A Visual Analytics approach to exploring television sitcom script dialogue: F.R.I.E.N.D.S.

Thanawat Thanaponpaiboon

Abstract— In this research, we shall use a visual analysis approach to visualize and analyse the text from the scripted dialogue of one of the most popular television sitcoms F.R.I.E.N.D.S. The objective of this study is to identify and visualize the insight from the text script including prioritising characters through dialogue, identify the relationship between each character, emotional and sentimental of the character using a combination of visual analytics and human reasoning. Network graphs are useful for visualizing connections between each character. However, understanding the relationship only surface through the network graph is not enough to display in-depth details such as a variety of dialogue and emotional communication. The challenge in this study is we wanted to compare how much important each main character played a role in their emotional expression; does it influence the overall dialogue? Indicate emotional communication and expression in characters are also a significant part of what draws people's attention to keep watching the TV series and wonder what will happen in the next episode.

1 Problem Statement

With the blossoming of technologies along with a strong flow of data trends accessing the data is seem to be easier than before. People start to aware of the benefits of using data to analyze and visualize to see through the trends from past events and predict what will happen in the next period. Including filmed and tv series production began collecting data on the script of each movie and series. In the hope that in the future people can leverage this data and use it to produce new fantasy scripts.

For this reason, we shall apply visual analytics techniques to exploring F.R.I.E.N.D.S. dialogue scripts that we expected to gain better insight from the text script rather than watching and listening through the dialogue through the television. Through this analysis our objective is to explore and answer the following research questions:

- Illustrate the proportion of the dialogue between main characters. Which character has the most dialogue and screen time?
- Whether using a network graph to represent the relationship is an effective way to illustrate how characters are connected through dialogue or not?
- Is it possible to pinpoint the main character's personality from the dialogue script through sentiment analysis?
- Is emotional sentiment from main characters possibly impact the direction of the plot in television series or not?

We also use human knowledge about familiarity with the content and dialogue of the characters along with visual programming and computing to avoid the probability leading to failed projects which we hope to achieve to gain fruitful insight through addressing these research questions.

2 State of the Art

The script, which serves as the core of the movie contains all the dialogue, directions, character development, and tone of the film [4]. It is important to understand the nature of our dataset [1] which we can study and reference from a similar study as a guideline to avoid problems or errors that may be encountered during the research. Text scripts from movies and television series become a broad and popular domain of text analysis. There are many methods for analyzing the text data like sentimental analysis or using the network to represent the relationship as a ubiquitous method across various areas such as social media can be used for developing insights from visualize.

All of these papers and articles below utilize these analytic approaches as we mentioned to address their analytic question. Hence, it is a good opportunity to investigate their study to replicate or leverage our data using their finding as our baseline.

Min and park [2] used text analysis methods from computational linguistics to analyze Les Misérables generating the network graph that represents the character interaction and using sentiment analysis to identify relationships between each character explicitly through human interaction and communication. Likewise, Park, et al. [3] used social network graphs to investigate character relationships using character dialogue, which claims in their study that investigation through dialogue is greater than any kind of data such as video or voice for creating social networks of characters. The research classifies each character into major or minor roles based on their centrality as well as

extracts the sequences via clustering allowing to observe of the most important characters and the correlation group. However, the major problem that they encounter is the lack of ability to extract story information from a temporal aspect. Another interesting article that is consistent with our study in this is from Rahman, et al. [4] study movie scripts which state that consist of diverse expressions. Word cloud has been used to visualize the frequency of words in each movie which has been said most of the time of the movie along with textblob for sentiment analysis to compute the sentimental polarity score and present the result by plotting the fluctuation of sentiments over time. The study points out that the sentiment score from the movie corresponds with the viewer's attitude to the movie.

We will apply a similar technique from these articles to study the emotional sentiment of the main characters through the dialogue text file and study the relationship between each character then create meaningful relationships through the networks graph. As we are using dialogue scripts from TV shows instead of using movie scripts and we have learned some techniques and tools that can handle text analyzing tasks from other previous research, we should expect to have success in this research and obtain fruitful insight for answering our analytic question.

