2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 19-20 December, Dhaka

2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) |978-1-6654-0489-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/STI50764.2020.9350525

Sentiment Analysis on Adventure Movie Scripts

Rifat Rahman[a*], Dr. Md Abdul Masud[b], Raonak Jahan Mimi[a], Mst. Nusrat Sultana Dina[a]
[a]Faculty of Computer Science and Engineering
[b]Department of Computer Science and Information Technology
Patuakhali Science and Technology University
rifat14@cse.pstu.ac.bd*, masud@pstu.ac.bd, mimiraonak14@cse.pstu.ac.bd, dina14@cse.pstu.ac.bd

***Abstract*—Sentiment analysis is the process of mining subjective information in different kinds of sources for playing an important role in making decisions. Movie scripts are an interesting text-domain including their diverse expressions for sentiment mining. In this paper, we have done sentiment analysis by applying Minimum Viable Product (MVP), Exploratory Data Analysis (EDA), and TextBlob sentiment analysis techniques. We compute the sentiment polarity and show the fluctuation of sentiments over time of a specified movie genre. We collected a dataset consisting of 20 adventure genre's movie scripts for sentiment analysis on movie scripts. The experimental results show that the fluctuations of the sentiment of the adventure genre's movie reflect the viewer's mindset as public opinion for social impact.**

***Index Terms*—Natural Language Processing, Sentiment Analysis, Minimum Viable Product, Exploratory Data Analysis**

I. Introduction

The script is like the soul of the movie which contains dialogues and directions, complete plots, characters, and tones for the whole film. As a multifarious exposition of sentiments expressed in movies, that's why movie scripts are the fascinating origin of the text. "Sentimental weight" [1] of the scripts is usually an instrument to attain the aforecited objective and indeed a very important and influential one. It is apparently the most perceptible point of resonance and communication with the audience. Films ordinarily comprise scenes where emotions change dynamically, between happiness and sadness, quiescence and anger as to collaborate the descriptive headway, while some works are characterized by a redundant emotional 'weight', such as adventuresome exploration in an adventure-based film [1]. For achieving this aim, the script requires to be written in a scheme that adopts the congenial sentiments and sanctions the actors to imprint it in their performances. These days most spectators don't just express their like or abhorrence of a film and rather inclined to the various emotions that the film evoked in them. It would be a fascinating request to examine the planned feeling of a film's content. The primary goal of this study is to accumulate movie scripts in order to analyze the emotions manifested in the movies. Opinions can be positive, negative, or unbiased or it can contain an arithmetical score articulating the adequacy of the sentiment [2]. Intuition can be alluded to as emotions. Our methodology, sensation, and judgment on a certain reaction or a specific function can be alluded to as assessments. Feelings can't be measured as specifics as they are close to personal limitations. On account of suppositions, twofold speculations are thought of, as against/for, good/terrible, and so forth. Semantic direction and polarity [3] are commonly deliberated as sentiment analysis jargons. Sentiment analysis can be characterized as the techniques used to coerce, perceive, or recognize the opinion substance of an original copy. Assessments can be unearthed just as pulled out in sentiment analysis. By applying various techniques like web scraping, data cleaning, EDA [4], and TextBlob module - this sentiment is analyzed from movie scripts.

II. Related Works

Sentiment analysis, as a field of study, has met a huge expansion in its materialness in different areas, as it can undoubtedly get cross-disciplinary and encourage various cycles [5, 6, 7]. Its effortlessness and clear methodology have set up it as a regarded factor of text investigation. Demonstrative areas where sentiment analysis is executed are business frameworks, advertising efforts, and recommender systems [8]. Lately, sentiment analysis related to Natural Language Processing (NLP) and Machine Learning (ML) methods has been utilized in a variety of various applications concerning movie scripts. They have been utilized to recognize designs in film structures [9] and figure out how to foresee the accompanying enthusiastic state dependent on the past [10], indicating individually that 'effective' motion pictures follow explicit story movements and have a specific "stream" and consistency in the way that passionate states unfold. Furthermore, it has been seen that the current binary ("positive/negative") sentiment analysis procedure have an 'energy inclination', preferring the learning of positive feelings over negative ones, thus "thinking little of" the last's presence. This can make a significant discrepancy in the precision of around 10 to 30%. It is indicated that con-

