INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

VOD and OTT Content as a Therapeutic Measure to Improve Self-esteem: A Retrospective Analysis of user Generated Opinion using NLP Techniques

Samuel Duraivel[1]*, **R Lavanya**[2]

1 Ph.D, Research Scholar, Department of Media Sciences, CEG Anna University, 600 025, Chennai
2 Assistant Professor, Department of Media Sciences, CEG Anna University, 600025, Chennai

Check for updates

OPEN ACCESS

Received: 21.09.2021

Accepted: 03.11.2021

Published: 07.03.2022

Citation: Duraivel S, Lavanya R (2022) VOD and OTT Content as a Therapeutic Measure to Improve Self-esteem: A Retrospective Analysis of user Generated Opinion using NLP Techniques. Indian Journal of Science and Technology 15(9): 395-407. https://doi.org/10.17485/IJST/v15i9.1753

* **Corresponding author**.

duraivelsamuel@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: [a] To understand how popular millennial media usage practices can be used to improve self-esteem among the millennial cohort; [b] To ascertain the effect of VoD and OTT streaming services in improving levels of self-esteem. **Methods:** Natural Language Processing techniques, namely, Term Frequency analysis, Network analysis, and Opinion Mining were used to identify and analyze the user generated Twitter Corpus dataset. Using Tweepy, a Python module, the data (n = 2566 tweets) were retrieved from the REST-API (Application Programming Interface) of Twitter. NLTK (Natural Language Processing ToolKit), NetworkX, and TextBlob were used for Opinion mining and visualization of the datasets **Findings:** Binge-watching of VoD and OTT content that address issues related to Body Positivity and Self-esteem resulted in improvement among the viewer cohort. Analysis of the Twitter corpus dataset implied that the following latent factors contributed toward the observed outcome: relatability (23:14%), upliftment(19.4%), gratitude(36.11%), and relief (21.2%). **Novelty:** Application of Natural Language Processing to identify latent factors that contribute toward either positive or negative outcome(s) has not been conducted before. Our previous research on similar areas identified latent factors that contribute toward behavioral outcomes in terms of attitude toward vaccines.

Keywords: Selfesteem; Bingewatching; Therapy; Netflix; Behavioral outcomes

1 Introduction

Recent developments in the field of Computational social sciences have made sure that this particular technique is applied to identify and solve problems that exist across various disciplines, especially in the field of behavioral sciences.

Natural Language Processing, a popular computational linguistic technique has been applied in the past to identify potential factors that contribute toward certain behavioral outcomes such as unwillingness to be vaccinated[1], etc. In this research, we attempted to extend the scope of this technique beyond identification of behavioral outcomes; here, we used NLP techniques namely, Term Frequency analysis, Network analysis, and Opinion mining to identify correlation between binge-watching using Video on Demand (VOD) and over-the-top (OTT) streaming services and self-esteem of the viewer cohort. In addition, we also identified the latent factors that contribute toward the observed effect.

To begin with, sentiment analysis or opinion mining is a natural language processing technique that involves finding and classifying opinionated parts of text data[2]. The application of sentiment analysis is more common in the field of Digital Marketing and social media analysis, where user opinion forms an integral part of the fact finding research - in order to develop effective solutions. More than anything, sentiment analysis as the name itself suggests, helps to identify the various sentiments and emotions expressed by users, which is a highly potential form of data from which excellent meaningful insights can be derived. In this research, various natural language processing techniques such as bigrams, network analysis, and sentiment analysis were applied to understand the effect of binge-watching mass media content on the self-esteem levels of the viewers. To be specific, we studied the effect of mass media contents that talk about issues related to body positivity, self-esteem, and body imagery on the viewers, using Natural Language Processing techniques.

1.1 Social media as a data resource

Social media platforms are an endless pool of unstructured text data, which when studied carefully can provide excellent insights. Discussion boards, forums, and groups specified to discuss selected topics, etc make social media platforms an excellent data source to harvest user opinion. However, many social media platforms do not allow or encourage this practice, Twitter lets users harvest Tweets (user generated texts) from its REST-API[3] Since only 280 characters can only be posted in a single tweet, tweets are a simple and effective way to express an opinion and the same makes them an effective source of data from which meaningful insights can be obtained. Twitter analysis has been very helpful to obtain the overall user opinion towards a specific product or an event - and it has found its application in the field of politics and diplomacy as well.

1.2 Social media analysis to understand human behavior

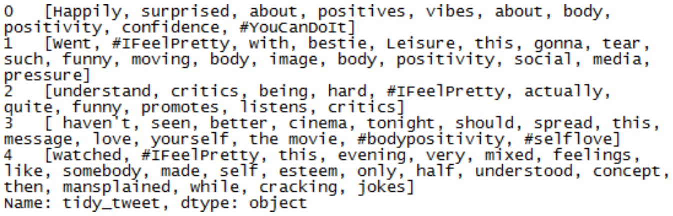
The application of Twitter analysis in the field of media effects and psychology is relatively new, although the practice has an enormous scope and the possibilities are seemingly endless. Besides, Twitter analysis has been used by media giants such as Netflix & Amazon Prime for content marketing - but, the application of Twitter analysis to understand the effects of media on the mental health of the viewers is unheard of at the moment[4]. The effects of mass media on the mental health of the viewers has been studied through various other techniques in the past, such as - surveys, experimental research, and content analysis. What makes Social media analysis using NLP techniques superior to any of these standard old practices is, social media analysis or Twitter analysis studies the data which is generated by the user itself, unlike surveys where the user answers the only questions asked by the researcher. Social media platforms let the users write whatever they want and therefore analysing such textual data comes with the perk of finding endless meaningful insights - given that the data is mined properly using appropriate techniques[5]. Since, human behavior is one of the complex areas of research where survey responses can easily be manipulated by the participant itself with a lot of false information and misrepresentation, analysis of user generated text data could be a better option.