3 Properties of the Data

The dataset that we shall use in this study is F.R.I.E.N.D.S. dialogue script [1] from Kaggle website which contains the text scripts of all the episodes of the ten seasons of F.R.I.E.N.D.S. aired between 1994 to 2004 as a .txt file. The actual data was scraped by Kansal [5] who used RNN model text generation to produce new scripts for Friends TV Show, extract from the transcripts all the episodes and store it as a .txt file.

We will be using this pre-processed version published on Kaggle for the analysis. This file contains key features such as the name of the character and the dialogue that represent their respective lines. Both properties are string formats that have not been arranged in an orderly manner. In order to take full advantage of this dataset, data transformation and data derivation are required. A key thing to note when working with large text data is we need to ensure that our data has to be arranged in order properly to allow us to perform a fully analyze. To gain a better understanding of the relationship between words and sentences, word tokenization has been applied to help us organize the string text. This step breaks down each sentence into words called tokens. These tokens can use to perform analysis tasks or computing tasks that we are going to do further.

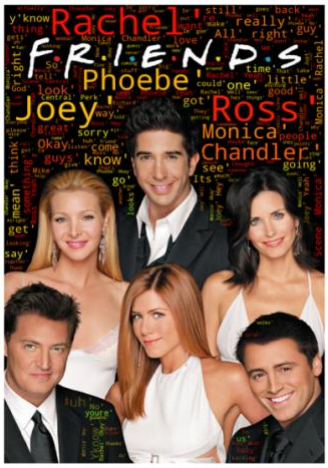

Fig. 1. Word clouds in F.R.I.E.N.D.S.

Once we extract the word from the sentence by tokenization, we can see the frequency and sequence of each word from the entire text. Figure 1 above shows an example of the most frequent word that has been found in this script. Apparently, the six main characters' name is the most apparent word in the script due to the whole story running around by these six characters. We can also link the distance between each character name to generate the relationship network illustrating the interactions between every pair of characters. In addition, feature engineering and derivation play an important role to change the data into an appropriate adjacency before doing further analysis. Since our data were collected in a text file we need to transform and collect it properly into a panda dataframe in order to fully exploit the information. The character name that appears before the dialogue has been derived into the object column as well as the sentence dialogue that the character said also used to create a new feature called sentence feature link through the character name. Each row in the dataframe represents each line or sentence that character said in the series. The sentence in each line is provided, constructing temporal analysis possible which means it allows us to create the visualization that is connected to the time period in the series from the length of the text and sentence. Finally, another benefit of transforming text file into dataframe allow us to perform the sentiment analysis in each sentence for each character in the dialogue which is an important part of answering research questions in this project.

4 Analysis

4.1 Approach

To answer the analytic question a visual analytics approach has to consider. It consists of computational algorithms to assist humans to gain better knowledge combined with judgement and reasoning by a human for replacing the part that the computer is unable to do.