978-1-6654-0489-1/20/$31.00©20XX IEEE

templating meta-highlights (capital letters, accentuation, and grammatical forms) mitigates this issue [11]. The methodology implemented to perform sentiment analysis involved Python's various Natural Language Toolkit (NLTK) modules like BeautifulSoup, pandas, re (Regular Expression), CountVectorizer, Counter, text, WordCloud, matplotlib, NumPy, and TextBlob. All these modules used for NLP procedures that include - scraping data from a website, text pre-processing, organizing the cleaned data into a way that is easy to input into other algorithms, finding most common words for creating word clouds, measuring the size of vocabulary for finding the number of unique words and also how quickly movies delivered their dialogues, calculating the amount of profanity for isolating bad words from organized data, polarity/subjectivity calculation for the final stage of sentiment analysis. The emerging significance of sentiment analysis in film assessment has likewise incited notable and regarded online rivalry networks like Kaggle to arrange activities to propose ingenious solutions for the issue. It is clear, however, that to deliver an all-around documented arrangement, highlights must be painstakingly chosen, thinking about the difference of terms, the treatment of refutation, and the treatment of sentiment words [12]. Our paper adds to the current logical system by exploiting information from the film contents and by applying subjectivity/objectivity identification, to lead notion and feeling investigation. We accept our research to be of eminent significance as it can demonstrate another beam of hope to the issue which we are looking for quite a while.

III. Methodology

Sentiments allude to mentalities, conclusions, and feelings. As such, they are abstract impressions instead of target realities. Various sorts of sentiment analysis utilize various procedures and methods to mark out the sentiments seized in a specific text. There are two primary kinds of sentiment analysis: subjectivity/objectivity identification and feature/aspect-based sentiment analysis. Here we utilized the strategy for subjectivity/objectivity marking off in our work. subjectivity/objectivity unification proof involves characterizing a sentence or a piece of text into one of two classifications: subjectivity or objectivity. In any case, there are difficulties with regard to leading this kind of analysis. The primary challenge is that the signification of the word or even an expression is regularly dependent upon its context. Our work depends on figuring the emotional extent of a film's transcript. As nowadays the younger generation inclined towards adventure movies where plots comprise the components of travel, they regularly include heroes who must depart their home or spot of complacency and go to distant terrains to satisfy an objective. That's why we chose adventure genre movie transcripts for applying our sentiment analysis techniques because it shows the fluctuations of emotions explicitly. In "Fig. 1" the full work-flow is shown.

A. Web Scraping

Web scraping additionally named web data extraction is a strategy utilized to take out a lot of information from sites through which the information is drawn out and saved in the local record in our PC or in the database in spreadsheet format. The information showed by most sites must be seen utilizing an internet browser. They don't grant the service to replicate this information for individual use. The main alternative at that point is to physically copy and paste the information - a repetitive act that can adopt numerous hours or occasionally days to finish. Web scraping is the strategy of computerizing this procedure, in order that rather than physically replicating the information from sites, the Web scraping programming will play out a similar task inside a small amount of the time. Barely any methods of web scraping are - physically scraping information utilizing browser extensions, picking content on a site with XPath, web scraping utilizing Python, HTML parsing, vertical collection, DOM parsing, text design coordinating, and so forth. At the information-gathering stage for gathering film transcripts, we utilized basic scrapers in Python to scrape specific web pages. The transcripts were assembled in .txt format. It utilizes the BeautifulSoup package and turns to the IMSDB site, finding all the transcripts titles, arranged in sequential order. BeautifulSoup is a Python library for getting information out of HTML, XML, and other markup languages. BeautifulSoup assists with pulling specific substances from a site's page, eliminate the HTML markup, and store the data. It is an instrument for web scraping that assists with tidying up and parse the records we have pulled down from the website.

