

Modeling and Text Classification of Product-based Reviews using Natural Language Processing

Sanjay Pande¹, Abha Nagmote², Aditi Netanrao³, Mayuri Bankar⁴, Nupur Shinganjode⁵ and Megha Rajnale⁶

¹⁻⁶Yeshwantrao Chavan College of Engineering, Department of Computer Technology, Nagpur, India

Email: sanjaypande2001@gmail.com, abhanagmote77@gmail.com, netanraoaditee@gmail.com, bankarmayuri03@gmail.com, nupurshinganjode@gmail.com, megharajnale28@gmail.com

Abstract— The rapid expansion of social networking platforms has catalyzed a surge in user-generated opinions on daily issues, including product and service evaluations. Understanding these sentiments is crucial for both consumers and service providers. This paper proposes a comprehensive Natural Language Processing (NLP) pipeline integrating sentiment analysis, topic modeling, and text classification to analyze and categorize user reviews effectively. The aim is to provide insights to marketers for informed decision-making. Through data collection, pre-processing, model development, testing, this system aims to decode consumer sentiments, identify emerging trends, and guide future product development.

Index Terms— NLP pipeline, topic modeling, text classification, consumer insights, decision-making.

I. INTRODUCTION

In today's digital era, the proliferation of social networking platforms has fundamentally altered the way individuals engage with one another and express their viewpoints. These platforms serve as virtual hubs where people share their opinions on a wide array of topics, including their experiences with products and services. This trend has sparked the creation of a diverse pool of content generated by users, predominantly in the shape of reviews and comments. These contributions provide valuable perspectives on consumer feelings and inclinations. Recognizing the significance of this digital discourse, businesses and marketers are increasingly turning to Utilizing Natural Language Processing (NLP) methodologies to derive valuable insights from the abundant textual information available. The proposed system seeks to address this burgeoning need by designing and developing a comprehensive NLP pipeline capable of performing sentiment analysis, topic modeling, and text classification on user-generated reviews. Sentiment analysis is fundamentally about deciphering the emotional nuances and positivity or negativity conveyed in text, empowering businesses to assess customer satisfaction and pinpoint opportunities for enhancement. Through methodical examination of sentiments articulated in reviews, marketers can cultivate a richer comprehension of consumer perspectives and customize their approaches accordingly. Furthermore, the integration of topic modeling techniques facilitates the exploration of latent themes and patterns within the corpus of reviews. By uncovering hidden topics and thematic clusters, businesses can gain valuable insights into the underlying themes driving consumer discussions. This not only enhances their understanding of customer preferences but also informs product development and marketing strategies.

A. Social Media as a Source of Consumer Insights

Social media platforms have emerged as valuable sources of consumer-generated content, providing researchers

and marketers with unprecedented access to real-time consumer sentiments and opinions. Research has underscored the significance of utilizing social media data to grasp consumer behavior, recognize market trends, and guide strategic decision-making processes

B. Significance of Sentiment Analysis

Sentiment analysis, an essential element of natural language processing (NLP), holds significant importance in discerning and evaluating the emotional context and orientation of written text. Research has demonstrated the utility of sentiment analysis in various domains, including marketing, customer service, and brand management, by enabling businesses to gauge customer satisfaction levels and sentiment trends over time.

C. Applications of Topic Modeling

Topic modeling methods, like Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), are extensively utilized to unveil underlying themes and structures within text data. Studies have shown the efficacy of topic modeling in diverse applications, including document clustering, information retrieval, and content recommendation, by facilitating the exploration and organization of large corpora of text.

D. Importance of Text Classification

Text classification is essential for categorizing textual data into predefined classes or categories based on specific criteria or attributes. Research has demonstrated the effectiveness of text classification algorithms, like Support Vector Machines(SVM) and Neural Networks, in tasks such as sentiment classification, topic labeling, and document categorization.