2 Procedure

A list of mainstream media content which directly addressed the issues related to self-esteem and body positivity were identified, and the user generated text data (Tweets) about these content were harvested from the REST-API of Twitter. A random sample of 2566 tweets (n = 2566) on Sierra Burgess is a loser (n = 512), Amy Schumer's I Feel Pretty (n = 661), 13 Reasons Why (n = 897), and Little Miss Sunshine (n = 496) were retrieved from the REST-API (application programming interface) of Twitter using the OAuth method of Tweepy, a python library that is authorized to access the platform's API. Relevant key terms were used as addends to every search query while harvesting the data, so as to retrieve tweets that are relevant and related to the contents mentioned before. For instance, the search query to retrieve tweets (n = 661) on I feel Pretty is as follows: [search_query() = I feel pretty + Amy Schumer -filter:retweets]. Retweets were filtered out in order to keep the dataset devoid of redundancy.

2.1 Data processing

Sentiment analysis is a method of identifying attitudes in text data about a subject of interest. It is scored using polarity values that range from 1 to -1. Values closer to 1 indicate more positivity, while values closer to -1 indicate more negativity. In this research, sentiment analysis was applied to Twitter data using the python package textblob. Initially the tweets in each of the datasets were tokenized, that is split into tokens or multiple substrings, using the tokenize function of NLTK (nltk.tokenize.casual) as shown in Figure 1.



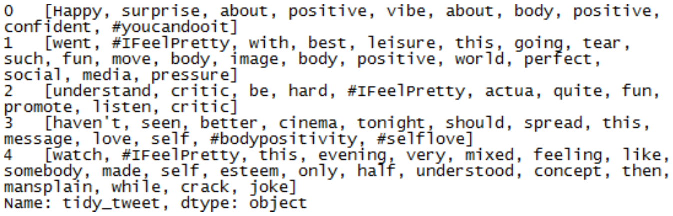
```
0    [Happily, surprised, about, positives, vibes, about, body,
positivity, confidence, #YouCanDoIt]
1    [Went, #IFeelPretty, with, bestie, Leisure, this, gonna, tear,
such, funny, moving, body, image, body, positivity, social, media,
pressure]
2    [understand, critics, being, hard, #IFeelPretty, actually,
quite, funny, promotes, listens, critics]
3    [ haven't, seen, better, cinema, tonight, should, spread, this,
message, love, yourself, the movie, #bodypositivity, #selflove]
4    [watched, #IFeelPretty, this, evening, very, mixed, feelings,
like, somebody, made, self, esteem, only, half, understood, concept,
then, mansplained, while, cracking, jokes]
Name: tidy_tweet, dtype: object
```

Fig 1. Sample of tokenized tweets. **Source: Twitter. Search query = #i feel pretty+amy schumer - filter:retweets, n = 661*

In Figure 1, it can be noted that each document is represented by a tuple (sentence, label). The objective of tokenizing the strings into substrings is to make the data set more favorable to be extracted into fine-grained sentiments, thereby predicting the polarity of the input with accuracy and precision.

2.2 Stemming and Lemmatization

The sentiment polarity of a tweet is the mean polarity of each word in the said tweet - for instance, the polarity of the tweet with index 0 that reads, "Happily surprised about positive vibes about body positivity and confidence, #you can do it" is the mean polarity [m] of the words [p(Happily) + p(surprised) + p(about) + p (body) + p(positivity) + p(and) + p (confidence)]. Further, the tokenized tweets were reduced to stems or distinct word forms, in order to reduce the vocabulary size, thereby sharpening the polarity scores of sentiment analysis. The process involves the reduction of a word to its word stem, which affixes to prefixes and suffixes or to the roots of words, also called a lemma. This procedure is one of the principal components in natural language processing and is an inevitable part of linguistic studies in morphology and artificial intelligence (AI) information retrieval and extraction. The lemmatized tweets are shown in Figure 2.



```
0    [Happy, surprise, about, positive, vibe, about, body, positive,
confident, #youcandooit]
1    [went, #IFeelPretty, with, best, leisure, this, going, tear,
such, fun, move, body, image, body, positive, world, perfect,
social, media, pressure]
2    [understand, critic, be, hard, #IFeelPretty, actua, quite, fun,
promote, listen, critic]
3    [haven't, seen, better, cinema, tonight, should, spread, this,
message, love, self, #bodypositivity, #selflove]
4    [watch, #IFeelPretty, this, evening, very, mixed, feeling, like,
somebody, made, self, esteem, only, half, understood, concept, then,
mansplain, while, crack, joke]
Name: tidy_tweet, dtype: object
```

Fig 2. Sample of lemmatized tweets. ** Source: Twitter. Search query = #i feel pretty+amy schumer - filter:retweets, n = 661*

As shown in Figure 2, the substrings in the tokenized Twitter corpus data set are reduced to their corresponding stems, that is, for instance, the words in the tweet with index 0 are reduced from 'Happily' to its basic form, which is 'Happy', 'surprised' to 'surprise', 'positives' and 'positivity' to 'positive', and so on. This would enable the Natural language processing dictionary to process the words much effectively with precision and thereby synthesizing polarity with accuracy. In this analysis, lemmatization was preferred over stemming because stemming reduced words to unrecognizable forms, in other words, to word forms that were not actually readable words, and use of which would lead to inaccuracies and skew the analysis.