B. Data Cleaning and Preprocessing

Data preprocessing includes the change of the crude dataset into a justifiable arrangement. Preprocessing information is a basic stage in data mining to improve information effectiveness. The data preprocessing techniques straightforwardly influence the results of any algorithm. Data preprocessing and clearing are commonly done in some straightforward advances like assembling the information, bringing in libraries, cleaning the information, linking information, combining information, and so on. We followed the Minimum Viable Product (MVP) [13] approach - start simple and repeat. To start with, we tokenized [14] all the content of the scripts. At that point we applied regular information cleaning steps like creating all the content into lower case, eliminating accentuation [1,2] marks, eliminating mathematical qualities, eliminating basic unreasonable content, and eliminating stop words [15] for cleaning our movie's transcript.

C. Pickling Data

"Pickling" is the cycle whereby a Python object pecking order is changed over into a byte stream. Python pickle module is utilized for serializing and deserializing a Python object structure. In machine learning, while working with scikit-learn library, we have to save the prepared models in a document and reestablish them to reuse it to contrast the model and different models, to test the model on new data. The saving of information is called serialization while reestablishing the information is called deserialization. We pickled different types of data for additional utilization of sentiment analysis.

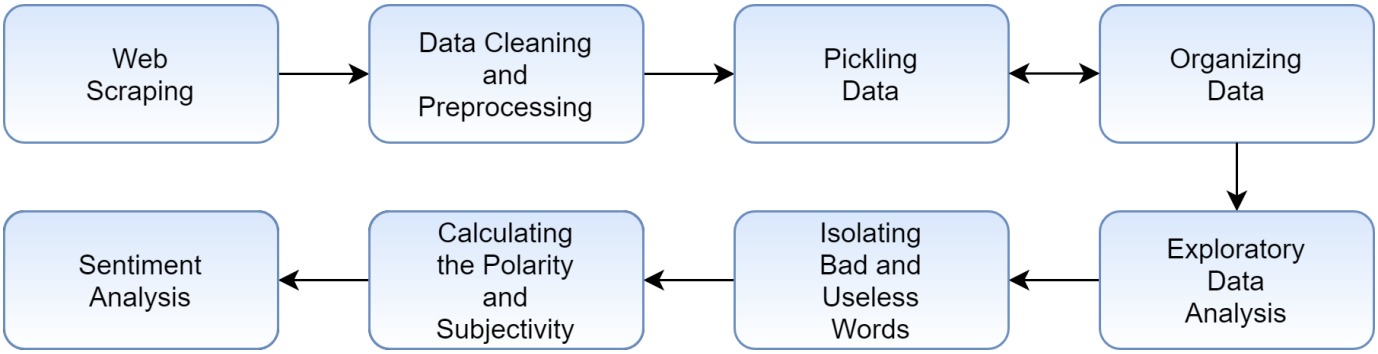

Fig. 1. Work Flow of Adventure Movie's Sentiment Analysis

D. Organizing Data

We coordinated our information in two standard structures like corpus and Document Term Matrix and pickled it for additional utilization of sentiment analysis.

E. Exploratory Data Analysis

Exploratory data analysis alludes to the basic cycle of performing introductory examinations on information to find designs, spot oddities, test theories and to check suppositions with the assistance of outline measurements and graphical portrayals. EDA should be possible in a few different ways like univariate analysis, bivariate analysis, multivariate analysis, and so forth. Prior to applying any extravagant calculation on our coordinated information, should have been investigated first for investigating the information and check whether what we're seeing bodes well.

F. Isolating Bad and Useless Words

In this progression, we will eliminate the swear and futile words that have no commitment to sentiment analysis by specifying those manually. By eliminating these sorts of words makes our printed information all the more new and real.

