages 17–19.

Ruihai Dong, Markus Schaal, Michael P O'Mahony, Kevin McCarthy, and Barry Smyth. 2012c. Unsupervised topic extraction for the reviewer's assistant. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 317–330. Springer.

Rob GH Driessen, Nanon FL Heijnen, Riquette PMG Hulsewe, Johanna WM Holtkamp, Bjorn Winkens, Marcel CG van de Poll, Iwan CC van der Horst, Dennis CJJ Bergmans, and Ronny M Schnabel. 2021. Early icu-mortality in sepsis–causes, influencing factors and variability in clinical judgement: a retrospective cohort study. *Infectious Diseases*, 53(1):61–68.

Priscilla E Greenwood and Michael S Nikulin. 1996. *A guide to chi-squared testing*, volume 280. John Wiley & Sons.

Young Ae Kim and Jaideep Srivastava. 2007. Impact of social influence in e-commerce decision making. In *Proceedings of the ninth international conference on Electronic commerce*, pages 293–302.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

Mohammad Salehan and Dan J Kim. 2016. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, 81:30–40.

Michel F Sanner et al. 1999. Python: a programming language for software integration and development. *J Mol Graph Model*, 17(1):57–61.

Raksha Sharma and Pushpak Bhattacharyya. 2013. Detecting domain dedicated polar words. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 661–666.

Bhargav Srinivasa-Desikan. 2018. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.