3 Findings and Results

3.1 Descriptive Analysis

The dataset on Sierra Burgess is a loser contained 512 tweets (n = 512), of which 23 profiles (4.49%) self-identified as men, 382 (74.60%) as women, and 107 (20.89%) did not wish to specify; the dataset on Amy Schumer's I Feel Pretty contained 661 tweets (n = 661), of which 18 identified as men (2.72%), 564 identified as women (85.32%), and 79 were unspecified (11.95%); the dataset on 13 Reasons Why had 897 tweets (n = 897), of which 288 identified as men (32.10%), 491 identified as women (54.73%), and 118 did not specify their gender (13.15%); similarly, the dataset on Little Miss Sunshine had 496 tweets (n = 496) of which, 9 self-reported as men (1.81%), 403 (81.25%) as women, and 84 did not specify (16.93%). Since the standard Tweepy API search has limitations in regard to scraping a large number of tweets and access to tweets older than a week, more number of Tweets could not be harvested. In summary, the dataset (n = 2566) had 338 tweets (13.17%) from profiles that self-identified as men, 1840 (71.7%) tweets from profiles that self-reported as women, and 388 (15.12%) tweets from Twitter users who did not wish to specify their gender. Table 1 shows the proportion of Twitter users in the data set based on their gender.

Table 1. Proportion of user based on their Gender

	Sierra Burgess	I feel Pretty	13 Reasons Why	Little Miss Sunshine
Male	23	18	288	9
Female	382	564	491	1840
N/A	107	79	118	84
n	512	661	897	496
n-w(g)	405	582	779	412

* N/A = Gender not specified. n-w(g) = number of tweets after removal of profiles without gender

3.2 Word Frequency Analysis

Counting the frequency of specific words in the list can provide illustrative data. The frequency of the most common words in each of the four datasets were studied using a bag of words (BoW) model, which is a way of extracting features from text for use in modeling, such as with machine learning algorithms[(6)]. In other words, the bag of words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity[(7)]. After the initial preprocessing or 'text cleanup' were URLs and special characters were replaced with 'nothing', the words in the datasets were made lowercase using the string method. lower(), since Python is a case sensitive language and treats words with capitalization different from words that are all lowercase, which would result in inconsistency and lack of precision in the analysis. The words in each tweet were then split into unique elements, so as to represent a mere 'bag of words' and not a set of meaningful tweets. In order to get the count of how many times each word appears in the sample dataset, the built-in Python library - 'collections', was used to create a special type of dictionary. Thereafter, using the functions 'collection. counter' and 'most_common', the most commonly used words and the number of times that they were used were returned. Further, the words were structured into a data frame for analysis and were plotted into a horizontal bar graph for visual representation. The results of word frequency analysis of the dataset on Sierra Burgess is a loser (n = 512) is displayed in Figure 3.

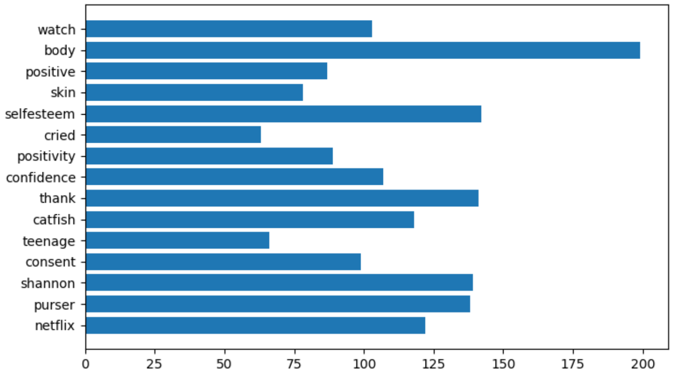


Fig 3. Common Words Found in Tweets on Sierra Burgess

As shown in Figure 3, the most common words in the dataset are as follows - Netflix, 122; purser, 138; shannon, 139; consent, 99; teenage, 66; catfish, 118; thank, 141; confidence, 107; positivity, 89; positive, 89; cried, 63; selfesteem, 142; skin, 78; positive, 87; body, 199; watch, 102. The most common word in the dataset is 'body' with a frequency of 199 times, followed by 'self esteem' with a frequency of 142 times. In the dataset, 'self esteem' was found in different variations, such as 'self-esteem', 'self esteem', and such, thus each form was neutralized and made into the form as shown in the Figure. Also, the term 'body' with the highest frequency co-occurred with either 'positive' or 'positivity', which would be 'body positive' or 'body positivity' as a phrase. 'Purser' and 'Shannon' co-occurred as 'Shannon Purser', the actor who played Sierra Burgess. The term 'catfish' relates to the storyline or the plot, where low self-esteem Sierra goes on a catfishing spree on her crush by using her friend as a proxy. Besides, the collections words, that is, the words which were used in the search query (search_query = sierra+burgess+loser : filter-retweets) were removed as they would be present in every tweet and skew the frequency analysis. In addition to the collection words, the stop words in the dataset were removed using python's Natural Language Processing Tool Kit (NLTK). In computing, stop words are words which are filtered out before or after processing of natural language data (text), prepositions - for instance. Figure 4 lists the most common words in the tweets on Amy Schumer's I Feel Pretty.