G. Calculating the Polarity and Subjectivity

Polarity is a float that lies in the scope of (-1,1) where 1 means a positive explanation and -1 means a negative articulation. Abstract sentences for the most part allude to sincere belief, feeling, or judgment though target alludes to verifiable data. Subjectivity is additionally a float which lies in the scope of (0,1). In our work for figuring the extremity and subjectivity of motion pictures, we select TextBlob's sentiment analysis process.

H. Sentiment Analysis

At the last stage to apply sentiment analysis on our corpus we utilized TextBlob module. TextBlob is a python library for NLP. TextBlob effectively utilized NLTK to accomplish its errands. TextBlob is a straightforward library that underpins complex investigation and procedure on textual data. TextBlob restores the polarity and subjectivity of a sentence. Subsequent to examining the assumption, all things considered, we attempted to show the sentiment changes in a couple of motion pictures by plotting.

IV. Experiment

The motive of our work is to understand how subjective a movie's script is. On the basis of the polarity variation, sentiment fluctuations are shown.

A. Dataset

In this stage for gathering film transcripts, we take advantage of the basic scraper of Python for scraping the particular web contents. To gather the film transcripts, we make use of the IMSDB site that is also familiar as the web's biggest film transcripts storehouses and hold in excess of 1100 film transcripts and drafts [1]. We assembled 20 film transcripts of the adventure genre. The contents were accumulated in .txt lay-out. It exploited the BeautifulSoup package and turned to the IMSDB site, finding all the transcripts titles, sorted in sequential order. BeautifulSoup assists with pulling specific content from a page, eliminate the HTML markup and stored the data. It is an instrument for web scraping that assists with tidying up and parse the data we have dragged down from the web. The scraper's initial activity is to reserve all the <pre> labels, resembling to film titles, in an inventory. Thereafter, it recapitulates the inventory of titles and redirects towards the URLs, where it loads transcripts in .txt patterns and makes a directory for depositing them. The scraper keeps just the information present in <td> labels where the predefined class was (”scrtext”) and selects to take no account of labels like <head> or <footer> [1].

B. Experimental Setting

For doing our sentiment analysis work, we used the Intel Core-i5 8265U processor which has 4 cores. The CPU clock speed is 1.6 - 1.8 GHz and the cache size is 6MB. We used the Python programming language for completing our research work. The methodology implemented to perform sentiment analysis involved Python's various NLTK modules and libraries like BeautifulSoup for parsing HTML documents, requests used to make HTTP requests simpler and more human-friendly, pickle for serializing and de-serializing python object structures, re (Regular Expression) can be used to check if a string contains the specified search pattern, CountVectorizer used to convert a collection of text documents

to a vector of term/token counts, Counter is a container that will hold the count of each of the elements present in the container, text, NumPy used for working with arrays, pandas for data manipulation and analysis, TextBlob for processing textual data, WordCloud for visually representing the text data, matplotlib provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits, and pillow for opening, manipulating, and saving many different image file formats.

C. Result and Discussion

Applying EDA, first, we find the top words of each movie which been said most of the time of the movie. In "Fig. 2" the most common words of few movies are shown through a word cloud. WordCloud is Python's library used for quickly perceiving the most prominent terms. This tool is quite handy for exploring text data and making any report more lively. WordCloud is a data visualization technique used for repre-

Fig. 2. Generated Word Cloud of Few Movies

senting text data in which the size of each word indicates its frequency or importance. It displays a list of words, the importance of each being shown with font size or color. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites. And then the top common words which half of the movies contains, that was excluded from the list of the top words because it contains less meaning of sentiment. For generating word cloud in Python, modules needed are – NumPy, pandas, matplotlib, pillow, WordCloud. The NumPy library is one of the most popular and helpful libraries that is used for handling multidimensional arrays and matrices. It is also used in combination with the Pandas library to perform data analysis. For visualization, matplotlib is a basic library that enables many other libraries to run and plot on its base including seaborn or word cloud. The pillow library is a package that enables image reading. Pillow is a wrapper for Python Imaging Library (PIL). After displaying the unique words of each movie using word cloud, through "Fig. 3" we can understand which movie's delivery is the fastest and slowest flow of emotion. The movie run times are collected from IMDB (The world's most popular and authoritative source for movie, TV, and celebrity information) in minutes. And at the end of the EDA technique, the amount of profanity is calculated and the bad word is been isolated for further sentiment analysis.

