Sentiment Analysis Using Yelp Review

Mohamed Ahmed*,Bahaa Eldin Moustafa*,Abdelrahman Elatrozy*, Mohamed Abdelnasser*,Ahmed Amr*,Karim Abdelkarim
{moham.mohamed,b.moustafa, A.Elatrozy, Ah.Helal, M.Abdelnassersenosy, k.moustafa}@nu.edu.eg

***Abstract*—In this paper, we look at Yelp reviews to understand people's opinions using a technique called sentiment analysis. We used PySpark, a tool for handling big data, to clean and prepare the reviews. Then, we applied three methods – Logistic Regression, Decision Tree, and K-Means Clustering – to analyze the sentiments in the reviews. Our goal was to find out which method works best for understanding what people feel or think based on their reviews. This study is helpful because it shows how we can use different computer techniques to analyze lots of reviews quickly and find out useful information from them.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The aim of this paper is to explore how we can understand people's opinions from Yelp reviews. Yelp is a popular website where people write reviews about businesses like restaurants and services. These reviews are very helpful because they tell us what people think and feel about these businesses. However, there are so many reviews that it's hard to read them all. This is where our project comes in.

We use a computer method called sentiment analysis to automatically find out if a review is positive, negative, or neutral. To do this, we first need to clean up the reviews using PySpark. PySpark is a powerful tool that can handle lots of data quickly and efficiently.

After cleaning the data, we use three different techniques to analyze the sentiments: Logistic Regression, Decision Tree, and K-Means Clustering. Each of these methods has its own way of understanding the reviews. By comparing these methods, we can find out which one is the best for this kind of work.

Our project is important because it shows how we can use technology to quickly understand large amounts of data, like Yelp reviews. This can help businesses know what their customers think and can also help customers make better choices.

II. Related Work

Sentiment Analysis: Predicting Yelp Scores: This study focuses on predicting the sentiment of restaurant reviews using a subset of the Yelp Open Dataset. It explores deep learning techniques like Hierarchical Attention Network (HAN) and Bidirectional Encoder Representations from Transformers (BERT) to analyze the text structure and sentiment.

Topic Modeling and Sentiment Analysis of Yelp Restaurant Reviews: This research employs a four-phase model, including data extraction and cleaning, with topic modeling using Latent Dirichlet Allocation (LDA) to identify key themes in Yelp restaurant reviews.

Sentiment Analysis of Yelp Reviews: A Comparison of Techniques: This paper analyzes over 350,000 Yelp reviews to compare various text preprocessing techniques and the effectiveness of machine learning models in predicting user sentiment.

Sentiment Classification and Aspect-Based Sentiment Analysis on Yelp Reviews Using Deep Learning: This study uses deep learning and word embeddings for sentiment classification and aspect-based analysis of Yelp reviews.

Sentiment Analysis of Yelp Reviews by Machine Learning: This research analyzes Yelp reviews to assign probabilities for positive or negative sentiment, focusing on aspects like food, service, price, and ambiance.

III. Methodology

1) Project Plan: Established milestones include data acquisition, preprocessing, model development, testing, and evaluation. Timelines are allocated for each milestone, and resources are designated for data handling and team roles.
2) System Architecture: Utilizes a layered architecture with separate components for data processing, analysis, and visualization. Chose PySpark for efficient data processing, machine learning libraries for analysis, and modern visualization tools for insights presentation.
3) Data Handling:
   - Acquisition: Sourced Yelp reviews dataset, focusing on a comprehensive variety of reviews.
   - Storage and Processing: Employed scalable storage solutions; data processing involved transforming raw data into a structured format.
   - Visualization and Security: Utilized intuitive tools for data visualization. Emphasized data security and privacy throughout the process.
4) Data Preprocessing:
   - Addressed missing data through imputation techniques.
   - Implemented outlier detection methods to ensure data quality.
   - Conducted feature engineering to enhance model input.
5) Algorithm Implementation:
   - Selected Logistic Regression, Decision Tree, and K-Means Clustering for their efficacy in sentiment classification.
   - Detailed their implementation, highlighting specific configurations and optimizations used.

A. Abbreviations and Acronyms

1) SA: Sentiment Analysis - A method used to analyze and categorize opinions expressed in text data.
2) Yelp DS: Yelp Dataset - The dataset of reviews and business information from Yelp.
3) PySpark: Python API for Apache Spark - A tool used for handling large datasets and performing data cleaning and analysis.
4) LR: Logistic Regression - A statistical method used for predicting binary outcomes.
5) DT: Decision Tree - A machine learning algorithm used for classification and regression tasks.
6) KMC: K-Means Clustering - An unsupervised learning algorithm used for clustering data into groups.
7) ML: Machine Learning - The study of computer algorithms that improve automatically through experience.
8) API: Application Programming Interface - A set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service.
9) NLP: Natural Language Processing - A branch of artificial intelligence concerned with the interactions between computers and human language.
10) JSON: JavaScript Object Notation - A lightweight data-interchange format.

B. Units

1) Byte (B): A unit of digital information storage. Often used in larger multiples such as kilobyte (KB), megabyte (MB), gigabyte (GB), etc.
2) Hertz (Hz): In computing, it often refers to the speed of processors, with one hertz indicating one cycle per second. It's commonly used in gigahertz (GHz).
3) Bit (b): The most basic unit of data in computing and digital communications. Like bytes, bits are often used in larger scales, such as kilobits (Kb) or megabits (Mb).
4) Percentage (%): Used to express proportions and likelihoods in statistical analysis.

IV. Results and Discussion

Our project achieved effective sentiment analysis on the Yelp reviews dataset. We found that Logistic Regression provided a good balance of accuracy and computational efficiency. Decision Trees offered more interpretability, but were slightly less accurate. K-Means clustering revealed interesting patterns in user sentiments. The integration of PySpark greatly enhanced data handling efficiency. Ethical considerations, such as bias minimization, were a key focus. Overall, this project not only advances the understanding of consumer sentiments in online reviews but also demonstrates the practical application of machine learning in real-world data analysis.

V. Conclusion

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [**?**].

References

[1] B. P. R. Guda, M. Srivastava, and D. Karkhanis, "Sentiment Analysis: Predicting Yelp Scores," 2022. [Online]. Available: arXiv:2201.07999 [cs.LG].

[2] Z. C., "Topic Modelling and Sentiment Analysis Yelp Reviews," GitHub repository, [Online]. Available: https://github.com/Zahivc/Topic-Modelling-and-Sentiment-Analysis-Yelp-Reviews.

[3] S. Liu, "Sentiment Analysis of Yelp Reviews: A Comparison of Techniques and Models," 2020. [Online]. Available: arXiv:2004.13851 [cs.CL].

[4] E. S. Alamoudi and N. S. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," Journal of Decision Systems, vol. 30, no. 2-3, pp. 259-281, 2021. [Online]. Available: https://doi.org/10.1080/12460125.2020.1864106.

[5] H. S. and R. Ramathmika, "Sentiment Analysis of Yelp Reviews by Machine Learning," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 700-704, doi: 10.1109/ICCS45141.2019.9065812.