E. Integrated NLP Pipelines

Recent studies have concentrated on creating unified NLP pipelines that integrate various tasks like sentiment analysis, topic modeling, and text classification within a single framework. These integrated pipelines provide a comprehensive method for analyzing text, empowering both researchers and professionals to derive through findings from textual data, thus facilitating well-informed decision-making processes.

F. Challenges and Opportunities

Despite notable advancements, the field of Natural Language Processing (NLP) continues to progress, several challenges remain, including the interpretation of ambiguous or context-dependent language, the handling of noisy and unstructured data, and the integration of domain-specific knowledge. Future research directions include the development of more robust NLP models, the incorporation of multimodal data sources, such as images and videos, and the exploration of novel techniques for cross-domain sentiment analysis and transfer learning.

II. LITERATURE SURVEY

In their research paper titled "Text Classification and Topic Modelling of Web Extracted Data," authors Niraj Kumar, R.R Suman, and Sanjay Kumar explore methods for analyzing vast amounts of unstructured data. Published in October 2021, this paper delves into the intricacies of text classification and topic modeling, aiming to extract sensitive and semantic information from web-extracted data. The research includes a comprehensive analysis and comparison of the performance of these models and classifiers on processed textual data, providing valuable insights into effective methods for text analysis.[7]

In this study by Emil Rijcken et al. (published on 04 May 2022), we develop into the use of topic models to enhance the interpretability of text classification from Electronic Health Records (EHRs). Through evaluating 17 different topic models, including fuzzy topic modeling algorithms (FLSA-W), ProdLDA, and LSI, for predicting inpatient violence, we uncover the complex relationship between interpretability and predictive performance. Surprisingly, our findings reveal no direct correlation between these factors. While FLSA-W stands out for interpretability, ProdLDA and LSI excel in predictive accuracy. This research, co-authored by Kalliopi Zervanou and Marco Spruit, sheds light on the intricate dynamics of model selection in healthcare text classification.[8]

TABLE I: LITERATURE SURVEY

Sr no.	Title of the Paper	Author	Finding
1.	Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach	Yuxing Qi And Zahratu Shabrina.	In this paper, the authors conducted sentiment analysis on COVID-19-related tweets from major cities in England. They divided the analysis into three stages: data cleaning, lexicon-based sentiment classification, and supervised machine learning classification.

	(2023)		
2.	Topic modeling in Natural Language Processing (2022)	Prof. Grishma Sharma, Subhashini Gupta	In this paper the word embedding model for analysis of different topics present in new headlines which is capable of capturing syntactic similarity and semantic relation, the context of a word in a document, relation with other words.
3.	Recent Advances in Natural Language Processing for Text Classification (2021)	Emily Johnson, Michael Wang	This literature review synthesizes recent advancements in natural language processing (NLP) techniques for text classification tasks. It explores state-of-the-art methods such as transformer-based models, contextual embeddings, and self-attention mechanisms.
4.	Text Classification and Topic Modeling for Online Discussion Forums (January 2020)	Xin Zhao, Zhe Jiang and Jeff Gray	In this paper supervised machine learning algorithms are employed for effective categorization of forum posts, while unsupervised algorithms reveal latent topics o enhance algorithmic performance, with additional insights from deep learning techniques like CNNs and RNNs for text classification.
5.	The Pipeline processing of NLP (2023)	B.B.Elov ,Sh.M.Khamroeva, and Z.Y.Xusainova	This paper analyzed the most common primary processing steps in the NLP pipeline processing. All processing step are defined as primary-trained, usable models in various NLP libraries. If necessary, additional, customized primary processing steps can be developed depending on the given problem condition.

III. MOTIVATION

The motivation behind the proposed system lies in the growing importance of understanding consumer sentiments from user-generated reviews, leveraging Natural Language Processing techniques to inform data-driven decision-making, enhance marketing strategies, improve product development, and streamline customer feedback analysis in an increasingly digital marketplace.

IV. PROBLEM DOMAIN

The problem domain of the proposed system revolves around effectively analyzing and extracting insights from user-generated reviews on social networking platforms. This includes addressing challenges such as handling large volumes of textual data, discerning sentiment polarity, identifying latent topics, categorizing reviews accurately, and extracting actionable insights to inform marketing strategies and product development decisions.