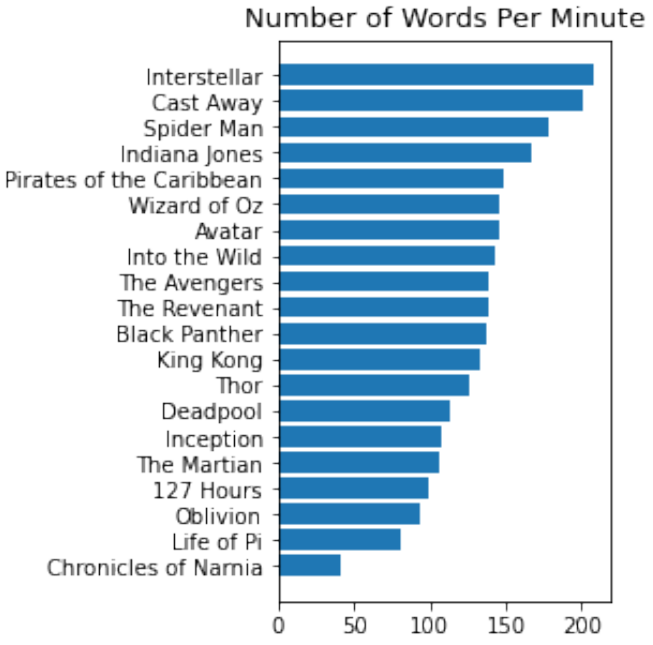

Fig. 3. Number of Words Per Minute

Now we apply sentiment analysis on our corpus we used TextBlob module. TextBlob is a python library for NLP. TextBlob effectively utilized NLTK to accomplish its errands. TextBlob is a basic library that upholds complex investigation and procedures on literary data. TextBlob restores the polarity and subjectivity of a sentence. Every term in a corpus is entitled regarding polarity and subjectivity (there are more marks too, yet we will overlook them for the present). A corpus' sentiment is the normal of these. Polarity - How positive or negative a term is. -1 is extremely negative. +1 is exceptionally positive. Subjectivity - How emotional, or obstinate a word is. 0 is a reality. +1 is a lot of feeling. TextBlob has semantic names that help with the fine-grained investigation. For instance emojis, exclamation marks, emoticons, and so forth. Subjectivity measures the extent of the closely-held conviction and genuine data contained in the content. The greater subjectivity implies that the content comprises a closely-held intent as opposed to authentic data. TextBlob has one more criterion—intensity. TextBlob computes subjectivity by taking a glance at the 'intensity'. Intensity decides whether a word changes the following word. In Table I, few data on the subjectivity and polarity of our analyzed movies have been shown from which we could figure out that the movies of today's generation are how subjective and opinionated. After seeing the data of Table I, it could explicitly understand that the "Deadpool" movie holds higher subjectivity (0.511080) among all the movies, so this movie is more opinionated among them. And "Chronicles of Narnia" movie contains the lowest subjectivity (0.417043) among all, so it could be clearly said that this movie is less opinionated. After showing the

TABLE I
SUBJECTIVITY AND POLARITY OF FEW MOVIES

Movie Name	Subjectivity	Polarity
Avatar	0.460565	0.024373
King Kong	0.456331	-0.000484
Chronicles of Narnia	0.417043	0.040720
Pirates of the Caribbean	0.462343	0.063743
Spider Man	0.436305	0.048594
Deadpool	0.511080	-0.010577
....	....	....
The Revenant	0.418798	-0.015828

polarity of each movie's, it can be seen that the "Pirates of the Caribbean" movie holds higher polarity (0.063743) amid all the movies, so this movie shows more positive vibes among them. And "The Revenant" movie contains the lowest polarity (-0.015828) amid all, so it could be clearly said that this movie shows negativity or gloomy. From the visualization of "Fig. 4," it is clear that most of the movies are showing positive wave of sentiment, some are showing neutral and only a few are showing the negative wave of sentiment. By applying the TextBlob sentiment analysis technique on data extracted from our data-set we find the average sentiment type which was shown by adventure genre's movies analysis. Through "Fig. 4", we can see that the movies in which polarity is greater than 0.05 are showing positive sentiment through the plot of their movies, which movie's polarity is from 0.00 to 0.05 they are neutral sentiment and which are less than 0.00, they are showing negative sentiment.

