Movie Review Analysis: Impact of User and Critic Reviews on Movie Release

*Influence of Critic and User Reviews on Movie Releases

1st Madni ali Hussain
Master's in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23158859@student.ncirl.ie

2nd Gaurav
Master's in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23189487@student.ncirl.ie

3rd Diya Srivastava
Master's in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23177608@student.ncirl.ie

***Abstract*—In the Movie Industry the likes and dislikes of audience and critics are a prominent ensemble that should be taken under consideration to forecast and analyze the performance of a movie. This work is to scrutinize "If these reviews make a difference? And if so, which one's prominent? Critic Reviews or User Reviews and what affect does this have on recent movie releases, if there exists any." The human analysis, however, lacks predictive analysis and critic reception along with the accessibility of quick and precise sentiment analysis. Such challenge requires a faster and accurate way to foster results. In this work, we propose the use of Bidirectional Encoder Representations from Transformers (BERT) to perform sentiment analysis of Movie Reviews from audience and critics. This in amalgamation with the recent movie releases and in making can give us a broad horizon of Genre preference, Actor popularity and much more. We entailed Distil-BERT model from Hugging Face for Sentiment Analysis on web scraped real-world IMDB user reviews and Rotten Tomatoes Critic reviews dataset, and performed data visualization to analyse movie trends and draw a correlation between movie releases and the sentiments. The Distil-BERT model yield about 95% accurate results being a subset of the BERT transformer.**

***Index Terms*—BERT, Sentiments, transformer-model, web-scrapping, critics**

I. Introduction

The inconsistent dynamics of audience preferences hold a great power to demonstrate the growth and prosperity of a product. Opinions can be expressed in different forms. One may be websites for reviewing products, like Amazon, or fetching a thousand comments from a YouTube video, or review sites such as Rotten Tomatoes which enable rating of products on some fixed scale and leaving reviews as well. These reviews are a comprehensive description of audience's experience with a movie. Generally, little status messages and comments on social networking sites like Twitter, Instagram also says a lot about movie performances and can be monitored to analyze the trends in the market.

In the Movie Industry the likes and dislikes of audience and critics are a prominent ensemble that should be taken under consideration to forecast and analyze the performance of a movie. There is a mélange of aspects that dominate the performance of a movie in the industry, constituting of actor preference, genre amiability, plot relevance and production companies and a lot more. Therefore, this provides a rigid motive to analyze consumer preference in association with movie making and release to gain fruitful results of the audience insights. We are working to see "If these reviews make a difference? And if so, which one's prominent? Critic Reviews or User Reviews and what affect does this have on recent movie releases, if there exists any." Several fields of Artificial Intelligence (AI) have gained enormous recognition for the purpose of Sentiment Analysis of such reviews. Transformer is a neural network model whose trend is increasing. ChatGPT, which has taken the world by storm because of its general answering and questioning system capabilities, is a form of pre-trained transformer called generative pre-trained transformer (GPT) developed by OpenAI is a transformer-based language model, generating coherent text based on given prompt. Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained transformer model, pre-trained on large text corpora and is finetuned for enormous tasks like question answering, text classification, sentiment analysis.

The objective of our project is to analyse the Rotten Tomatoes Critic Reviews and IMDB User Reviews and combine them with the released movie and find out the correlation between these review, ratings and the amount of movies released, to find out If and How they affect the recent movie releases and In making. For datasets, we have [4] Rotten Tomatoes Critic Reviews, which contain comprehensive reviews by the critics along with rotten movie links, actors, genres and date of reviews etc. Next dataset is [2] web scraped [14] from IMDB, this has the user reviews with actors, genres and date of reviews etc. Finally, we combine this with a Movies metadata [3], that has the movie title, release date, actors, genres and status of movie. We have first saved the data in MongoDB database using Nosql as per the requirement of our project and data. We perform Sentiment Analysis on Rotten Tomatoes Critic Review dataset and on the movie review data web scraped from IMDB User Reviews leveraging Distil

BERT from Hugging Face, this is a pre-trained model. This is a pre-trained model that yields Sentiment label- 0 and 1 (0 for negative and 1 for positive review classification) and a Confidence Score (range [0-1] higher score, closer to 1 indicates high confidence in label classification). This analysis is then anchored to further study, analyse and visualize the influence of these reviews on movie release and production trends. The Distil-BERT model yields about 95-97% of BERT's performance while being 40% smaller 60% faster.