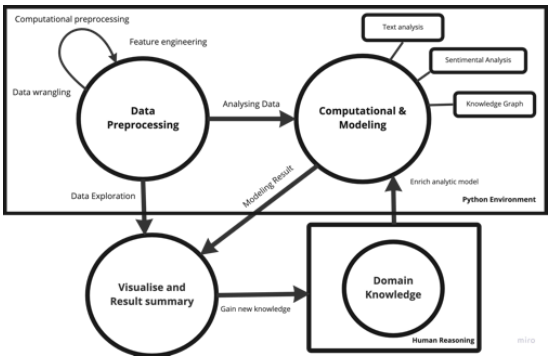

Fig. 2. Visual analytic approach diagram

The initial step for carrying on in this project since we obtain the raw text file data is data pre-processing. Regarding how we use different techniques to make it suitable for analysing and building Machine Learning models. This includes tokenization and lemmatization to utilize the full potential of the text data and leverage it to gain insight from the data when we do text analysis in the next further step. The next step after we derive our data into the proper format, we shall compute and analyse the data using machine learning approaches along with Natural Language Processing (NLP) due to the majority of our data is a string text so we can analyse through the frequency of the word or apply sentimental analysis which is a very popular tool in ML for analysing the emotional from the text allow us to evaluate how positive or negative of the sentence in the entire text [6]. The knowledge graph has been used to describe the relation and connection between each character. To create the network graph, the data need to be transformed in pre-processing step. The shortest distance between both characters' names will be defined as an interaction. Since we know the emotions in each sentence and the connection in each character, therefore, we can understand the pattern of the story or even the emotions, thoughts, and personalities of the characters. After we obtain satisfactory results from our data through modelling and computing the visualise step shall take part to present the result and finding in the form of graphs or images that represent the result and allow people to understand the direction of this project. Besides, using visual representations of the data allow people to understand the pattern easier by seeing than reading the whole text. Understanding the phenomenon from visualisation also leads to gaining a new set of domain knowledge which enables us to analyse and enrich our research. To ensure the robustness of our approach domain knowledge needs to apply along with another step during the research. Human reasoning and judgement play a role to sets the boundaries of our actions and keeping our research from failing. For example, when we get the result about sentiment score from each dialogue, we require a little bit of familiarity with the story and the personality of each character first to ensure and validate the result after computing by the program based on user domain knowledge The last step of this approach would be extracting the results of the analysis and addressing the research question. Report the finding as well as reflect on any areas of improvement and further study.

4.2 Process

After we obtain the data the initial step to do in order to understand the structure of data is to explore our data and preprocess unstructured data to make it more efficient for use in computational analysis [7].

Data pre-processing

In this step, we need to organize our dataset due to unwell format after we read from the text file. Word tokenization has been applied to split a chunk of text into smaller unit as well as Lemmatization which work as a text normalization that transforms a word to its base root mode. Converting all letters to lowercase, removing punctuation and stop stopwords also include in this step. Since we already prepare data in the proper format, we shall store it as a data frame because of the familiarly and manageable of computing for further analysis.

Initial exploration

We will start to explore and analyze the data to gain a better understanding of the properties of the dataset. After preprocessing data, it allows us to observe the frequency of the word in the entire text. To see the most use of the word in context word cloud serves as a popular tool to visualize the frequency of the text as we have seen in figure1. Notice that most of the word that appears in the dialogue is the name of characters because every dialogue must mention which character belongs. Therefore, prior knowledge and familiarity with this series play an importance part to extract the character name from the text file.

Lack of ability to present the actual number of frequencies instead of using font size based on their frequency of appearing [8] word clouds does not provide precise insight. Hence, the funnel chart in figure 3 was introduced to visualize the frequency based on the proportion of the dialogue of each character. The width of the bar is the number of dialogues in entire series, and the bar with the highest number of dialogues sits at the top.

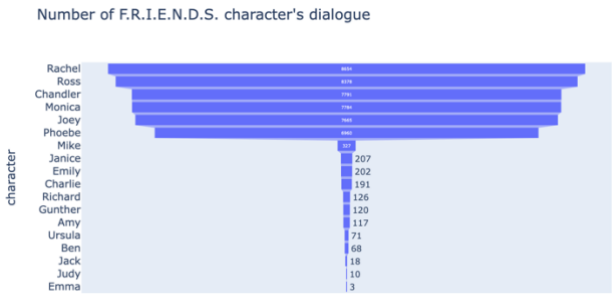

Fig. 3. Number of dialogues all characters

More than 95% of the dialogue through 10 seasons of F.R.I.E.N.D. was dominated by six main characters as we have some prior knowledge that this story is about six friends living and sharing their apartment together, other supporting characters appear sometime during the series and does not so

much impact the primary storyline, most of the story progresses around the main character.

