RRS: Rating Reviews System Based on Sentiment Analysis

Saeed Al-Qarni
School of Computing and Engineering
University of Missouri - Kansas City
saacfb@mail.umkc.edu

Khalid Dhabbah
School of Computing and Engineering
University of Missouri - Kansas City
kmdk2t@mail.umkc.edu

Anshit Saxena
School of Computing and Engineering
University of Missouri - Kansas City
asm5y@mail.umkc.edu

***Abstract*—Nowadays, Social media has become popular places for sharing the best description of a movie through texts. Analyzing reviews or text helps in recognizing the sentiment stated by people. IMDb movie website has a lot of reviews that have been expressed by reviewers towards a movie. In the proposed work, we first get the data (the reviews) dynamically, do sentimental analysis on them and giving out the positive and the negative ratings. Getting positive and negative sentiment based on the reviews plays a significant role in text classification using sentiment analysis. As a result, we will see what the difference between our rating and IMDb is. Moreover, our system will help people find the easiest way to find the top words that describe a movie other than the genre. Also, the model will get an overall sentiment for all the reviews and will not be limited to reading a few of them.**

I. INTRODUCTION

Nowadays, people want to watch movies after they see the rating or read the reviews for that movie. However, people can only view 10-20 reviews at a time and the rating is based on what the registered users feel about the movie and not all the people watching it. Through sentiment analysis, people can know the top words that are used to describe a movie, if the overall sentiment is positive or negative, we can get the review ratings from above 5000 reviews. It is an easy way to describe and find the best movie instead of reading a lot of text in the reviews. In this paper, we explore ways to analyze these reviews and compare them with IMDb rating, in the context of text classification. We will grab the reviews for a movie dynamically and do the sentiment analysis. After that, we will give the rating based on that sentiment and compare it with IMDb rating. Once we get our rating, we can find the top words which describe a movie.

II. RELATED WORK

Sentiment analysis has become a significant research direction in natural language processing (NLP), which predicts people's opinion. At present, most of the research in sentiment analysis have been applied using NLP, ML, and DL.

CNN and LSTM has been used in a research [1] to enhance the accuracy and improving the overfitting. The dataset used was IMDb reviews and they used Dropout layer with max Pooling and Batch Normalization, to deal with overfitting issues, achieving best accuracy of 89.5%.

Another research with sentimental analysis on IMDb reviews [2] improved the vanishing gradient issue. The model was a mix of CNN and residual connection, and was tested on four different models, achieving the best accuracy of 90.02%.

With NLP, one can rapidly progress on text data processing, specially in sentiment analysis. On the other hand, a combination of deep learning approaches LSTM and CNN to perform analysis of movie reviews was done on the IMDb movie reviews and two models were proposed [3]. The first model was a hybrid LSTM-CNN, which performs better than the second hybrid CNN-LSTM model. The accuracy achieved was 79%. Preprocessing techniques were used to enhance the model's performance and embedded layer was used to embed the words into a vector.

There has been a research using different algorithms, such as SVM, KNN, Maximum Entropy, and Naïve Bayes [4], using (TF, TF-IDF) beside Lexicon features to get a better result. Preprocessing and feature extraction were applied to the dataset for better results.

A new way of sentiment analysis was proposed [5] by using Bayesian Rough Decision Tree, which uses the Decision Tree and Bayesian Rough set combined. By doing text preprocessing, such as tokenization and conversion to lowercase, as well as feature extraction, such as POS, they achieved about 95% accuracy.

Research has been conducted to develop a technique that can classify movie reviews [10].After classification, the results were compared with eight different classification algorithms which were NB, BN, DT, KNN, RRL, SVM, RF, and SGD classifiers.

Machine learning classification approaches, namely, Naive bays (NB) and Random forest (RF), was used for measuring positive, negative, and natural [11]. As a result of applying these two techniques, RF performed better than NB algorithm in terms of time and memory to get the best movie to users.

All research done on movie reviews used dataset which was predefined, or static (did not have the capability to add more real time data), which is shown in table 1. In this paper, we are extracting the dataset in real time (dynamic) and then performing sentimental analysis on the dataset.

III. ARCHITECTURE

We have followed several steps to implement the sentiment analysis as shown in Figure 1.