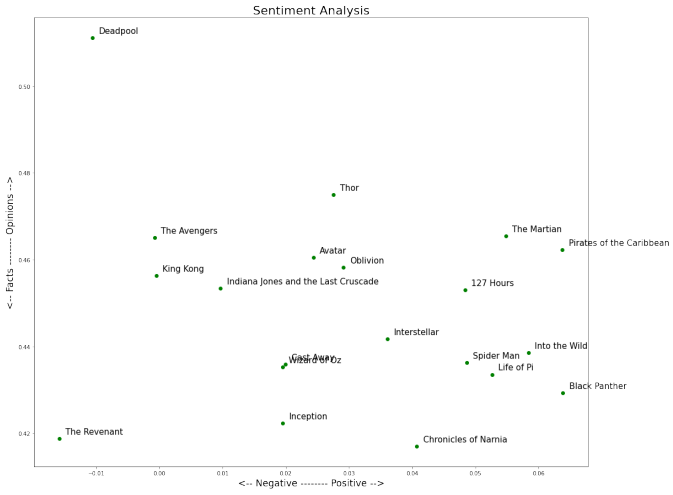

Fig. 4. Sentiment Analysis

Following all these steps of sentiment analysis the vicissitudes of sentiment extracted from each movie's script. The findings of our whole work can be shown in "Fig. 5". We have shown a few movie's plots of sentiment fluctuation. "Fig. 5" indicates a few movies flow of emotion from start to end clearly. In our work for calculating the polarity and subjectivity of movies, we select TextBlob's sentiment analysis function.

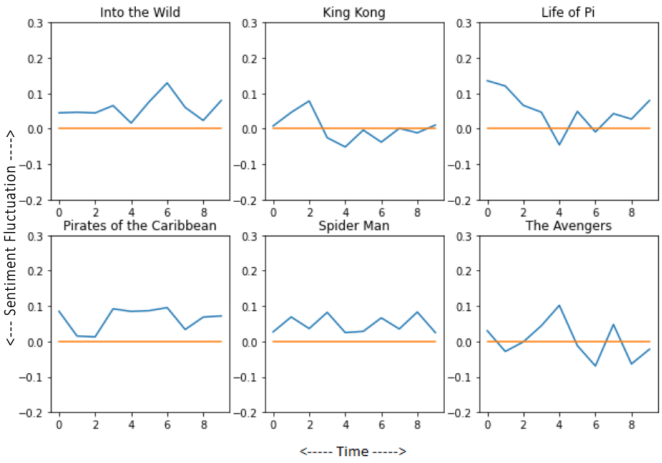

Fig. 5. Sentiment Fluctuation of Few Movies

The explanation behind picking it is that they are much the same as Python strings. In this way, we can change and trifle with it as the same we did in Python. Whereas it is based on the foundation of NLTK and pattern, thusly it gives an instinctive interface to NLTK. Additionally gives language interpretation and detection which is controlled by Google translate. As it provides the language translation and detection feature so we can detect various languages of movies and could work with the scripts of different languages of different origins. TextBlob is well known on account of the manner in which it permits us to just work with literary data with no problem with intricate API calls.

V. RELATED TO INDUSTRY 4.0

Industry 4.0 determines the 4th Industrial Revolution which is a combination of development in the Internet of Things (IoT), Robotics, Artificial Intelligence (AI), Quantum Computing, Genetic Engineering, as well as different technologies. The fourth industrial revolution is constructing on the third, the technological revolution which has been happening since the middle of the remaining century. Expanding productivity in film industries generally in the motion picture industry is a significant issue. Sentiment analysis of NLP on film's script can assist to accomplish the sustainable advancement in the film industry.