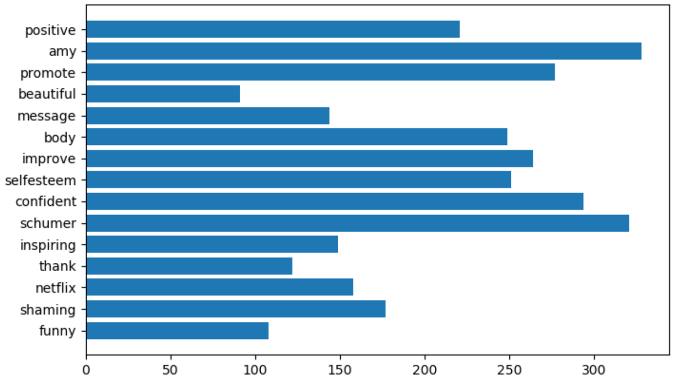


Fig 4. Common Words Found in Tweets on I Feel Pretty

As shown in Figure 4, the most common words in the tweets of Amy Schumer's I Feel Pretty are as follows - funny, 108 times; shaming, 177; netflix, 158; thank, 122; inspiring, 149; schumer, 321; confident, 294; self esteem, 251; improve, 264; body, 249; message, 144; beautiful, 91; promote, 277; amy, 328; positive, 221. The most frequent word in the dataset was 'amy' with a frequency of 328 times, which co-occurred with Schumer, which has a frequency of 321 times. Besides, the term 'confident' is a result of lemmatized contraction of its various forms, for instance - the noun confidence which was stemmed to 'confident' in order to stop splitting of frequencies, perhaps enough to skew or disturb the precision of the analysis. Similar to that of the results in Sierra Burgess is a loser, the term 'body' co-occurred with 'positive' and 'positivity', but with an addition, 'shaming'. Also, the term 'thank' co-occurred with 'you' which was removed by NLTKs list of stopwords. Besides, as a phrase, 'thank you' was followed by 'Amy Schumer'. Figure 5 lists the most common words in the tweets on 13 Reasons Why.

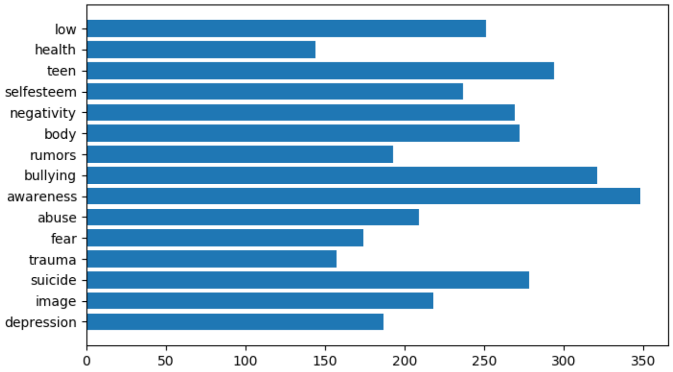


Fig 5. Common Words Found in Tweets on 13 Reasons Why

Unlike the analysis of tweets on Sierra Burgess is a loser and I Feel Pretty, the analysis of tweets on 13 Reasons Why required the addition of a set of new words to the list of NLTKs stopwords, as the ten most frequent words in the data were found to be name of the characters in the series. Therefore, the names of the characters which include, but are not limited to, 'hannah+baker', "clay+jensen", "justin+foley", and "jessica + davis" were appended as addends to the list of stopwords. Also, the search query was modified from 'search_query = 13+reasons+why -filter:retweets' to 'search_query = 13+reasons+why+mental+health -filter:retweets', as the original search term did not yield relevant results.

With the altered query, tweets that were relevant to the objective of the research, that is, the therapeutic effect of binge-watching, were retrieved from the API of Twitter. Thus, as shown in Figure 6, the most common words in the tweets on 13 Reasons Why are as follows - depression, 187 times; image, 218 times; suicide, 278; trauma, 157; fear, 174; abuse, 209; awareness, 348; bullying, 321; rumors, 193; body, 272; negativity, 269; self-esteem, 237; teen, 294; health, 144; low, 251. Despite the high volume of tweets (n = 897), the results included a lot of other factors related to mental health, in addition to self-esteem and body positivity, as the series talks about a wide range of complex issues ranging from suicide to body imagery. However, the effect of 13 Reasons Why on self-esteem was studied by network analysis of words through construction of a co-occurrence matrix. Figure 6 lists the results of frequency analysis of tweets on Little Miss Sunshine.

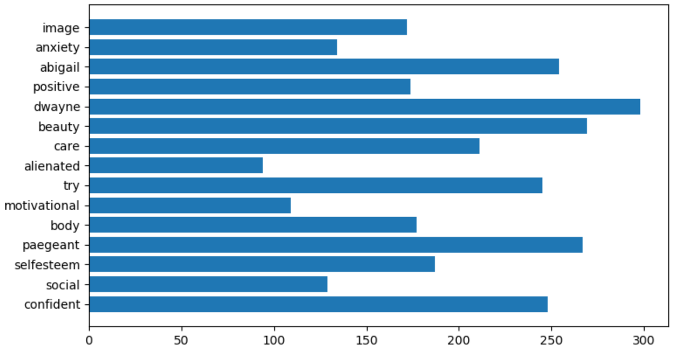


Fig 6. Common Words Found in Tweets on Little Miss Sunshine

As shown in Figure 6, the most frequent words in the tweets on Little Miss Sunshine are as follows - confident, 248 times; social, 129; selfesteem, 187; pageant, 267; body, 177; motivational, 109; try, 245; alienated, 94; care, 211; beauty, 269; dwayne, 298; positive, 174; abigail, 254; anxiety, 134; image, 172. Unlike the usual OAuth method of Tweepy, which was used to retrieve tweets from the REST-API, the datasets on Little Miss Sunshine were harvested using 'GetOldTweets3' - a Python 3 library and a corresponding command line utility for accessing old tweets. Although accessing the Twitter API using this method is not always recommended, the usual OAuth method did not return a significant volume of tweets, as it does not scrape tweets older than a week, and the majority of tweets on Little Miss Sunshine were posted little more than a year ago during the time of retrieval. Further, the search query was modified with an addend, 'mental+health', similar to that of the query that was used to retrieve tweets on 13 Reasons Why, as the original query, 'search_query = little+miss+sunshine -filter: retweets' returned diversified results, of which most did not relate to the research objective. Thus, the results of word frequency analysis on the four datasets, namely, Sierra Burgess is a loser, I Feel Pretty, 13 Reasons Why, and Little Miss Sunshine, with a sample sum of 2566 tweets (n = 2566) indicate that the contents had a direct and immediate positive effect on the following factors related to mental health, which include but are not limited to, body image issues, depression, self-esteem, and confidence, although it is assumed that the effects are short lived. In other words, the overall opinion of viewers with respect to mental health, as expressed on Twitter, seems to be 'overall-positive', as the list of identified frequent words in the data set contains key terms of the research area such as depression, self-esteem, and anxiety, and body positivity.

