Enhancing Consumer Decision-Making: An Integrated Approach to E-commerce Platforms

PERAM SREE KEERTHAN REDDY
EE20BTECH11040
ee20btech11040@iith.ac.in

Abstract

The exponential growth of e-commerce has ushered in a new era of convenience and accessibility for consumers worldwide. However, amidst the abundance of choices, customers often grapple with confusion, leading to challenges in making informed purchasing decisions. This phenomenon, termed as 'e-confusion' or customer confusion, has become increasingly prevalent, primarily driven by similarity, overload, and unclarity. In e-commerce, understanding customer sentiments from product reviews is pivotal for businesses to gauge consumer satisfaction. This report delves into a comprehensive analysis methodology employing Natural Language Processing (NLP) techniques, particularly sentiment analysis using Support Vector Machines (SVM), to decipher sentiments expressed within textual product reviews.

1. Introduction

The proliferation of e-commerce platforms has revolutionized the retail landscape, offering consumers unparalleled access to a myriad of products and services. However, this abundance of choices has ushered in a new psychological challenge termed 'e-confusion' or customer confusion. This phenomenon, characterized by the difficulty in making informed purchasing decisions amidst the overwhelming array of options, poses a significant hurdle in the online shopping experience.

The past decade has witnessed an unprecedented surge in e-retail sales, crossing the staggering mark of 5.2 trillion USD globally in 2021. This meteoric rise in online commerce, while emblematic of convenience and accessibility, has also contributed substantially to the conundrum of customer confusion. The complexity of navigating numerous e-commerce websites to make a purchasing decision often results in ill-informed choices, potentially leading to mistrust in online shopping and financial repercussions for consumers.

The underlying principles of e-confusion encompass similarity confusion, overload confusion, and unclarity confusion. In emerging markets like India, additional factors such as product pricing, ratings, reviews, and availability significantly influence consumer behavior, contributing substantially to their purchasing decisions. Recognizing these challenges, this paper focuses on addressing the core issue of consumer indecision through an integrated approach that harnesses the power of data analytics methodologies.

This report delves into the domain of sentiment analysis, a facet of Natural Language Processing (NLP), particularly focusing on the application of Support Vector Machines (SVM) for discerning sentiments within product reviews. The objective is to harness machine learning techniques to analyze and categorize sentiments expressed by customers towards various products, thereby aiding businesses in making data-driven decisions to enhance customer satisfaction and optimize product offerings.

The advent of SVM, a robust supervised learning algorithm known for its efficacy in classification tasks, presents a promising avenue for sentiment analysis. Leveraging this model, the study endeavors to classify textual product reviews into distinct sentiment categories, including positive, negative, and neutral, based on the underlying sentiments expressed by customers.

In view of time constraints and the focus on a comparative analysis, the study narrows down the evaluation to one prominent e-commerce platform: Amazon. The methodology involves web scraping product queries from both platforms and extracting product reviews, enabling individual sentiment analysis, particularly on Amazon.

This project work is done under the guidance of Dr. Saumya Jana for the course EE2802 (Machine Learning).

2. Literature Review

Many publications are trying many models to solve this problem. Some of the most popular papers are Unsupervised Data Augmentation for Consistency Training [1], Re-

visiting LSTM Networks for Semi-Supervised Text Classification via Mixed Objective Function [2], XLNet: Generalized Autoregressive Pretraining for Language Understanding [3].. etc. These papers use LSTM, Bert, XLNet models for achieving good accuracy. In this paper we are following SVM model for sentimental analysis and compare the accuracies along with other models. The paper in which we followed reference is Real-time Sentiment Analysis On E-Commerce Applications [4], as we want to apply this for real-time applications.

3. Motivation

The motivation behind this research stems from several pivotal factors:

Navigating the Complexity of Choice: The proliferation of e-commerce platforms presents consumers with an extensive array of products, brands, prices, and reviews. This abundance contributes to decision-making complexities, hindering consumers' ability to identify the most suitable products or the optimal platforms for their needs.

Impact on Consumer Trust and Satisfaction: Ill-informed purchases resulting from customer confusion can lead to dissatisfaction, erode trust in e-commerce platforms, and sometimes incur financial losses. Addressing this issue is crucial in fostering trust and enhancing consumer satisfaction in the online shopping experience.