II. Related Work

A. SVM and MNB

"Support Vector Machine" is a machine learning model that classifies some given data into pre-defined categories, and thus it is a supervised learning method. "Multinomial Naïve Byes" MNB is also a popular algorithm for text classification. There have been researches that used the respective machine learning models for task of sentiment analysis of Rotten Tomato Reviews. [8], [9] are the works that have introduced several methods for sentiment analysis and have compared their results. The data is prepared initially performing, Tokenization, Normalization and Noise Removal of the textual data to finally serve to the model. Constructed on the Bayes Theorem and the simplistic independent hypotheses across features, the Naive Bayes (NB) classifier is a straightforward stochastic-based supervised machine learning classifier. The NB classifier utilizes the independent feature assumption to calculate the posterior probability of a specific category based on the given text data in sentiment classification. To achieve higher levels of efficiency and generalization capacity than the individual basis learners, the ensemble learning approach trains many base learners and aggregates their predictions. Both SVM and MNB outperformed deep learning techniques in [8], achieving relatively low accuracy 58.6%. Although the model was slightly bent towards predicting number of positives more than the number of negatives, but was fine-tuned to achieve accuracies of 76.1% for MNB and 74.7% for SVM.

B. Convolutional Neural Network

Leveraging CNN can be very fruitful for sentiment analysis, this requires construction of a neural network architecture that can be trained on textual data and perform classification. For deep learning, 1D convolutional neural network (CNN), 2D-CNN, stacked long short-term memory (LSTM), and Bidirectional LSTM algorithms have been utilized. Shaung et al. [11] proposed an approach using CNN-LSTM to detect polarity in English and Chinese product reviews. Their proposed CNN-LSTM model outperformed individual CNN and LSTM classifiers and achieved 81.86% accuracy. The CNN classifier consists of input, output, and hidden layers. The hidden layers constitute of convolutional, pooling, fully connected and normalization layers [11]. In experiments, the best results were obtained using an 11-layered CNN architecture. The results demonstrate that the stacked-BiLSTM model outperformed all other methods. In addition, it is observed that CNN and LSTM classifiers can effectively detect the polarity of the movie reviews. Furthermore, they also help in detecting contextual information as compared to traditional classifiers. It is shown that deep learning approaches are more optimal for sentimental analysis due to their having less overfitting and better generalization. stacked-bidirectional-LSTM achieved the highest accuracy, of up to 95.61%, for the movie dataset.

C. BERT, RoBERTa, DeBERTa

Learning from the outcomes of the SVM and CNN trained models, this part explores the use of some BERT pre-trained models. According to (Liu et al., 2019),the BERT model is undertrained, and an upgraded model termed the RoBERTa model has been developed (Robustly optimized BERT), this outperforms the BERT model because it was pre-trained effectively using better pre-training procedures. The RoBERTa model has proved to be trained for 40 epoch layers and takes about 160 gb of data to pre-train, in light of this it's a great model to use for text generation and works like categorizing of spamming. DeBERTa (Decoding-enhanced BERT) with disentangled attention on the other hand is a powerful new strategy that has been put out (He et al., 2020). On Considering the tasks involving natural language processing, this model outperforms the RoBERTa and BERT model. It is pre-trained is pre-trained utilizing masked language modeling(MLM). the DeBERTa model represents the position and the content using two vectors. To get the weight for these vectors, the disentangled matrix is used. This demonstrates how the word's position and meaning are taken into account when determining the attention weight. It offers the absolute word position embedding technique to decode the masked word and is fine-tuned to normalize the word embeddings and transform them into a stochastic vector is called scale-invariant fine tuning (SiFT). From [9], the DeBERTa model with 2 epochs has achieved an accuracy of 93% and is considered the best model for this dataset. Hence, helping us choose the appropriate transformer model for our task, i.e. Distil-BERT which is a smaller sized, with reduced computation requirements provides good results where language and nuances are less critical, it offers a good balance between efficiency and performance.