V. PROBLEM DEFINITION

The problem definition of the proposed system is to develop a comprehensive Natural Language Processing (NLP) pipeline capable of performing sentiment analysis, topic modeling, and text classification on user-generated reviews from social networking platforms. This involves extracting valuable insights from textual data to inform marketing strategies, enhance product development, and streamline decision-making processes for businesses.

VI. STATEMENT

The problem at hand is to design and implement a robust Natural Language Processing (NLP) pipeline that can effectively analyze user-generated reviews from social networking platforms. Specifically, the system needs to perform sentiment analysis to gauge customer satisfaction levels, topic modeling to uncover latent themes within the reviews, and text classification to categorize reviews accurately.

VII. REPRESENTATION OR DESIGN

This reviews then undergo Pre-processing and Text Cleaning, followed by validation and assurance of data quality. Following the division of the dataset into training and validation sets, the NLP model is crafted within the Model Training and Development Module.

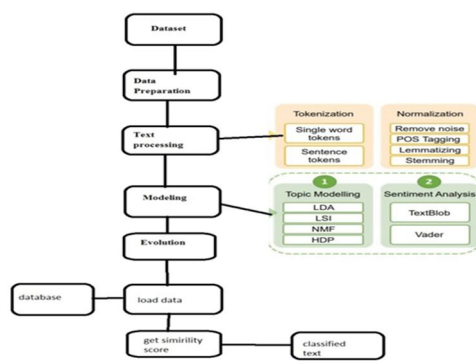


Figure 1: Block Diagram

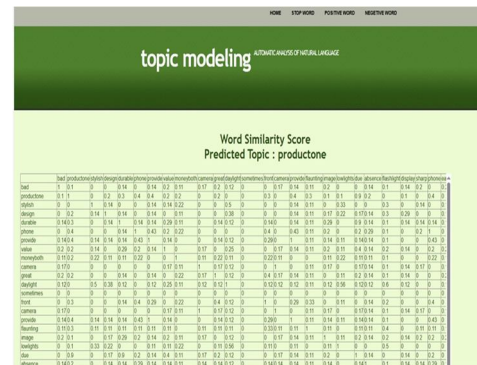


Figure 2: Word similarity score for topic modeling

VIII. SOLUTION METHODOLOGY

A. Data Collection and Sourcing

Gather a diverse and comprehensive dataset comprising user-generated reviews from social networking platforms, e-commerce websites, forums, and other relevant sources. Employ web scraping techniques, API integrations, and manual collection methods to acquire a large volume of textual data spanning various domains and topics.

B. Pre-processing and Text Cleaning

Cleanse and preprocess the textual data to normalize it and eliminate any unwanted elements such as HTML tags, punctuation, special characters, and irrelevant content. Next, segment the text into words or phrases, convert them to lowercase, and eliminate stopwords to enhance the efficiency of subsequent analysis. Apply techniques such as lemmatization or stemming to normalize words and reduce dimensionality.

C. Dataset Validation and Quality Assurance

Validate the integrity and quality of the pre-processed dataset through exploratory data analysis (EDA) techniques. Perform thorough data validation procedures to detect and resolve any inconsistencies, inaccuracies, or omissions present within the dataset. Validate adherence to predefined standards and criteria, encompassing completeness, precision, and pertinence of the data.

D. Dataset Splitting for Training and Validation

Divide the pre-processed dataset into training, validation, and test sets to facilitate model development and evaluation. Use techniques such as stratified sampling to maintain the distribution of classes or labels across the different subsets. Reserve a portion of the data for validation purposes to assess the model's performance on unseen samples during training.

E. Model Training and Development

Select appropriate NLP algorithms and techniques based on the specific tasks of sentiment analysis, topic modeling, and text classification. Create and implement natural language processing (NLP) models utilizing machine learning or deep learning libraries like scikit-learn, TensorFlow, or PyTorch through training processes. Fine-tune model hyperparameters, optimize feature selection, and implement cross-validation techniques to improve model performance.