Emerging Market Dynamics: In emerging markets such as India, consumer behaviour is notably influenced by factors like product pricing, reviews, and availability. Understanding and addressing these market-specific influences are imperative to cater to consumers' unique needs and preferences in these regions.

The potential of Data-Driven Solutions: The application of data analytics methodologies like Knowledge Discovery in Databases (KDD) and sentiment analysis presents an opportunity to extract valuable insights from vast datasets. Leveraging these insights can aid in deciphering consumer sentiments, preferences, and behaviours, ultimately facilitating informed decision-making in e-commerce.

Empowerment of Consumers: Developing tools or applications that streamline decision-making by offering comprehensive information about products, platforms, and customer reviews empowers consumers to make informed choices. This empowerment is pivotal in enhancing consumer satisfaction and confidence in online retail transactions.

This paper addresses the multifaceted challenges posed by customer confusion in e-commerce. By employing data-driven methodologies and focusing on consumer-centric solutions, this study aims to enhance the online shopping experience, foster trust, and aid consumers in making well-informed purchasing decisions.

4. Data Collection Methodology

In order to accommodate the impromptu and customizable nature of this project, a dynamic approach to dataset generation was adopted, necessitating the creation of three distinct modules for web scraping tailored to specific purposes.

Sources and Collection Procedure:

The primary sources for the dataset generation were the official websites of leading e-commerce platform namely Amazon. The collection process involved the development of a generalized algorithm for web scraping product information from these platforms. This algorithm facilitated the extraction of comprehensive details about various products, including their names, prices, ratings, review counts, product images, and URLs.

A dedicated module was specifically designed for scraping product reviews for the previously obtained urls. This module systematically traversed through all customer comments related to the products. Post data scrubbing and preprocessing, the information gleaned from the reviews was stored in a structured dataset format.

Generated Datasets:

Throughout the data collection process, two main datasets were constructed:

1. **Web Scraping Amazon:** This dataset was formed based on specific product queries provided by the client, extracting pertinent information from Amazon's platform.

    - `product_name`
    - `product_price`
    - `product_rating`
    - `rating_count`
    - `product_image_url`
    - `product_url`

2. **Web Scraping Reviews:** Similar to the Amazon product scraping, we are using the beautiful soup library to extract the product reviews. The dataset is collected for each product name, user name, review content, and related fields.

    - `User_name`
    - `product_name`
    - `product_rating`
    - `rating_count`
    - `NoA_Phrase`
    - `NoV_Phrase`

    These are the major dataset columns used for analysis.

5. Data Preprocessing

a) Tokenization: Tokenization is the initial step in breaking down entire product review sentences into individual tokens, such as words, phrases, or symbols. During tokenization, special characters like exclamation marks and semicolons are removed to create a cleaner text corpus suitable for analysis.

b) Removing Stop Words: Stop words, commonly occurring words (e.g., "the," "and," "is") that do not contribute significantly to the meaning of text, are eliminated from the product reviews. This step aids in enhancing the efficiency and accuracy of sentiment analysis. The selection of stop words varies according to language and country-specific formats.

c) POS Tagging: Part-of-speech (POS) tagging involves assigning specific parts of speech (e.g., noun, pronoun, verb, adjective) to individual words within the product reviews. POS tagging provides linguistic context and assists in identifying word categories, aiding in deeper semantic analysis.

d) Stemming: Stemming, a text normalization technique, involves reducing words to their root or base form. It aims to unify related words into a common root, thereby reducing the dimensionality of the text data. Stemming is crucial in text processing for information retrieval, assisting in recognizing variations of words within reviews for better analysis.

6. Methodology

a) Data Collection: The initial phase involves collecting raw textual product reviews from various e-commerce platforms. These reviews may contain multiple sentences expressing customers' opinions and sentiments towards specific products.

b) Data Preprocessing: As described above, the collected textual data undergoes cleaning, tokenization, stop words removal, and TF-IDF vectorization to prepare it for sentiment analysis.

c) Model Selection: For sentiment analysis, Support Vector Machines (SVM) are chosen due to their efficacy in text classification tasks. SVM is a supervised learning algorithm known for its ability to classify text into different sentiment categories based on learned patterns.

d) Model Training: The preprocessed dataset is divided into training, validation, and testing sets. The SVM model is trained using the training data and fine-tuned using the validation set. During training, the SVM model learns the patterns and relationships between textual features and sentiment labels.

e) Model Evaluation: The trained SVM model's performance is evaluated using the test dataset, measuring metrics such as accuracy, precision, recall, and F1-score. These metrics gauge the model's effectiveness in correctly classifying sentiments from unseen reviews.

f) Prediction and Sentiment Labeling: After the model's successful evaluation, it is used to predict sentiment labels (Positive, Negative, Neutral) for new, unseen product reviews.

g) Deciding the best product: After evaluating the positive, negative and neutral reviews the program gives the Product_name and Product_URL od the best product.