3.3 N-Grams and Network Analysis

In order to study the co-occurence of the words identified through frequency analysis, thereby understanding the association between the factors, the data were studied using network analysis of bigrams by plotting the co-occurrence matrices, on a character co-occurrence network diagraM[8]. In general, a bigram is a sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words. A bigram is an n-gram for n=2. The frequency distribution of every bigram in a string is commonly used for simple statistical analysis of text in many applications, including in computational linguistics, cryptography, speech recognition, and so on[9]. Using library-bigrams from the Natural Language Processing Tool Kit (NLTK), the co-occurence of words in the dataset were identified, which were captured by a 'collection counter' that holds the 'word-pairs' as keys of the dictionary and their corresponding counts as the values. The dictionary was structured into a data frame, in order to visualize the top 20 occurring bigrams as networks using the Python package NetworkX. The plot in Figure 7 displays the network of co-occurring words in the tweet dataset on Sierra Burgess is a loser.

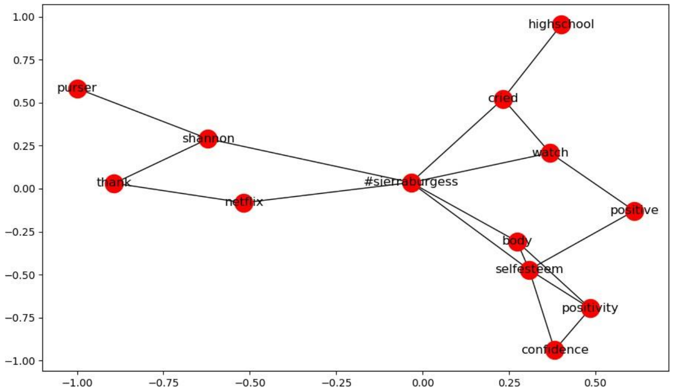


Fig 7. Network of Bigrams in tweets on Sierra Burgess is a loser

The plot displayed in Figure 7 displays the correlation between the bigrams, where it can be seen that the largest single node #sierraburgess connects with the key terms observed in the frequency analysis, such as "cried" and "watch", which as a phrase would roughly mean, "cried while watching #sierraburgess". The most frequent bigram in the dataset is #sierraburgess + positivity, which has a co-occurrence frequency of 127 times, followed by body + positive, which has a co-occurrence rate of 98 times. Besides, a chain of gratitude can be observed in the plot. #sierraburgess has co-occurred with 'netflix', while 'netflix' has co-occurred with the term 'thank'; 'thank' has co-occurred with 'shannon' and 'shannon' has co-occurred with "purser". This would mean that the viewers had expressed a sense of gratitude towards the content, character, and the actor who played Sierra Burgess - Shannon Purser. Thus, the results of the network analysis of tweets on Sierra Burgess indicate a strong correlation between the content and its effect on the self-esteem levels of the subjects, or in other words, the viewers. The following section briefs on the results of the analysis of tweets on Amy Schumer's I Feel Pretty. Figure 8 displays the network of co-occurring words in the dataset (n = 661).

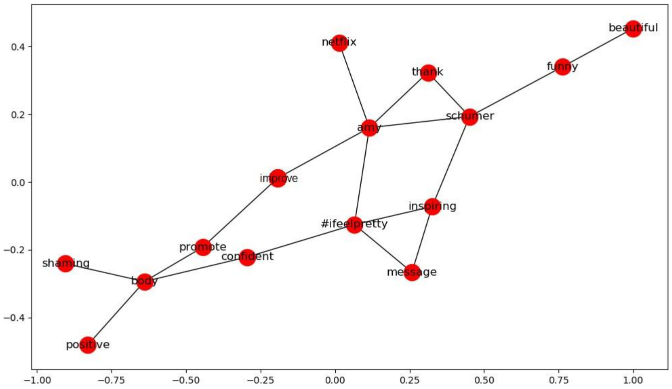


Fig 8. Network of Bigrams in tweets onI Feel Pretty

The most frequent bigrams observed in the Twitter corpus dataset on I Feel Pretty, as shown in Figure 3.6 include, but are not limited to, (#ifeelpretty, inspiring) 51 times; (#ifeelpretty, message) 39 times; (#ifeelpretty, amy) 66 times; (#ifeelpretty, confident), 31 times; (body, positive) 59 times; and (body, shaming) 26 times. Besides, the most frequent bigram was (amy + schumer), 189 times - which has been muted to highlight the bigrams that reflect the effect, rather than bigrams with null values. The overall observation of the co-occurrence network, shows a positive correlation between Amy Schumer's I Feel Pretty and body positivity, confidence, and self-esteem. Similar to that of the network analysis results of Sierra Burgess is a loser, a 'chain of gratitude' is observed in the dataset corresponding to I Feel Pretty as well, where the words have co-occurred in the following pattern - (amy, thank), (schumer, thank), (netflix, thank), or (thank, netflix). Thus it is concluded that the content had a positive effect on the viewers, with respect to self-esteem and body image, the key factors of the research objective. Similarly, the tweets on 13 Reasons Why (n = 496) were studied through the analysis of bigrams in the dataset. Figure 9 displays the network of co-occurring words.