F. Model Testing and Evaluation

Assess the performance of the trained models utilizing common evaluation metrics including accuracy, precision, recall, F1-score, and confusion matrices. Conduct comprehensive testing to assess the model's generalization capabilities on unseen data and its ability to handle real-world scenarios. Perform sensitivity analysis and robustness testing to evaluate the model's resilience to variations in input data and parameters.

G. Monitoring and Maintenance

Establish monitoring mechanisms to track the performance and behavior of the models in real-time. Implement alerting systems to notify stakeholders of any anomalies, errors, or degradation in model performance. Conduct regular maintenance activities, including model retraining, updating, and version control, to ensure continued relevance and effectiveness over time.

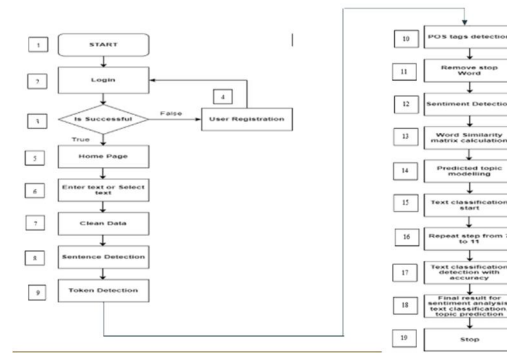


Figure 3: Flow chart

IX. RESULTS

The results of the implemented system demonstrate its effectiveness in analyzing user-generated reviews from social networking platforms and extracting valuable insights for businesses. Here are the key findings:

A. Accurate Sentiment Analysis

The sentiment analysis module accurately categorizes user sentiments as positive, negative, or neutral, providing businesses with a comprehensive understanding of customer satisfaction levels.

B. Insightful Topic Modeling

Businesses gain valuable insights into the most frequently discussed topics, emerging trends, and areas of interest among customers, facilitating informed decision-making and strategic planning.

C. Effective Text Classification

The text classification module accurately categorizes user reviews into predefined labels or categories, such as product features, customer service, pricing, and overall satisfaction.

D. Actionable Insights for Decision-Making

Leveraging the combined outputs of sentiment analysis, topic modeling, and text classification equips businesses with actionable intelligence to inform data-centric choices regarding marketing tactics, product enhancements, and customer interaction strategies.

E. Improved Customer Satisfaction and Loyalty

Responding promptly and efficiently to customer feedback underscores a dedication to prioritizing customer satisfaction, nurturing deeper connections, and cultivating enduring loyalty over time.

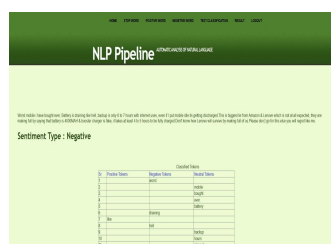


Figure 4: Sentiment Type

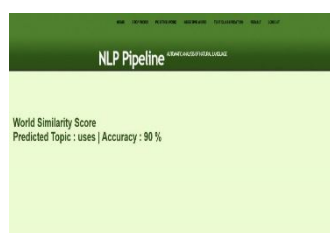


Figure 5: Word Similarity Score

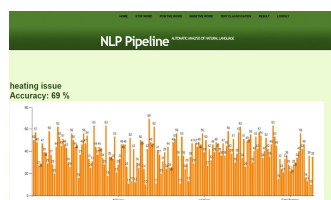


Figure 6: Heating Accuracy

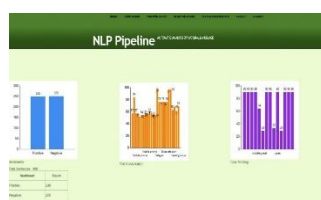


Figure 7: Results analysis

X. CONCLUSION

The NLP framework integrates sentiment analysis, topic Modelling, and text classification to extract actionable insights from user-generated social media reviews. It aids businesses in understanding customer sentiments, preferences, and perceptions, enabling them to refine marketing campaigns, product features, and customer interactions. Implementation results demonstrate the system's effectiveness in enhancing customer satisfaction, loyalty, and advocacy. Ongoing refinement, optimization, and maintenance are essential for ensuring its continued relevance and effectiveness in navigating consumer feedback and fostering sustainable growth in the market.