The detailed workflow architecture is shown in figure-1 (a).

The Data Processing workflow is shown in figure-1 (b).

6.1. Sentiment Sentence Extraction and Pos Tagging

The process of sentiment sentence extraction involved isolating sentences within product reviews that conveyed users' sentiments towards the products. Sentiment sentences were identified through various techniques, including sentiment analysis algorithms or rule-based methods. These sentences encapsulated customers' opinions, feelings, or evaluations about the products, providing valuable insights into their sentiments.

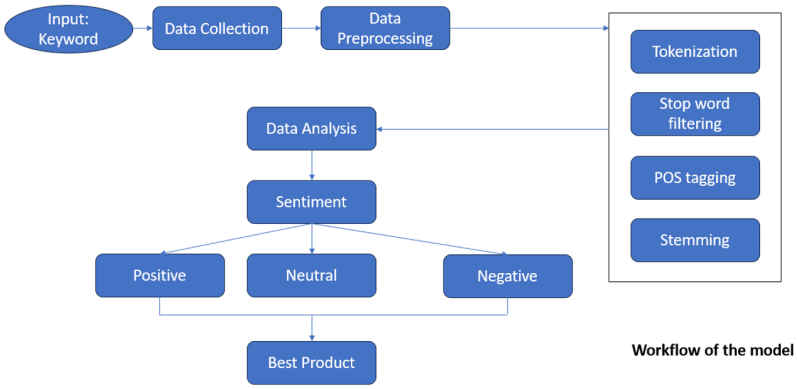
Simultaneously, the review text underwent preprocessing steps to enhance the sentiment analysis. Initially, the text was tokenized into individual English words. Subsequently, stop words, such as "in," "that," "is," "are," "so," "but," etc., were removed, filtering out noise and irrelevant words. Stemming, a technique used to reduce words to their root forms, further streamlined the analysis process.

Once the sentences were cleansed and prepared, they underwent Part-of-Speech (POS) tagging. In natural language processing, POS tagging is a prevalent technique used to determine the role of words in sentences and categorize them based on their parts of speech. This process organizes words into distinct categories such as nouns, pronouns, verbs, adjectives, etc.

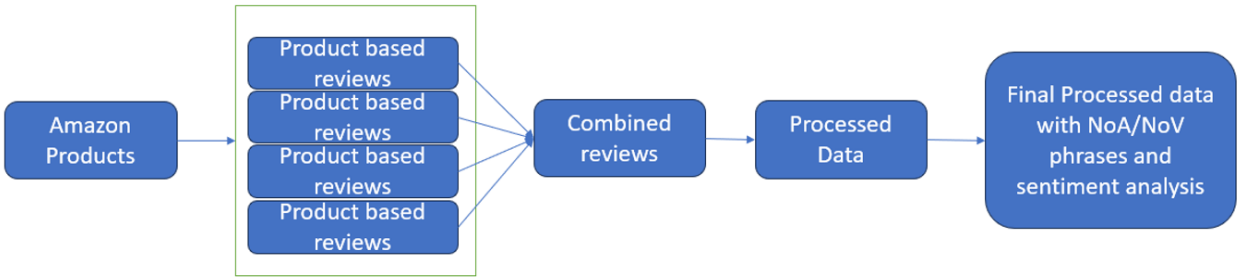
The POS tagger played a pivotal role in opinion mining due to several reasons. Firstly, it facilitated the elimination of non-opinionated words, such as nouns and pronouns, which typically lack sentiments regarding the product. Secondly, it enabled the differentiation of words used in various parts of speech, allowing for a more nuanced understanding of the language used in the reviews.

By combining sentiment sentence extraction with rigorous preprocessing and POS tagging, the analysis aimed to uncover and comprehend the sentiments expressed in product reviews, ensuring a comprehensive understanding of the opinions shared by customers. This multi-step approach facilitated the effective extraction and categorization of sentiment-bearing words within the reviews, enriching the sentiment analysis process.