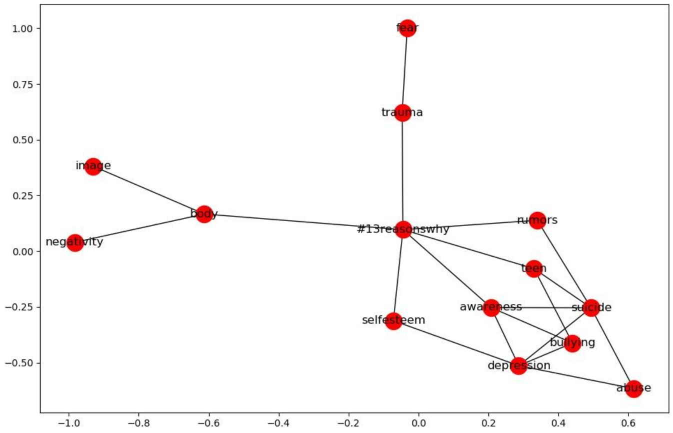


Fig 9. Network of bigrams in tweets on 13 Reasons Why

The frequent bigrams observed in the tweets on 13 Reasons Why seem to address a wide range of issues, in addition to self-esteem and body image, so does the series. In Figure 9, it can be noted that #13reasonswhy has been associated directly with self esteem, awareness, and 'body' which is further associated with 'image'. 'Awareness' is paired with 'depression', which would be 'depression awareness', likewise, 'awareness' + 'bullying', which would be 'bullying awareness', and 'suicide awareness'. Although a significant number of 'direct effects' can not be identified between self-esteem and 13 reasons why owing to the presence of high number of other factors in the dataset, overall observations reveal that the content had a positive effect on the subjects in terms of mental health, - that is on depression, suicide awareness, and bullying; however, the effect on self-esteem and body positivity are limited as shown in Figure 9, if at all. This could be attributed to the fact that 13 Reasons Why addresses a wide range of issues, in addition to body positivity and self-esteem, unlike Sierra Burgess is a loser and I Feel Pretty, where the said factors were pivotal parts of the plot/storyline. The following section briefs on the results of the analysis of tweets on Little Miss Sunshine. Figure 10 displays the network of co-occurring words in the dataset (n = 496).

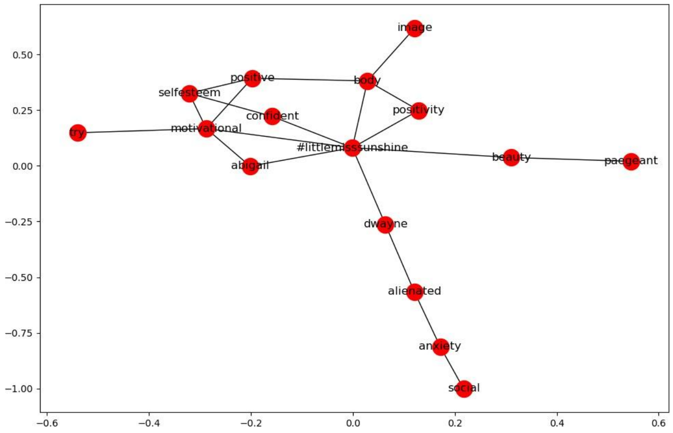


Fig 10. Network of bigrams in tweets onLittle Miss Sunshine

The observation of the network plot as displayed in Figure 10 reveals that #littlemisssunshine had co-occurred with 'body' 44 times; 'positivity, 32; whereas 'body' & 'positivity' had co-occurred 35 times. Further, #littlemisssunshine had co-occurred with motivational with a frequency of 21 times. In addition, the lower chain of the bigram network that extends through the nodes 'dwayne', 'alienated', anxiety', and 'social', indicates that the dataset contains tweets on social anxiety as well, as these mentioned nodes represent the socially anxious teenager, Dwayne Hoover, played Paul Dano, in Little Miss Sunshine. Thus, the analysis of bigrams in the dataset suggests that the overall opinion of the viewers of Little Miss Sunshine, with reference to self-esteem and body positivity are positive, as the chain of words displays a pattern of positivity in the text dataset.

3.4 Prediction of Sentiments and Polarities

The polarity value for each tweet on a given subject was calculated and then plotted in a histogram to identify the overall sentiments toward the subject of interest. The following section discusses the results of the sentiment analysis of tweets on Sierra Burgess is a loser (n = 512). Table 2 provides an overview on the prediction of polarity for tweets in the dataset.

Table 2. Sentiment scores of sample tweets from the dataset

Tweet	Polarity	Subjectivity
"this sierra burgess movie on netflix is good perfect self esteem bully movie for all to watch"	0.85 (as estimated by Python's textblob sentiment Reasoner)	0.8
"Sierra Burgess is a loser is a terrible movie and we should not celebrate a plot line that revolves around low self esteem being an excuse for being a shitty person and still getting the fairytale ending thanks for coming to my ted talk"	-0.2125 (as estimated by Python's textblob sentiment Reasoner)	0.3875
"Sierra Burgess is a loser - felt weird and also romantic. I feel they just had to end it on a good note but damn, catfishing is a huge turnoff"	0.15 (as estimated by Python's textblob sentiment Reasoner)	0.75
Sierra Burgess Is A Loser has me in an emotional mess. I laughed, cried and got angry with quickfire emotions. Highly relatable teen movie for anyone that has felt like they weren't/aren't good enough. Noah Cent and Shannon Purser - well done! Great casting.	0.103 (as estimated by Python's textblob sentiment Reasoner)	0.446

*(search query = # sierra+burgess+is+a+loser :filter - retweets, n = 118, Source: Twitter)

The polarity and degree of subjectivity for each tweet was studied using Python's textblob, and the mean polarity, or in other words, the compound sentiment score of the dataset was used to predict the overall sentiment of the viewers towards the subject, namely Sierra Burgess is a loser with respect to self-esteem and body image. The plot in Figure 11 displays a histogram of polarity values for tweets in the dataset (n = 512).