To answer our curiosity about the proportion of the dialogue between main characters from the first analytic question, the pie graph in figure 4 needs to apply to demonstrate the Segmentation of dialogue in each character considering only main character. Each of the main characters has quite the same number of dialogues. Rachel is the highest number of dialogues compared to others which are typical because she is a protagonist since the series started in season one. Phoebe, however, has the lowest number in terms of dialogue because from the middle of the story to the end Phoebe doesn't live in the same apartment as the other main characters leading to less interaction in the end.

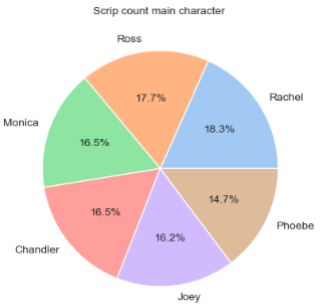

Fig. 4. Proportion of dialogue for main character

Supporting characters who play the role of adding color to the story, we can analyze by comparing their frequency of appearances with the main character from figure 5. Our six-protagonist appear in every episode from the beginning to the end of the season. However, some characters like Ursula who is Phoebe's sister, Gunther the owner of the coffee shop that the main protagonist is most likely to go or even ross's son Ben. All of these name as we can see in the figure, appear sometimes during the entire series shows that the story doesn't revolve around just only main characters.

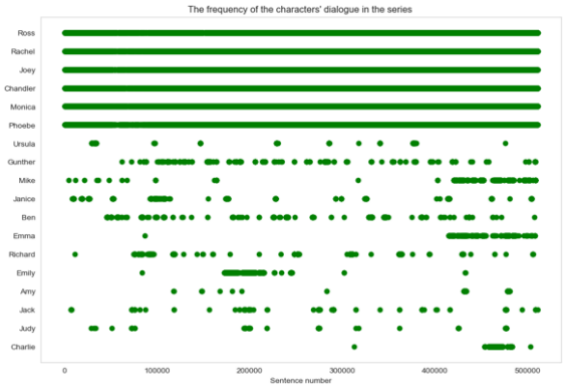

Fig. 5. Appearance of all characters in the story

Since we know that the dialogue script is not just only six main characters, we can do further analysis such as graph theory and network analysis to identify the connection between each character using the distance between each character's name represent the interaction.

Network graph character relationship

"A picture speaks a thousand words" It's a sentence that we are familiar with. Understanding the text through the frequency of the word occurrence sometimes might not be enough to provide insight from our data such as the relationship between one character to other characters.

A network graph which is the subset of Graph theory is used to illustrate the picture of the connection between each character in order for us to understand the story plot and the relationship between each character. The shortest distance between two characters' names in dialogue was used to describe the interaction between the pair of characters. Once we obtain all interactions of each character in the series, NetworkX the package in python has been employed to create, manipulate, and study the dynamic of the complex network graph.

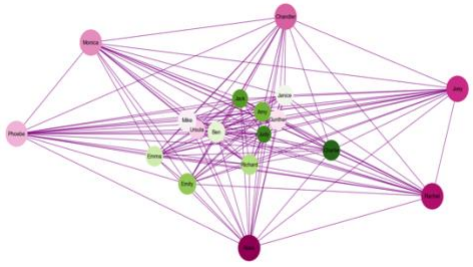

Fig. 6. Network graph F.R.I.E.N.D.S. relationship

Figure 6 is a network graph of F.R.I.E.N.D.S.'s character, we use the character names as nodes, and edges as an interaction between each character through the dialogue. Color means the significance of that character pink node is more important than green node. The more color intensity is the more dialogue and screen time they have, and node size also represents the number of dialogues using logarithm transformation to scale node size to make our main character node not too big than other nodes. From figure 6 we know that the main six-character nodes surround the support characters. It shows that our six main characters interact with every supporting character. At the same time, the supporting characters have relatively little interaction with each other.