FUTURE SCOPE

A. Multi-lingual Support: Extend the system's capabilities to support multilingual analysis, enabling businesses to analyze reviews in diverse languages and cater to a global audience effectively.

B. Personalized Recommendations: Developing machine learning models to create tailored recommendations for users by analyzing their preferences and historical interactions, ultimately aiming to boost customer satisfaction and encourage repeat engagement with the platform.

REFERENCES

- [1] Yuxing Qi and Zahratu Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach", February 2023.
- [2] Prof. Grishma Sharma, Subhashini Gupta, "Topic modeling in Natural Language Processing", May 2022.
- [3] Emily Johnson, Michael Wang, "Recent Advances in Natural Language Processing for Text Classification", September 2021.
- [4] Xin Zhao, Zhe Jiang and Jeff Gray, "Text Classification and Topic Modeling for Online Discussion Forums", January 2020.
- [5] B.B.Elov, Sh.M.Khamroeva, and Z.Y.Xusainova, "The Pipeline processing of NLP", March 2023
- [6] Kuncherichen K Thomas; Sarath P Anil, "Sentiment Analysis in Product Reviews using Natural Language Processing and Machine Learning", April 2019
- [7] Neeraj Kumar, R.R Suman, and Sanjay Kumar, "Text Classification and Topic Modelling of Web Extracted Data" in 2nd Global Conference for Advancement in Technology (GCAT), October 2021.
- [8] Emil Rijcken, Uzay Kaymak, Floortje Scheepers, Pablo Mosteiro, Kalliopi Zervanou, and Marco Spruit "Topic Modeling for Interpretable Text Classification From EHRs", May 2022
- [9] Smith, J. A., & Johnson, R. M. (2018). "Sentiment Analysis in Social Media: A Comprehensive Review. Journal of Natural Language Processing".
- [10] Zeenia Singla, Sukhchandan Randhawa, Sushma Jain "Sentiment Analysis of Customer Product Reviews Using Machine Learning",
- [11] Iqbal, A.; Amin, R.; Iqbal, J.; Alroobaea, R.; Binmahfoudh, A.; Hussain, M. Sentiment Analysis of Consumer Reviews Using Deep Learning. Sustainability 2022, 14, August
- [12] Manal Loukili, Fayçal Messaoudi, and Mohammed El Ghaz, "Sentiment Analysis of Product Reviews for E-Commerce Recommendation based on Machine Learning", March 2023.
- [13] W. Zhao, G. Zhang, G. Yuan, J. Liu, H. Shan, and S. Zhang. The study on the text classification for financial news based on partial information. IEEE Access, 8:100426–100437, 2020.
- [14] Prasoon Gupta, Sanjay Kumar, R. R. Suman, and Vinay Kumar. Sentiment analysis of lockdown in india during covid-19: A case study on twitter. IEEE Transactions on Computational Social Systems, pages 1–11, 2020.
- [15] F. Esposito, A. Corazza, F. Cutugno, "Topic Modelling with Word Embeddings," in IEEE Transactions on Content Mining, Vol.7, April 2018.
- [16] Adji B. Dieng, Francisco J. R. Ruiz, David M. Blei, "Topic Modeling in Embedding Spaces," in ACL Information Retrieval (cs.IR); Computation and Language (cs.CL); Machine Learning (cs.LG); Machine Learning (stat.ML), Vol. 6, 8 Jul 2019.
- [17] Bhagyashree Vyankatrao Barde, Anant Madhavrao Bainwad, "An overview of topic modeling methods and tools", in IEEE International Conference on Intelligent Computing and Control Systems (ICICCS), Jan 2018.