Certainly! Here's the content for the "Negative Phrase



(a) Model Workflow.



(b) Data Processing Workflow.

Figure 1. Workflows.

Identification" section along with the pseudo code for the algorithm:

—

6.2. Negative Phrase Identification

In sentiment analysis of product reviews, identifying negative phrases is crucial to comprehend the polarity of customer opinions accurately. Negative phrase identification involves recognizing phrases that convey negative sentiments despite the presence of negations or modifiers that alter their meaning.

6.3. Algorithm Description

Algorithm 1 outlines the process of identifying negative polarities in text data using Part-of-Speech (POS) tagging. The algorithm detects two types of phrases: negation-of-verb (NOV) and negation-of-adjective (NOA). It relies on the presence of negative prefixes in sentences to infer negation and identifies negative phrases based on the associated adjectives or verbs.

The algorithm systematically scans through tagged sentences and identifies instances where negative prefixes are followed by adjectives or verbs, signifying negation-of-adjective (NOA) or negation-of-verb (NOV) phrases, respectively. These phrases are crucial in capturing the nuanced negative sentiments expressed by users in reviews.

The identification of NOA and NOV phrases assists in recognizing subtle negative expressions, enhancing the accuracy of sentiment analysis by understanding the modified or negated meanings of words in the context of product reviews.

6.4. Support Vector Machine

For a linearly separable dataset, the decision boundary is represented as a hyperplane given by:

$$w^T x + b = 0$$

where: - w is the weight vector perpendicular to the hyperplane. - x is the input vector. - b is the bias term.

For nonlinearly separable datasets, SVMs use kernel functions to map the input space into a higher-dimensional space, allowing for the creation of nonlinear decision boundaries. The decision function for nonlinear SVM is given by:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b\right)$$

Algorithm 1 Negative Polarities Identification

Require: Tagged Sentences, Negative Prefixes
Ensure: NOA Phrases, NOV Phrases

```
for each Sentence in Tagged Sentences do
    Tokenize the Sentence into words
    Apply Part-of-Speech (POS) Tagging to identify word categories
    for each Word in Sentence: do
        Get the POS Tag of the Next Word
        if Word is a Negative Prefix: then
            if Next Word is an Adjective: then
                Capture the NOA Phrase
            else if Next Word is a Verb: then
                Capture the NOV Phrase
            end if
            if Next Two Words form an extended pattern: then
                Save the Extended NOA and NOV Phrases
            end if
        end if
    end for
end for
Return NOA Phrases, NOV Phrases
```

where: - N is the number of support vectors. - α_i is the Lagrange multiplier. - y_i is the class label. - x_i is the support vector. - $K(x, x_i)$ is the kernel function.

Common kernel functions include: - **Linear kernel:** $K(x, x_i) = x^T x_i$ - **Polynomial kernel:** $K(x, x_i) = (x^T x_i + c)^d$ - **Gaussian RBF kernel:** $K(x, x_i) = \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)$ - **Sigmoid kernel:** $K(x, x_i) = \tanh(\alpha x^T x_i + c)$

The hyperparameters such as the penalty parameter C and kernel-specific parameters are optimized through cross-validation to achieve better generalization. SVMs aim to find the optimal hyperplane that separates classes while maximizing the margin, which is the distance between the hyperplane and the support vectors. This process involves solving the optimization problem with constraints, often formulated as a quadratic programming problem.

7. Experimental Results

We utilized Support Vector Machines (SVM) for sentiment analysis on this dataset. The SVM model was trained using 80% of the dataset, with the remaining 20% reserved for testing.

The model outputs the best product from the collected data with the most positive sentiment. It also analyses the total number of positive , neutral and negative sentiments.

The user can choose the product according to the outcome.

7.1. Accuracy & performance metrics

From the provided output of various evaluation metrics and the confusion matrix, we can interpret the following insights:

1. **Accuracy:**The model achieved an accuracy of approximately 85 - 91.3%, signifying the overall correctness in classifying sentiments across all classes.

2. **Precision:**

   (a) - **Micro-average Precision:** It represents the precision calculated globally by considering the total true positive, false positive, and false negative values across all classes. In this case, the micro-precision is also 91.3%.

