Project Proposal: Big Data-Intensive Emotion-Based Sentiment Analysis and Intelligent Movie Recommendations

Siddharth Bhatt Nishchay Nilabh

Abstract—**This paper introduces an innovative approach in the realm of big data analytics, targeting a deeper understanding of sentiment analysis within the context of movie reviews. Rather than adhering to the conventional binary classification methods, our methodology is designed to pinpoint and classify nuanced emotions, such as happiness, anger, and sadness, as expressed in the reviews. To handle the sheer volume and complexity of the data, big data technologies like Hadoop and PySpark are employed. The significance of this refined emotion-based classification transcends mere academic interest; it serves as the foundation for a sophisticated recommendation engine that can be integrated with IMDb. This combination ensures movie suggestions that align closely with the feelings of viewers. The end results are displayed through an interactive front-end, making the experience more immersive.**

Index Terms—**Big Data Analytics, Hadoop, PySpark, Sentiment Analysis, Movie Reviews.**

I. Introduction

SENTIMENT analysis in movie reviews has conventionally been restricted to simple binary classifications: positive or negative. While this provides a rudimentary understanding of public opinion, it falls significantly short of capturing the various other human emotions and reactions that are often expressed in textual reviews. To add to the complexity, existing methods have not been effectively scaled to handle the overwhelming volume of reviews generated daily across various platforms. For these reasons, the problem at hand can be solved through big data techniques and technologies such as Hadoop and PySpark. A very Large Movie Review dataset(link) provided by Stanford University and sourced from IMDb (Internet Movie Database) shall be used to perform the proposed research. The dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. There is unlabeled data of about 50,000 reviews in raw text and already processed bag of words formats. A number of previous investigations already exist on the dataset in question and these have been discussed in brief followed by our proposed research idea.

A. Importance

1) For Production Houses:

- **In-Depth Audience Understanding**: The entertainment industry has seen a surge in content production, making it increasingly crucial for production houses to comprehend nuanced audience feedback. Simple binary feedback fails to capture this diversity of emotions, leading to a content gap.

- **Targeted Marketing**: Understanding the emotional impact of a film could be invaluable for marketing strategies. For example, if a movie evokes strong feelings of nostalgia, marketing campaigns could be designed to highlight this.

- **Content Refinement**: Precise emotional feedback can guide in tweaking existing content or in the conceptualization of new projects.

2) For Consumers:

- **Enhanced User Experience**: While existing recommendation systems based on genres or directors can be effective to some extent, they lack the emotional granularity. If the sentiment analysis can be done on a more nuanced emotional scale, the recommendation algorithms can potentially offer titles that match the user's emotional preferences, resulting in a more personalized user experience.

- **Time Efficiency**: By receiving more accurate recommendations, users can make more informed choices without having to sift through a myriad of options, making the process time-efficient.

B. Relevance of This Project

1) Scalability: One of the primary gaps in the current research landscape is the lack of scalability in sentiment analysis models. This project aims to utilize big data technologies like Hadoop and PySpark to perform sentiment analysis on a dataset of 50,000 movie reviews. The methodologies could then be easily scaled to analyze millions of reviews.

2) Complexity: This research will move beyond binary sentiment classifications by delving into more complex, emotion-based categories such as happiness, sadness, and anger. This will add a layer of depth to the analysis which has been missing from most large-scale studies so far.

3) Real-world Applications: Beyond the immediate scope of movie recommendations, the methodologies developed in this project have potential applications in various fields that rely on public opinion or feedback, such as retail, politics, and healthcare.

4) Technological Contributions: Lastly, this project aims to make significant contributions to the field of big data analytics, as it presents a case study for handling high-volume textual data, feature extraction, and efficient data processing. The successful implementation of this project could serve as a blueprint for future research endeavors in large-scale sentiment analysis.

By addressing these elements, this project will tackle the existing shortcomings in sentiment analysis, offering a more nuanced understanding of human emotions, all while effectively dealing with large-scale data. This will not only advance the field of big data analytics but will also have immediate practical applications in improving the quality of recommendations and feedback interpretation.

II. Previous Investigation

A. Machine Learning Methods

There have been plenty of machine learning algorithms used to classify between positive and negative reviews, such as :

*1) **Naive Bayes** [1]:* It classified the text based off of a predefined set of positive and negative words. It performed best in case of an equal split of positive and negative reviews of the training data. The accuracy changed depending on the number of training samples, but lied in the range of 70 - 75% on the nltk corpus "movie reviews" dataset.

*2) **SVMs** [2]:* This showed better accuracy than Naive-Bayes on "Pang Corpus" and "Taboada Corpus" datasets. The accuracy varied from 80-93% depending on the hyperparameters. SVMs can perform binary classification by finding a hyperplane that best separates the dataset. They perform high flexibility by allowing the fine-tuning of many hyperparameters.

*3) **Random Forest Ensemble**[3]:* It had an accuracy of 97% (compared to SVMs 92%) on the reviews extracted from flipkart.com. This could be because random forests are not prone to overfit and the data need not be linearly separable. The ensembling of decision trees allows the model to learn the weights of words more accurately.