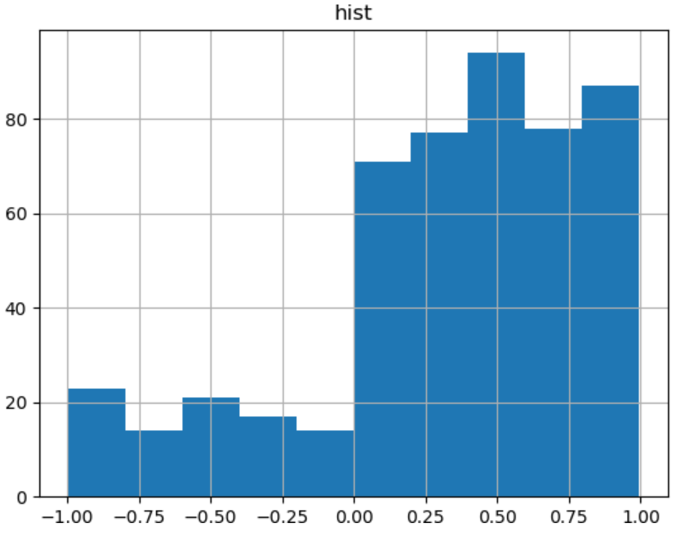


Fig 11. Sentiments from tweets on Sierra Burgess is a loser

In Figure 11, the polarity values equal to zero have been removed, and a break has been added at zero, to better highlight the distribution of polarity values. The dataset (n = 512) has 407 positive tweets (79.49%), 89 negative tweets (17.38%), and 16 neutral tweets (3.12%). Further, the distribution of polarity is extremely right skewed, that is more positive, which indicates that the opinion expressed by the subjects on Sierra Burgess is a loser with respect to self-esteem and body image is extremely positive, despite a minimal number of negative tweets. When compared with the results of the word frequency analysis and network analysis of the bigrams where a series of positive impact indicators were identified, the results of the sentiment analysis confirm that Sierra Burgess is a loser indeed improves the self-esteem levels in the viewers, given that the observed effect is susceptible to change based on the degree of acceptability and perceivability of the user towards the content. The following section presents the results of the sentiment analysis of tweets on Amy Schumer's I Feel Pretty. The polarity of each tweet in the dataset was calculated as the mean polarity of the lemmatized contraction of words in each tweet. The plot in Figure 12 displays a histogram of polarity values for tweets in the dataset (n = 661).

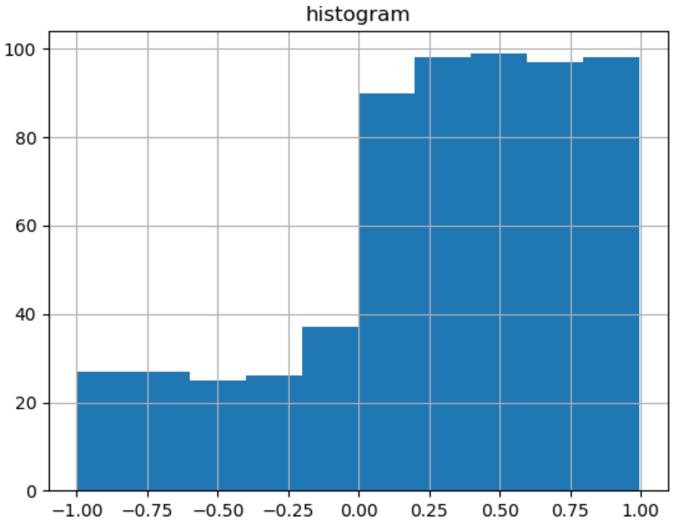


Fig 12. Sentiments from tweets on I Feel Pretty

The dataset (n = 661) contained 482 positive tweets (72.91%), 142 negative tweets (21.48%), and 37 neutral tweets (5.59%). In Figure 13, the polarity values equal to zero have been removed (the neutral tweets), and a break has been added at zero, to better highlight the distribution of polarity values. Further, similar to that of the sentiment analysis results of Sierra Burgess is a loser, the histogram of polarity corresponding to I Feel Pretty is extremely right skewed as well, indicating that the opinions expressed by the subjects in the dataset with respect to self-esteem and body positivity are significantly more positive than negative. When compared with the results of the word frequency analysis and network analysis of the bigrams where a series of positive impact indicators were identified, the results of the sentiment analysis confirm that Amy Schumer's I Feel Pretty indeed improves self-esteem in the viewers, given that the observed effect is susceptible to change based on the degree of acceptability and perceivability of the user towards the content, whilst the longevity of the effect remains unknown. The forthcoming section details the results of sentiment analysis of tweets on 13 Reasons Why dataset. The plot in Figure 13 displays a histogram of polarity values for tweets in the dataset (n = 897).

The dataset (n = 897) had 503 positive tweets (56.07%), 173 negative tweets (19.28%), and 221 neutral tweets (24.63%). In Figure 4.32, the polarity values equal to zero have been removed (the neutral tweets), and a break has been added at zero, to better highlight the distribution of polarity values. The histogram of polarities corresponding to the tweets referring to 13 Reasons Why with respect to self-esteem are right skewed, with a mean (m) of 0.260 and a standard deviation of 0.533, indicating that the results are more positive than negative, although a significant percentage of tweets are neutral and negative. The results of the frequency analysis and the network analysis of bigrams indicated that the dataset on 13 Reasons Why contains a lot of other factors referring to mental health in addition to self-esteem, which include, but are not limited to depression and recovery. Thus it is concluded 13 Reasons Why improves mental health in the viewers, given that a significant percentage of tweets indicate that the effect is reverse, although the percentage of this observation is minimal (19.28%). The following section briefs on the results of the sentiment analysis of tweets on Little Miss Sunshine. The plot in Figure 14 displays a histogram of polarity values for tweets in the dataset (n = 496).