III. Methodology

A. Dataset

We are using [4] Rotten Tomatoes critic review open-source dataset acquired, this has columns constituting rotten tomato links, critic names, review date and review content. We have merged this with the Rotten Tomatoes movie dataset that has movie genres, title, director, cast, release date and production accompany, through the rotten tomato links. For User reviews [14] we have entailed IMDB User Review dataset that has the URL's of each movie reviews of IMDB web site, we have Web Scraped these reviews from IMDB. Also, for more insights we have fetched movie details, i.e. movie cast from TMDB (The Movie Database). Finally, the Movie metadata [3], that has the movie title, release date, status of the movies, cast and genres to infer the influence and trends of movie reviews if there exist any. All the datasets have initially been acquired from Kaggle.com.

B. Loading in Database

We have first saved the raw and unprocessed data in databases. To accomplish this, we have made use of MongoDB [15] database since it is an exceptional NoSQL open-source database to store and manage enormous volumes of data. For the user review data, we had different csv files for diverse genres that has movie titles, release time and the review links to IMDB user reviews. We have web scraped data using 'Beautiful Soup' from these URL's and saved them in database, 25 reviews per movie, i.e. around 28,600 user reviews. Also, we have fetched cast information for the respective movies from TBDM open-source database and saved in our database. For rotten tomatoes critic reviews that are around 10 million reviews and movie metadata around 45,000 movie data, that were also csv files initially, we have saved this data in the database prior to processing.
MongoDB is our preference for the following reasons:

- **Document-Oriented-** MongoDB stores data in flexible, JSON like format for easy representation of complex and comprehensive data structures.
- **Relational Independency-** We have entailed MongoDB over traditional relational databases (RDBMS) to store non-relational data, since we do not intend to draw any relation between the movies and acknowledging the possibility of multiple reviews for a single movie. We are working to store multiple reviews of both critics and user for every individual movie and hence this is not a good practice to store this data making an individual row for each movie review. Also, storing and managing multiple reviews for a single movie would hinder the ACID (Atomicity, Consistency, Isolation, and Durability) properties of the relational database.
- **Scalability-** MongoDB is designed to scale across multiple servers, making it suitable for handling large amounts of data, we are working on around 10 million movie reviews in this project.

C. Pre-processing

After fetching the data from database, one cannot use raw data and hence, preprocess is one of the most significant tasks in data analysis. We have performed the following tasks for pre-processing of data:

1) Drop columns: Dropped columns like: directors, movie-info, run-time etc. , since they were not playing a significant role in the carrying out and finding the impacts of reviews.
1) Drop rows: Dropped rows, due to missing values in review and genre columns that were a pre-requisite of our project and maintain the data quality.
1) Drop duplicate values: dropped duplicate values, to avoid biased and inaccurate results, and maintain the data quality.
1) Convert original release date to datetime: This has been done to convert the date in 'yyyy-mm-dd' for better understanding and uniformity.
1) Sort by original release date: This is to arrange the movies in ascending order of the release date, supporting the purpose of our research.
1) Reset Index and return the processed data: Finally, this to reset the index according to all the alterations made.

D. Using pre-trained model for Sentiment Analysis

*1) **Why Distil-BERT** :* Acknowledging the better results leveraged by transformer models, we decided to make use of BERT. Distil-BERT is a small and light weighted model, which has very less resource computational requirements. This also makes it a lot more accessible to train on small data. It has proven to offer good results in resource-constrained environment with good inference speed. Despite its small size, it retains most of the significant performance of the original BERT model. Finally, with reduced computation requirements and good results where language and nuances are less critical, it offers a good balance between efficiency and performance. This research exploits how the novel pre-trained model DistilBERT, can be used for analysing the movie reviews. The transformer models work on sequence transduction enabling a connection between encoder and decoder. Unlike deep learning, these models allow parallelization and produce good results despite such short training periods. Having 6 encoders-decoders. BERT has the capacity to comprehend the context of the words in a bidirectional manner with the aid of the masked language mode. The BERT model tokenizes the words using word piece tokens. The BERT Base has 12 Transformer blocks , a hidden size of 768, and 12 attention heads with a total of 110M parameters. We have entailed DistilBERT model from Hugging Face transformers library [1]. The model can be fine-tuned for sentiment analysis tasks On Considering the tasks involving natural language processing, this model outperforms the RoBERTa and BERT model. Distil is a BERT that goes through a distillation process, making the model smaller and lighter. The model has the input size of 512 characters. The model was fine-tuned for 2 epochs with a batch size of 8, Adam optimizer with a learning rate of 2e-5. We have entailed DistilBERT for Critic review and User review to leverage sentiment analysis to obtain Sentiment label- 0 and 1 (0 for negative and 1 for positive review classification) and a Confidence Score (range [0-1] higher score, closer to 1 indicates high confidence in label classification). This analysis is then anchored to further study, analyse and visualize the influence of these reviews on movie release and production trends. The Distil-BERT model yields about 95% accurate results being a subset of the BERT transformer. The distil BERT-base-uncased model is much faster model and requires very less resources in comparison with other bert models. Hence, The distil-BERT model yields about 95-97% of BERT's performance while being 40% smaller 60% faster.