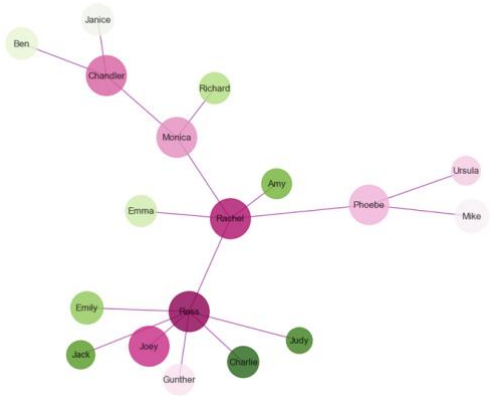

Fig. 7. Maximum spanning tree relationship network.

We expand our network by using the maximum spanning tree figure 7 to find the highest frequency that interacts between each character due to the network from figure 6 being a bit messy and cryptic to interpret. Centrality is used to identify the most important nodes in this graph, color and node size remain the same meaning. We found that Rachel is the center node of the graph as we mentioned that she is the main protagonist Emma and Amy who has the node-link are her daughter and sister respectively. Likewise Phoebe, Mike and Ursula are her husband and her sister. It is turn out that Jack and Judy who are the father and mother of Ross and Monica interact with Ross more than Monica because in the story they both love their children unequally. In the same way, Emily, and Charlie, who were once Ross's lovers, etc.

Apparently, using human knowledge merged with graph visualization allow us to describe the meaningful pattern in terms of relation to each character. However, as we stated earlier using a knowledge graph to represent the connection between each character still does not deliver the sentimental emotions of the character toward other characters that said there is still a lack of ability to extract profound story information from the dialogue.

Sentiment analysis from dialogue

Sentiment analysis applies to analyze the dialogue to fulfil the analysis that the network graph is unable to do. Besides, we can gain additional information by using these two methods to help us understand the nature of our dataset, detecting characters' emotional through their dialogue text. As we stored line by line along with character names, it acts as a social media post such as Twitter allowing us to apply sentiment analysis to detect how positive or negative a sentence said by that character is. VADER (Valence Aware Dictionary for Sentiment Reasoning) is the sentimental model that we use to predict the sentimental score in each dialogue to observe and classify the sentimental from a short text. It is widely used in Sentiment Analysis due to its reputation that even outperforms human observation by detecting sentimental in the sentence [9].

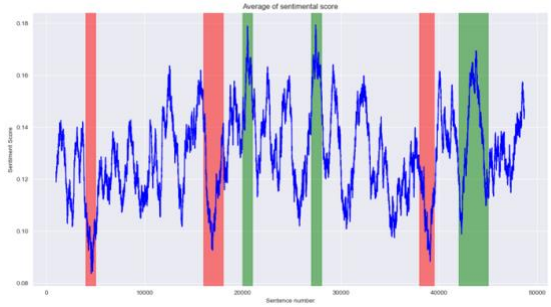

Fig. 8. Average sentiment score during the series.

The VADER score assign score from-1,1 which in this script the average score lies more than 0, that mean the tone of the story and the sentence which is said by character in the story quite positive. Meanwhile, the graph from figure 8 shows the fluctuations through time from start to finish which reflexes that these series have various emotional maybe we can assume that because the director of this series wants to create this series that does not look too bland, causing emotional ups and downs throughout the story.

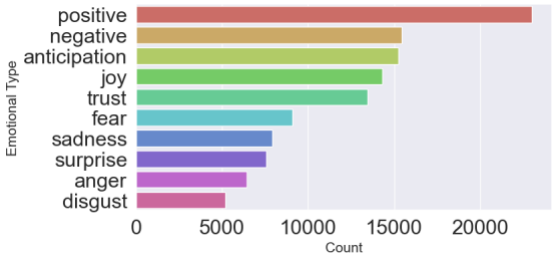

Fig. 9. Overall emotional sentiment in the story.