B. Deep Learning Methods

The onset of deep learning allowed the models to become more complex and greatly understand the context of words. Neural neworks have been excellent at understanding the sentiment of sentences. Some of these deep learning methods are :

*1) **CNNs**:* While Convolution Neural Networks (CNNs) are typically associated with image processing, they can also be applied to text data by treating it as a 1D sequence of words. They are effective at capturing local patterns and can be used for text classification, including sentiment analysis.

*2) **RNNS**:* Recurrent Neural Networks(RNNs) use the prior word as well as the current word embeddings to find out the context of the sentence. This allows the model to understand comments that don't directly have a negative word, but still have a negative impact. This is how RNNs are different from simple feed-forward networks. They can be further modified to be more robust (Ex : LSTM, GRU).

*3) **Word Vector Model**:* Word Vector Model uses the features of words to relate them using strings. Essentially a graph that has edges between similar words. Word2Vec (developed by Google) is a tool that gives vector embeddings of any word depending on the context. The output is an embedding matrix.

These models performed extremely well, even on datasets where Naive Bayes failed to do so. Ex [4] : On data scraped from Traveloka (Travel company) website, the accuracy of the various models were :

- RNN + Word2vec - 92%
- CNN + Word2vec - 89%
- Naive Bayes - 44%

III. Proposed Research

The primary aim of this project is to process a large dataset of nearly 50000 reviews using Big Data techniques such as Hadoop or PySpark. The methodology has been discussed to begin with.

A. *Methodology*

1) **Data pre-processing**: Hadoop’s HDFS or PySpark’s resilient distributed datasets can be used to import initially the raw format of the movie reviews and then the bag-of-words format for the same reviews. We then clean the data and normalize the text by applying various techniques such as lower-casing, removing special characters and stopwords and stemming the text. The data is partitioned after this to prepare it for sentiment analysis and distribute it into different emotional labels.

2) **Feature Extraction**: The feature extraction from the clean data is a bit more complex and the exact procedure remains a little tough to crack. After consulting various other projects, the idea is to convert the textual data into feature vectors using Term Frequency-Inverse Document Frequency(TF-IDF). Principal Component Analysis can be applied on these feature vectors to ease the computational loads without significant loss of information.

3) **Emotion Classification**: We have to use a suitable method from the previously used methods such as Random Forests, Naive Bayes or SVMs to classify the text. During the training phase, we can utilise the MapReduce architecture from Hadoop or the MLlib from PySpark to distribute the computational load. We can further apply Deep Learning techniques such as tuning our hyperparamter using a method like GridSearch or RandomizedSearch and k-fold cross-validation to validate our model’s performances. These techniques will enchace our model’s performances greatly.

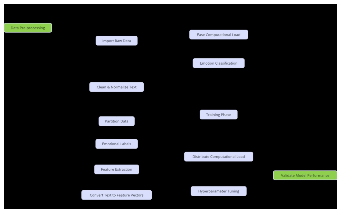

Fig. 1. Flow chart of the proposed solution

B. *Challenges*

- **Computational Complexity:** The intricate nature of text data and the inclusion of multiple emotional labels can make the classification task computationally expensive.

- **Data Sparsity:** High dimensionality resulting from TF-IDF vectorization could make the data sparse, affecting the efficiency of machine learning algorithms.

- **Integration Complexity:** Merging the emotion-based classification data into IMDb’s existing systems seamlessly will be challenging. Also, since the reviews are limited in number, it will be challenging to come up with a wider range of recommendations solely based on this dataset.

C. *Significance*

- **Real-Time Recommendations:** The proposed system would be capable of providing real-time, emotion-based movie recommendations, which is an advancement over existing, more static models. The recommendations, as mentioned earlier, shall be integrated with the IMDb, which provides a far more extensive recommendation system.

- **Extensibility:** The methodologies developed can serve as a blueprint for sentiment analysis in other sectors like retail or politics.

- **Personalization:** The IMDb-integrated recommendation engine will significantly improve the personalization and relevance of suggested movies, thereby enriching the end-user experience. The time saved in searching for relevant and personalized movie, is the end goal to top it all off.

References

[1] V. Nama, V. Hegde, and B. Satish Babu, “Sentiment analysis of movie reviews: A comparative study between the naive-bayes classifier and a rule-based approach,” in *2021 International Conference on Innovative Trends in Information Technology (ICITIIT)*, 2021, pp. 1–6.

[2] N. Zainuddin and A. Selamat, “Sentiment analysis using support vector machine,” in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 2014, pp. 333–337.

[3] P. Karthika, R. Murugeswari, and R. Manoranjithem, “Sentiment analysis of social media network using random forest algorithm,” in *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2019, pp. 1–5.

[4] L. Kurniasari and A. Setyanto, “Sentiment analysis using recurrent neural network,” *Journal of Physics: Conference Series*, vol. 1471, no. 1, p. 012018, feb 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1471/1/012018