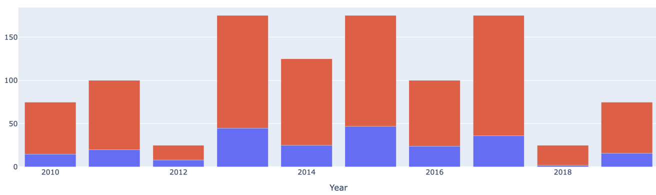

Fig. 1. Sentiment distribution for Genre: Crime over the Years
1. red: positive 2. blue: negative

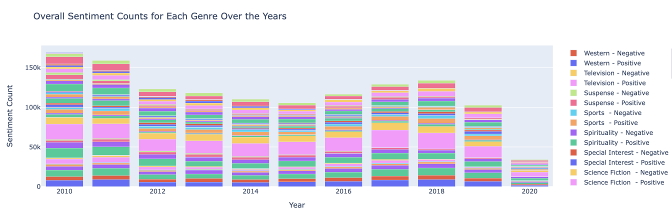

Fig. 2. Overall Sentiment Counts for each Genre over the years for Critic Reviews from Rotten Tomatoes dataset

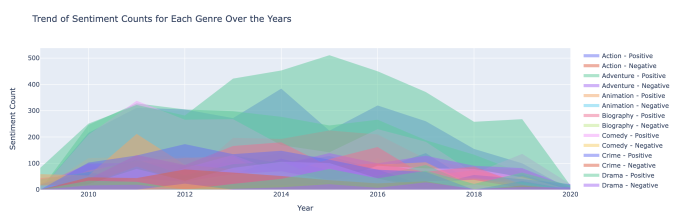

Fig. 3. Trends of sentiments Count for each Genre over the years for User Reviews from IMDB use review dataset

E. Data Visualization

Moving towards the data visualization, which is the core objective of our project, to find out the influence of the critic and user reviews on movie releases and production over the years. We have also tried to infer the correlation between these reviews in association to each genre over the release of movies over the years. For this sole purpose we have drawn a number of graphs acknowledge the trends of these reviews. The first graph is a histogram representing the "Percentage of positive and negative reviews by each genre", which illustrates high positive review percentage than negative irrespective of genre. Then, we have plotted "Sentiment distribution of each genre over the years", which portrays the reduction of negative reviews maintaining a high percentage of positive reviews in all genres like Musical, Drama, Suspense, Adventure etc. For this we have accounted sentiment distribution over the period of 2010-2020, of 31 genres for rotten tomato critic reviews and 18 genres for IMDB user reviews. We have decided to work on the 18 Genres that are present in the 'movie metadata' dataset, since we have the movie release information of these 18 genres. Further, we have made account of 'Overall', 'Positive' and 'Negative' to calculate the Sentiment count for each genre and have plotted a bar graph for the same over the years 2010-2020. Plotted a graph for "Trend of Sentiment Over the years of each genre", here the sentiment reviews go till 20,000 Positive reviews for the Drama genres in year 2017.
In the movie metadata, to take this forward we have done visualization of Movies 'Robert Downey Jr.' has worked in over the years, portrayed in 10 different pie charts for years 2010-2020. This explains his work in movies of different genres over ten years, like in 2016 Robert Jr. has worked in adventure genre (33.33%), action (33.33%) and science fiction (33.33%). We have plotted "Seasonal trends by genre" which demonstrates the number of movie release in a year (i.e over the months) of each genre in a bar graph, finally plotting this again but throughout the period of 10 years. To embark the affects of these reviews on the movie releases, have plotted sentiment count over the years graph for each genre acknowledging both User- positive, negative and Critic- positive, negative reviews. The graphs provided here are, Figure 1. "Sentiment distribution for Crime genre, over the years", as you can see in the graph the positive reviews have always been higher than negative ones, although there is an extreme downfall in reviews on Crime in the year 2018, but again its risen for the upcoming years. Next, Figure 2. represents the overall sentiment count for each genre for critic reviews, as you can see there is a continuous hype for television positive reviews and same goes for suspense genre positive reviews. Then in Figure 4. represents the Seasonal trends by Genre, its visible that the number of movies have risen over the years and a subsequent rise in the genre 'Comedy' and 'Drama'. In the conclusion, we have plotted the influence of these reviews on the movie releases that we'll be covering in evaluation.