In addition, instead of providing only negative or positive which we cannot tell what kind of positive it is whether happy, joyful or surprise from the sentence, we can provide more profundity to our analysis by applying NRCLexicon by Mark M. Bailey to detect 10 varies emotions in the sentence which can see in figure 9. Obviously, the story consists of various emotional state which is correspond with sentiment score from VADER. F.R.I.E.N.D.S. TV series is a sitcom comedy show, therefore, not surprising that most emotions are more positive like joyful, anticipation, and trust than negative like fear, anger, and disgust.

This sentiment analysis allows us to understand the tone of the entire story. Nevertheless, we expected to see the emotional state of each character and answer the analytic question to identify their personality from the analysis the result of this we shall discuss in the next section.

4.3 Results

The pie chart from figure 4 can answer the first analytic question that there is the dialogue is divided evenly among the six main characters, but the most prominent is Rachel who is the main protagonist. Using the combination between network graphs and human reasoning to interpret the connection of each character provides fruitful insight. However, it has a limitation that it cannot reach in-depth into the feeling towards other characters.

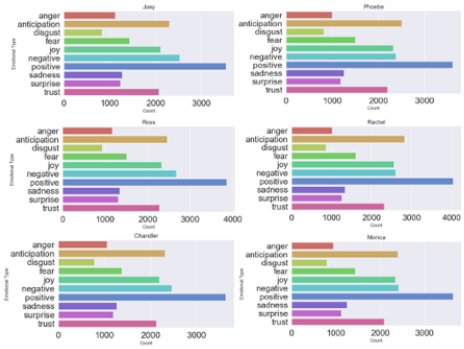

Fig. 10. Protagonist overall emotional sentiment.

For this reason, we applied VADER and NRCLexicon which is the well-known for sentiment analysis tools to observe the

emotions in dialogue once we know the character name in each sentence, we can extract the sentence and allow us to track the emotions in each sentence said by the characters precisely. Figure 10 shows that even different characters have seen not much difference in emotion. Causing unable to specify the personality of that character as expected through the emotion in the dialogue because every character has a mixture of emotions based on the true story that people have many sides.

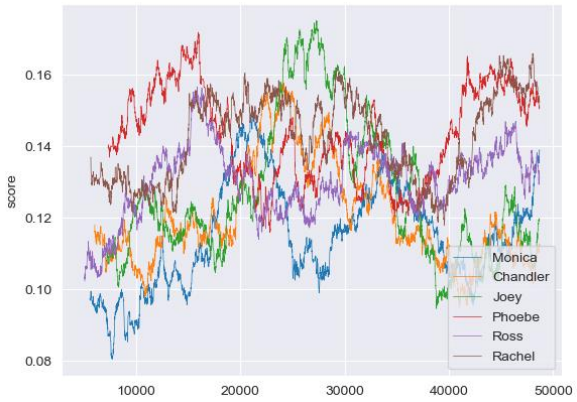

Fig. 11. Emotional sentiment score of the main character.

But if we observe throughout the time in the series, in figure 11 every protagonist in the story has different emotions depending on each moment which we can assume that main characters affect the tone of the story.

5 CRITICAL REFLECTION

In recent years, text analysis become widespread the use of natural language processing (NLP) and machine learning (ML) approaches like Sentiment analysis has become the field that growing rapidly in data science research areas [10]. Similarly, this research applied the above techniques to address the research questions. Despite the fact that there are not many different emotions in every main character, we can still interpret that because each character has many sides and consist of various emotions. From the observation when comparing the line graphs of each main character's sentiment score found that throughout the series different characters show different emotion Although we cannot provide an answer about identifying the character's personality through the dialogue in the study, we can summaries that it is not necessary that one character has only one main emotion and has only one specific personality just like people in the real world, they tend to get emotionally complex over time. The reason why it has happened is that the director of the series wants to make a comedy that is based on the real world, there is no villain or hero everything is a shade of grey. Otherwise, it may make the story seem bland and predictable.