1) Collecting the Dataset

For the purpose of collecting the data from the IMDb for a particular movie, we are using the Chrome Browser. We use a driver dedicated for chrome for opening up the IMDb reviews page for that particular movie and wait for it to load all the

Paper	Dataset Used	Purpose	Methods	Focused	Accuracy
1	IMDb	Applying sentiment analysis	CNN and LSTM	Enhancing the accuracy and improving overfitting	89.5
2	IMDb	Applying sentiment analysis	A mix of CNN and residual connection	Improving the vanishing gradient	90.02
3	IMDb	Applying sentiment analysis	Two models in deep learning: 1- a hybrid LSTM-CNN. 2- a hybrid CNN-LSTM	Enhancing the model's performance and using embedded layer to embeds the words into a vector	79
4	IMDb	Using Tf, TF-IDF and lexicon features	SVM, KNN, Maximum Entropy, and Naïve Bayes	Applying preprocessing and feature extraction	75
5	IMDb and Twitter Data	Applying sentiment analysis	Bayesian Rough Decision Tree	Applying text preprocessing	95
10	IMDb	Classifying the movies	Using eight different classification algorithms	NB, BN, DT, KNN, RRL, SVM, RF, and SGD	RF performs the best, RRL performs the worst
11	Movies Reviews	Applying sentiment analysis	Navie Bays(NB) and Random Forest(Rf)	Applying these methods	RF perform better than NB

Table I: The related Work summery

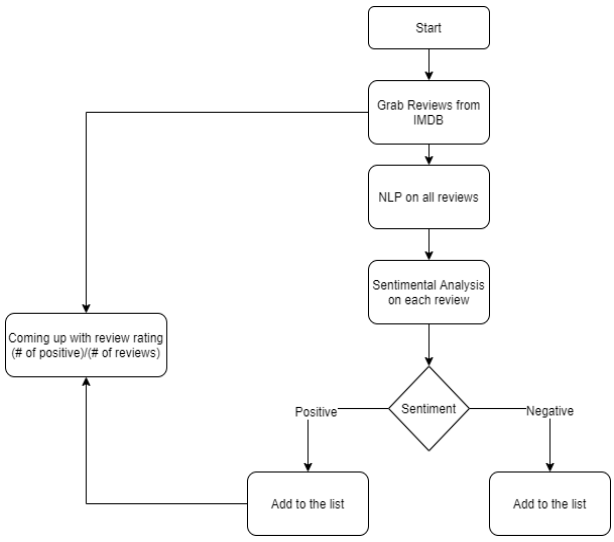

Figure 1: Flowchart for the proposed architecture

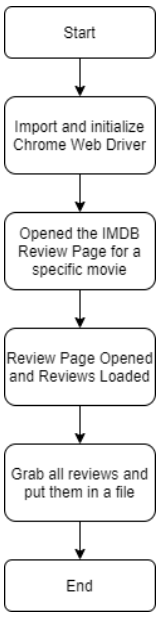

Figure 2: Flowchart for getting the IMDb dataset

reviews. After the page has loaded all the reviews, we get all the reviews using BeautifulSoup and HTTP Parser.

Figure 2 shows our approach for collecting the IMDb Dataset.

2) Performing NLP

Removing stop words Encoding the data in proper format (UTF-8) Tokenization and Capitalization/De-Capitalization Breaking attached words Lemmatization/Stemming

3) Performing Sentiment Analysis

After running sentiment analysis on each review, we classify them as positive and negative and separate them.

4) Calculating the Rating

We calculate our rating using the total number of reviews and the number of positive reviews by:

Total number of Positive reviews / Total number of reviews

IV. Implementation and Evaluation

We propose a Rating Reviews System named RRS. RRS will do sentiment analysis for all reviews, both one at a time and combined reviews for a specific movie.

The flow chart illustrated in Figure 1 is the working system architecture for RRS. RRS can aid the person who is reading the reviews to have a better insight to the movie through the reviews without reading them. It is like a review on the reviews. RRS loads the reviews from IMDb and then applies NLP on all the reviews after cleaning them. Then it will do sentiment analysis on each review as well as all reviews. For tackling sentiment analysis, there are two methods that we followed: 1- Using StanfordCoreNLP [6]. it is a tool for natural language processing to implement some functions like finding the base form of a word, sentiment analysis, etc. 2- Using Text Blob [7], which is another tool for processing text built in NLTK, a python library.

We decide to use StanfordCoreNLP for RRS since it makes it easy to implement linguistic analysis and has improved features like named entity recognition [8] and better grammatical functions.

In order to get the dataset, we are using a WebDriver for Chrome browser, in which we will open the web page with all the reviews for the respective movie and load all the reviews on the page. After loading all the reviews on the browser page, we use a python library called BeautifulSoup [9] to parse all the content to 'utf-8' format and find out how many reviews are there on the page. We write all the reviews in a file by grabbing the reviews one by one from the page. Then we start cleaning up the reviews by first removing the stop words, and

we use python core NLP from StanfordCoreNLP to do our sentimental analysis on the reviews and then segregate the reviews based on the sentiment. As a result, we are providing the negative rating and the positive rating for the movie.