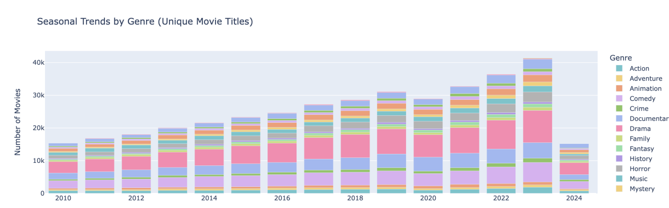

Fig. 4. Seasonal Trends by Genre of Unique Movie Titles

F. Results and Evaluation

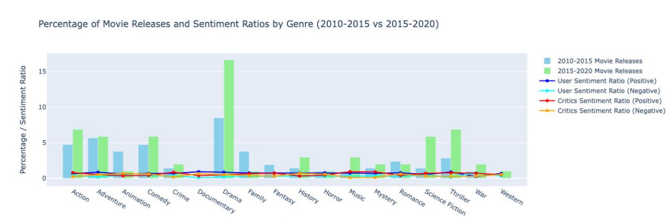

Fig. 5. Impact of Positive/ Negative Reviews on Movie Releases by Genre from 2010-2015 and 2015-2020

To determine the result of the influence of both user and critic reviews, we have first taken into consideration the 18 genres that are common in the three datasets and these 18 genres for which we have the movie release and status

```
The number of releases for the genre 'Action' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Adventure' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Animation' decreased from 2015 to 2020 compared to the period from 2010 to 2015 due to negative reviews.
The number of releases for the genre 'Comedy' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Crime' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Documentary' decreased from 2015 to 2020 compared to the period from 2010 to 2015 due to negative reviews.
The number of releases for the genre 'Drama' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Family' decreased from 2015 to 2020 compared to the period from 2010 to 2015 due to negative reviews.
The number of releases for the genre 'Fantasy' decreased from 2015 to 2020 compared to the period from 2010 to 2015 due to negative reviews.
The number of releases for the genre 'History' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Horror' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Music' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Mystery' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Romance' decreased from 2015 to 2020 compared to the period from 2010 to 2015 due to negative reviews.
The number of releases for the genre 'Science Fiction' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Thriller' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'War' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
The number of releases for the genre 'Western' increased from 2015 to 2020 compared to the period from 2010 to 2015 due to positive reviews.
```

Fig. 6. Influence of positive and negative reviews on movie release

information. Next, we have separated the data by year to scrutinize the trends, rise or fall in the movie releases with respect to genres, for this we have taken into account, the periods 2010-2015 and 2015-2020. First, we have calculated the total number of releases in the two time-frames, then have calculated the change in the number of movie releases from the years 2010-2015 and 2015-2020 based on genre. Finally, have determined the impact influence of positive and negative reviews on the movie releases. Taking a look at figure 5, it portrays the increase in the movie release in the years 2015-2020, majorly in 'Drama', 'Sci-Fi' and 'thriller' especially, where you can also see a significant rise in the positive rise by both critics and users. And in Figure 7, there is a visual hike in the positive and negative critic reviews till year 2014 and then there's a conical fall in the same. Which we can confidently say is quiet a correlation, stated in Figure 6, that tells you the Number of Release by each Genre from 2015-2020 compared to the movie releases in the time-frame 2015-2020 impacted due to positive reviews.