Human reasoning plays a crucial part whether understanding the pattern of each character and pushing the research in the right direction such interpretation from the graph in Figure 11, shows that sentences in dialogue can express the emotion in each character over time. Each period has fluctuations due to the diversity of emotions. To interpret the story of what happened from the computation and why it was so. For instance, Joey and Phoebe got the higher positive score than others because both characters are hilarious and humorous throughout the film. It is possibly linked to their non-serious and playful nature for both. In contrast, Monica during the show tends to be serious more than every main character due to struggling with love, a job or even an organized personality. As a consequence, make the sentiment score from her sentence tend to be lower than other characters. These are the part that human knowledge and familiarity with the domain replace the stuff that computational programs unable to do. It is important to take into an account that domain knowledge is often required to confirm the veracity of results in the analysis

One thing to note is we can do further analysis by creating a network graph with edges and hue to represent the emotion toward other characters by analyzing through sentiment analysis to create a more meaningful visualization but with time constraints, therefore, this study is an initial point for other studies to inspire when working with text dialogue script whether it is movie or TV series using various techniques of text analysis approaches. In addition, these approaches are also applicable to other text domains due to people start to realize that sequential text documents now have become a bright and broad field to gain fruitful insight.

Table of word counts

	Expected	Actual
Problem statement	250	259
State of the art	500	476
Properties of the data	500	503
Analysis: Approach	500	488
Analysis: Process	1500	1509
Analysis: Results	200	209
Critical reflection	500	498
Total	**3950**	**3942**

REFERENCES

[1] Agrawal, D. (2020) *Friends TV show script, Kaggle*. Available at: https://www.kaggle.com/datasets/divyansh22/friends-tv-show-script (Accessed: December 23, 2022).

[2] Min S, Park J. (2019) *Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling.* PLoS One. Vol. 14(12). doi: 10.1371

[3] Park, S.-B., Oh, K.-J. & Jo, G.-S., 2012. *Social network analysis in a movie using character-net. Multimedia Tools and Applications*, 59(2), pp. 601-627. Doi: https://doi.org/10.1007/s11042-011-0725-1

[4] R. Rahman, M. Abdul Masud, R. Jahan Mimi and M. Nusrat Sultana Dina, (2020) "*Sentiment Analysis on Adventure Movie Scripts*," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp. 1-6, doi: 10.1109/STI50764.2020.9350525.

[5] Kansal, A. (2020) *Uragirii/Friends-generator*., GitHub. Available at:https://github.com/uragirii/Friends-Generator (Accessed: December 24, 2022).

[6] [Dataquest. (2022) *Tools for text analysis: Machine learning and natural language processing*, Dataquest. Available at: https://www.dataquest.io/blog/using-machine-learning-and-natural-language-processing-tools-for-text-analysis/ (Accessed: December 27, 2022).

[7] Kosaka, M. (2020) *Cleaning & preprocessing text data for sentiment analysis*, Medium. Towards Data Science. Available at: https://towardsdatascience.com/cleaning-preprocessing-text-data-for-sentiment-analysis-382a41f150d6 (Accessed: December 29, 2022).

[8] Alida. (2021) *The Pros and cons of word clouds as visualizations*, Alida. Vision Critical Communications Inc. Available at: https://www.alida.com/the-alida-journal/pros-and-cons-word-clouds-visualizations (Accessed: December 29, 2022).

[9] Hutto, C. J. & Gilbert, E., (2015). *VADER: A Parsimonious Rulebased Model for Sentiment Analysis of Social Media Text*. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, pp. 216-225.

[10] Mäntylä, M.V., Graziotin, D., Kuutila, M., (2018). *The evolution of sentiment analysis—A review of research topics, venues, and top cited papers*, Computer Science Review, Vol. 27, pp. 16-32. Doi: https://doi.org/10.1016/j.cosrev.2017.10.002.