VI. CONCLUSIONS

Our paper inquired into the blending emotion of the adventure genre's movies and analyzed the fickleness of the movie plots. This work shows the sentimental pattern of this specified genre's movies, which clearly illustrates the reason behind the huge acceptance among nowadays audiences. Among various complex techniques of sentiment analysis, we used some simple techniques to extract the exact sentiment from the

corpus of our data. At the time of processing our data, the EDA technique gave the accurate result through the word cloud technique which affects the further subjectivity and polarity calculation terms in the TextBlob module to bring out the expected result from sentiment analysis. A good movie, which plot is very subjective gives the audience a great experience and a movie with low subjectivity gives the audience the worst experience. Nowadays before watching a movie people check the IMDB rating for escaping from any worst experience from the movie. But when the movie releases, the first audience don't know how subjective the movie's plot would be. As in today's era, people don't like time wastage and time consumption. For this reason, in near future, a web application could be made where people could know about the subjectivity and polarity of a movie. Higher subjectivity could be the symbol of a good movie, and the findings of our research could work in the back-end of that web application.

ACKNOWLEDGMENT

This research was supported by the Research and Training Center (RTC) under project Code No. 5921 in Patuakhali Science and Technology University, Bangladesh.

REFERENCES

[1] P. Frangidis, K. Georgiou, and S. Papadopoulo, "Sentiment Analysis on Movie Scripts and Reviews Utilizing Sentiment Scores in Rating Prediction," 2020, p. 43.

[2] K. Chakraborty, S. Bhattacharyya, R. Bag, and A. A. Hassanien, "Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Techniques," 2019, p. 12.

[3] J. Kim, Y. Ha, S. Kang, H. Lim, and M. Cha, "Detecting Multiclass Emotions from Labeled Movie Scripts," in *IEEE International Conference on Big Data and Smart Computing* 2018, p. 590.

[4] C. H. Yu, "Exploratory data analysis in the context of data mining and resampling," 2015.

[5] G. Mesnil, T. Mikolov, M. Ranzato, and Y. Bengio, "Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews," 2014.

[6] B. Pang, and L. Lee, "Opinion Mining and Sentiment Analysis," 2008, pp. 1–135.

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal* 2014, pp. 1093–1113.

[8] K. Chakraborty, S. Bhattacharyya, R. Bag, and A. E. Hassanien, "Comparative Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Approach," in *The International Conference on Advanced Machine Learning Technologies and Applications* 2018, vol. 723, pp. 311–318.

[9] S. Lee, H. Yu, and Y. Cheong, "Analyzing Movie Scripts as Unstructured Text," in *IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)* 2017, pp. 249–254.

[10] T. P. Sahu, and S. Ahuja, "Sentiment analysis of movie reviews: a study on feature selection and classification algorithms," in *International Conference on Microelectronics, Computing and Communications (MicroCom)* 2016, pp. 1–6.

[11] J. Kim, Y. Ha, S. Kang, H. Lim, and M. Cha, "Detecting Multiclass Emotions from Labeled Movie Scripts," in *IEEE International Conference on Big Data and Smart Computing (BigComp)* 2018, pp. 590–594.

[12] A. Manek, P. D. Shenoy, M C. Mohan, and V. K R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," 2016, pp. 135–154.

[13] A. N. Duc, and P. Abrahamsson, "Minimum Viable Product or Multiple Facet Product? The Role of MVP in Software Startups," 2016, p. 118.

[14] D. Kim, S. Lee, and Y. Cheong, "Predicting Emotion in Movie Scripts Using Deep Learning," in *IEEE International Conference on Big Data and Smart Computing* 2018, p. 530.

[15] D. Ananda, and D. Naorema, "Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering," in *7th International conference on Intelligent Human Computer Interaction, IHCI* 2015, p. 88.