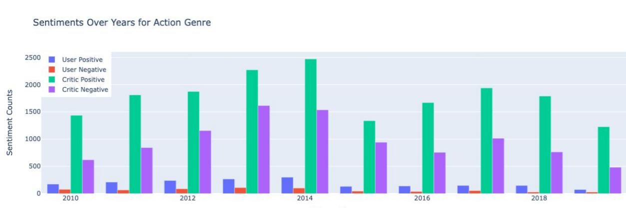

Fig. 7. Sentiments Over Years for Action Genre

G. Conclusion and Future work

From this project we have inferred the relevance and Influence of the Reviews (Positive and Negative) both from Critics and Users. We have analysed the significant changes in the Movie Releases in the time-frame 2010-2015 and 2015-2020. As you can infer from Figure 5. that there has been a significant inclination in the movie releases belonging to the Genres: 'Drama', 'Sci-Fi' and 'thriller' in the years 2015-2020 the based on the increment of the Positive Reviews on the same, and a declination on the Genre 'Romance' based on the decrease in the positive reviews. Also, in Figure 6. one can demonstrate clearly that the movie releases have been increased in the specified genres like 'Action' in the years 2015-2020 in comparison to 2010-2020. In conclusion, we have discovered the positive influence of the reviews on Movie Releases which can be fruitful to production houses and opens opportunities for new and innovative ideas of Movies in the specified booming genres.

References

[1] Hugging Face. (2024, April). sarahai/movie-sentiment-analysis. [Online]. Available: https://huggingface.co/sarahai/movie-sentiment-analysis/tree/main

[2] Jai Dalmotra. (2024, March). Movies Reviews. [Online]. Available: https://www.kaggle.com/datasets/jaidalmotra/movies-review?resource=download

[3] Rounak banik. (2018). The Movies Metadata. [Online]. Available: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

[4] Stefano Seone. (2021). Rotten Tomatoes Movies and Critic Reviews dataset. [Online]. Available: https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

[5] G. Putrada, N. Alamsyah and M. N. Fauzan, "BERT for Sentiment Analysis on Rotten Tomatoes Reviews,". 2023rd International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2023, pp. 111-116.

[6] M. G. Sousa, K. Sakiyama, L. d. S. Rodrigues, P. H. Moraes, E. R. Fernandes and E. T. Matsubara, "BERT for Stock Market Sentiment Analysis," 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 2019, pp. 1597-1601, doi: 10.1109/ICTAI.2019.00231.

[7] M. S. Başarslan and F. Kayaalp, "Sentiment analysis with ensemble and machine learning methods in multi-domain datasets", TUJE, vol. 7, no. 2, pp. 141–148, 2023, doi: 10.31127/tuje.1079698.

[8] Vinayagamurthy, S. "Utilizing the Transformer models for Analysing Deceptive Reviews and Aspects of the reviews "(Doctoral dissertation, Dublin, National College of Ireland) 2022.

[9] Raghunathan, N. and Kandasamy, S. "Challenges and Issues in Sentiment Analysis: A Comprehensive Survey".2023.

[10] Baid, P., Gupta, A. and Chaplot, N. "Sentiment analysis of movie reviews using machine learning techniques". International Journal of Computer Applications, 2017,179(7), pp.45-49.

[11] Dashtipour, Kia, Mandar Gogate, Ahsan Adeel, Hadi Larijani, and Amir Hussain. 2021. ”Sentiment Analysis of Persian Movie Reviews Using Deep Learning” Entropy 23, no. 5: 596. https://doi.org/10.3390/e23050596

[12] Shuang, K.; Ren, X.; Chen, J.; Shan, X.; Xu, P. “Combining Word Order and CNN-LSTM for Sentence Sentiment Classification”. In Proceedings of the 2017 International Conference on Software and e-Business; ACM: New York, NY, USA, 2017; pp. 17–21.

[13] K. Soni, P. Yadav and Rahul, ”Comparative Analysis of Rotten Tomatoes Movie Reviews using Sentiment Analysis,” 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1494-1500, doi: 10.1109/ICICCS53718.2022.9788287.

[14] TMDB. (2024). The movie database. [Online]. Available: https://developer.themoviedb.org/reference/intro/getting-started

[15] Győrödi, R. Győrödi, G. Pecherle and A. Olah, ”A comparative study: MongoDB vs. MySQL,” 2015 13th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, 2015, pp. 1-6, doi: 10.1109/EMES.2015.7158433.