   (b) **Macro-average Precision:** It computes the average precision for each class without considering class imbalance. The macro-precision is approximately 96.1%, suggesting good precision across classes.

   (c) **Weighted-average Precision:** It calculates the precision for each class, considering the number of samples in each class. The weighted precision here is around 92.3%.

   (d) **Class-specific Precision:** It indicates precision values for each class separately. The classes have precision values of 100%, 100%, and approximately 88.2%, respectively.

3. **Recall:**

   (a) **Micro-average Recall:** It denotes the recall calculated globally, considering total true positive, false positive, and false negative values across all classes. The micro-recall achieved is 91.3%.

   (b) **Macro-average Recall:** It computes the average recall for each class without considering class imbalance. The macro-recall is approximately 77.8%, indicating some variability in class-specific recall values.

   (c) **Weighted-average Recall:** It calculates the recall for each class, considering the number of samples in each class. The weighted recall here is 91.3%.

   (d) **Class-specific Recall:** The recall values for classes are approximately 83.3%, 50%, and 100%, respectively.

4. **F1 Score:**

(a) **Micro-average F1 Score:** It represents the harmonic mean of precision and recall calculated globally across all classes. The micro-F1 score achieved is 91.3%.

(b) **Macro-average F1 Score:** It calculates the average F1 score for each class without considering class imbalance. The macro-F1 score is approximately 83.8%.

(c) **Weighted-average F1 Score:** It computes the F1 score for each class, considering the number of samples in each class. The weighted F1 score here is around 90.7%.

(d) **Class-specific F1 Score:** The F1 score values for each class are approximately 90.9%, 66.7%, and 93.8%, respectively.

5. **Confusion Matrix:** The confusion matrix displays the model's classification results. It indicates that:

   (a) Class 0: 5 samples correctly predicted, 0 samples incorrectly predicted as other classes, and 1 sample misclassified.

   (b) Class 1: 1 sample correctly predicted for this class, 0 samples incorrectly predicted as other classes, and 1 sample misclassified.

   (c) Class 2: 15 samples correctly predicted for this class, with no misclassifications.

Summary: The model demonstrates high accuracy, especially in correctly classifying samples for Class 2, while exhibiting some misclassifications for Classes 0 and 1. Further analysis and potentially fine-tuning the model could be beneficial to address misclassifications and improve performance, especially for Classes 0 and 1.

8. Conclusions & Future developments

The accuracy of our model turns out to be 85-90% which is very good for the purpose.

There are other models as well, which have an accuracy of more than 95

Further, this model can be extended to compare products between various e-commerce platforms, such as comparison between Flipkart and amazon, which can help consumers in a great way. Also, the accuracy should be improved slightly by choosing a better model.

Note: Accessing/Scraping e-commerce data may lead to the blocking of your IP address by the platform due to huge traffic in less time, So please take care of that while collecting the data.

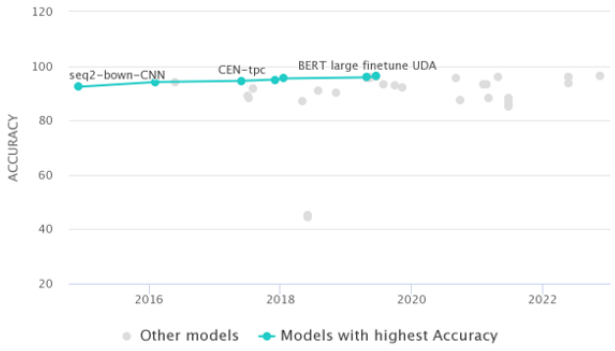


Figure 2. https://paperswithcode.com/sota/sentiment-analysis-on-imdb

References

[1] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," 2020. 1

[2] D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Revisiting lstm networks for semi-supervised text classification via mixed objective function," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 6940–6948, July 2019. 2

[3] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2020. 2

[4] J. Jabbar, I. Urooj, W. JunSheng, and N. Azeem, "Real-time sentiment analysis on e-commerce application," in *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 391–396, 2019. 2

9. Other References

1. https://www.geeksforgeeks.org/web-scraping-amazon-customer-reviews/

2. https://www.geeksforgeeks.org/scraping-amazon-product-information-using-beautiful-soup/

3. https://www.datasciencecentral.com/how-to-scrape-amazon-product-data/