To get the top words, we are just using the frequency for now to find most frequently used words in both positive and negative reviews. Then after getting the top words we are masking those top words onto images for both positive and negative reviews to make word clouds. The full source code can be found here: https://github.com/SaeedAlmohisen/KDM_Project.

A. Evaluation and Result

1) Datasets

The dataset we have is automatically loaded to the model, which means we are having the recent reviews using the Web Driver tool. This makes sure that we take in even the most recent reviews.

2) System Specifications

The RRS system can be integrated with any website. When integrated with the website, some icons can be added to apply the RRS, anyone will be able to select a movie from the list: "Movie name". They have the option to read all the reviews. If they want to have one review, they also can do that. The RRS system is applied on these reviews. The person can find options beside them. If the person who is surfing the site selects the analysis icon, he can apply RRS functionality on the review to get the sentiment analysis for it. If the person selects the same icon but the one under the movie trailer, the RRS is also applied to ask the person if he wants a sentiment analysis feature in all the review available. If the movie does not have a review, the RRS features goes hidden by hiding the icons.

3) Statistics

We have chosen a movie as a sample called the dark Knight, which was released on July 18, 2008. We collected 6664 reviews and cleaned them by doing stop word removal. Then we applied the sentiment analysis. The table 2 shows the details about this movie.

Name	∫ reviews	∫ positive
The dark knight	6664	2203
∫ Negative	**∫ Neutral**	**∫ Very Positive**
3011	479	480
∫ Very Negative	**% of all Positive**	
490	4.026714693081194	

Table II

So according to the reviews, this movie has 40% positive percentage rating and 52% negative percentage rating.

In addition, Figure 3 shows the data visualization for the result. Then we have filtered the data and tokenized the positive and negative reviews to find the top words using Word Cloud as shown in Figure 4.

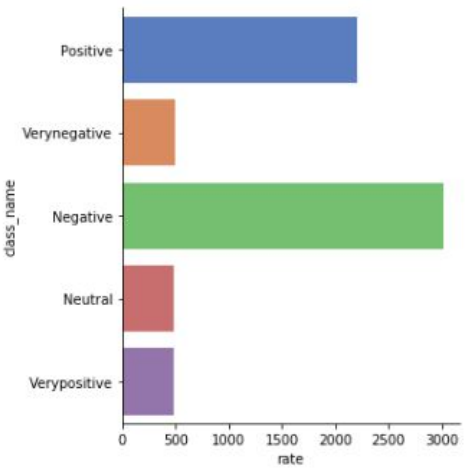

Figure 3: The sentiment analysis

Figure 4: Providing the most frequently used words

V. Conclusion

Our project extracts the movie reviews (dataset) from the IMDb website in real time, and then provides ratings for movies based on the sentiment of those reviews. This project not only gives the positive rating but also negative ratings. Also, this project provides word clouds for both positive and negative sentiment to the users for them to post on social media.

VI. Future work

We will be extending this project and will develop a web app, on which users can go and type in any movie they would like to get the rating for and also get the word clouds to share on their social media.

References

[1] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), New York, NY, 2017, pp. 540-546.

[2] N. K. Thinh, C. H. Nga, Y. Lee, M. Wu, P. Chang and J. Wang, "Sentiment Analysis Using Residual Learning with Simplified CNN Extractor," 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 2019, pp. 335-3353.

[3] A. Sajeevan and L. K. S, "An enhanced approach for movie review analysis using deep learning techniques," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019, pp. 1788-1794.

[4] Kumar, Himanshu, B. S. Harish and H. K. Darshan. "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method." IJIMAI 5 (2019): 109-114.

[5] İ.Tarımer, A. Çoban, A.E. Kocaman. "Sentiment Analysis on IMDB Movie Comments and Twitter Data by Machine Learning and Vector Space Techniques" 2019.

[6] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing-toolkit. InProceedingsof52ndAnnualMeetingof theAssociationforComputationalLinguistics: System Demonstrations, pages 55–60.

[7] TextBlob, https://textblob.readthedocs.io/en/dev/2017

[8] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian,KazuyaKawakami,andChrisDyer.2016. Neural architectures for named entity recognition. In NAACL-HLT.

[9] Beautiful Soup, https://readthedocs.org/projects/beautiful-soup-4/downloads/pdf/latest.

[10] M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 860-865.

[11] T. M. Untawale and G. Choudhari, "Implementation of Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1197-1200.