The dataset contained 496 tweets (n = 496) of which 38 were identified as positive (681.4%), 75 were identified as negative (15.12), and 83 were identified as neutral (16.73) by the 'textblob' sentiment analyzer of Python. In Figure 4.33, the polarity values equal to zero have been removed (the neutral tweets), and a break has been added at zero, to better highlight the distribution of polarity values. The histogram of polarities are right skewed, which indicates that the tweets are more positive

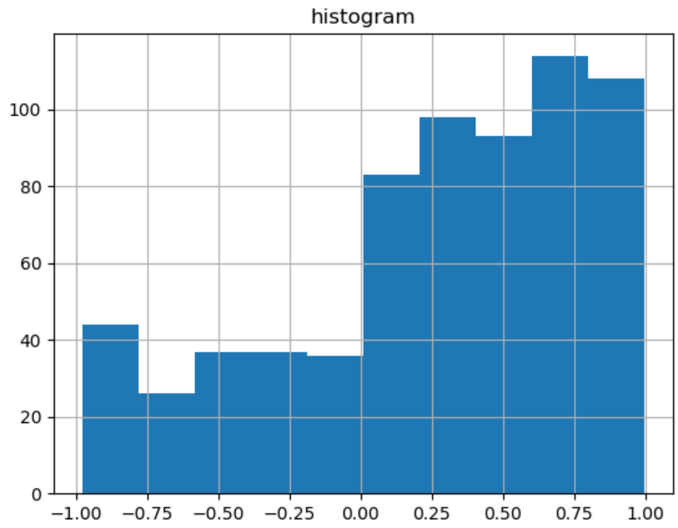


Fig 13. Sentiments from tweets on 13 Reasons Why

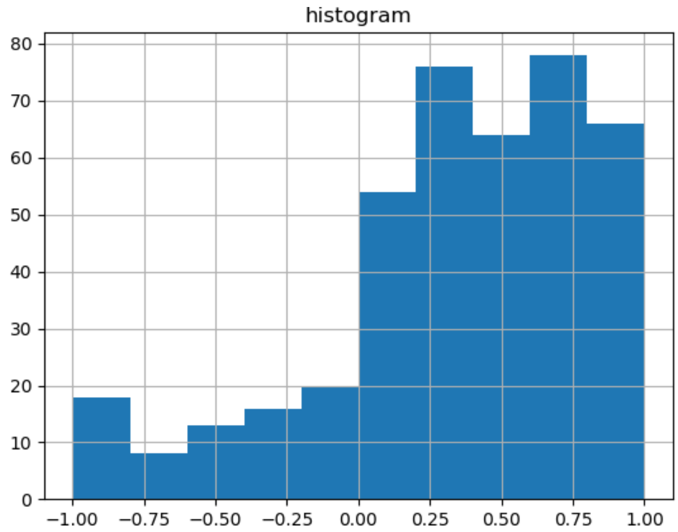


Fig 14. Sentiments from tweets on Little Miss Sunshine

than negative with 16.73% of neutral tweets in the dataset. When compared with the results of the frequency analysis and network analysis of bigrams, the results of the sentiment analysis indicate that Little Miss Sunshine indeed improves self-esteem levels in the subjects with respect, although the longevity and proactivity of the approach remain unknown.

3.5 Identification of positive impact indicators

The positive causal effect was ascertained by systematized identification of potential impact indicators in the linguistic data. The tweets in each of the four datasets that were identified as positive statements by the sentiment analyzer (n = 1730) based on polarity were further studied manually and grouped into four categories of positive impact indicators or factors, namely, namely relatability with the content, sense of upliftment, sense of gratitude, and sense of relief. The tweets were classified into any of the four best fitting categories, as shown in Table 3.

The factors listed in Table 3 were assumed by studying the tweets manually in addition to analysis of the most frequent hashtags in the dataset. In general, the use of a Hashtag in a Tweet is considered to reflect the overall opinion expressed in the said Tweet, and therefore the most frequent Hashtags were used to classify or group the positive tweets (n = 1730) into the aforementioned categories. Also, in order to ensure that the tweets fit into the right category of identified positive

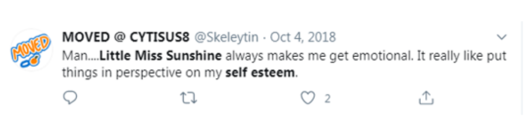
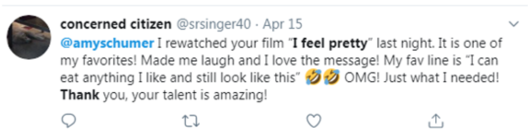
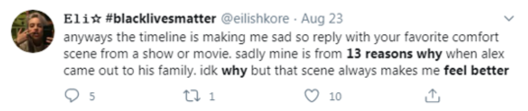
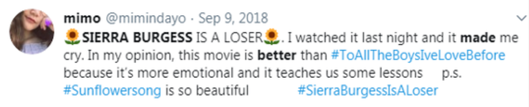
Table 3. Positive impact indicators (n = 1730)

Relatability	Upliftment	Gratitude	Relief
23.14% (n = 401)	19.4% (n = 331)	624 (n = 36.11%)	21.2% (n = 366)

*Here n represents the number of positive tweets identified by the sentiment analyzer

impact indicators, the tweets were studied using a NLP Thesaurus so as to improve the accuracy and precision of the overall categorization procedure. Table 4 displays a sample of tweets and their assumed categories.

Table 4 Sample tweets with assumed categories

Tweet	Category
 **MOVED @ CYTISUS8** @Skeleytin · Oct 4, 2018 Man....**Little Miss Sunshine** always makes me get emotional. It really like put things in perspective on my **self esteem**. 2	Relatabilit y
 **concerned citizen** @srsinger40 · Apr 15 @amyschumer I rewatched your film "**I feel pretty**" last night. It is one of my favorites! Made me laugh and I love the message! My fav line is "I can eat anything I like and still look like this" 🤣🤣 OMG! Just what I needed! **Thank** you, your talent is amazing!	Gratitude
 **Eli☆ #blacklivesmatter** @eilishkore · Aug 23 anyways the timeline is making me sad so reply with your favorite comfort scene from a show or movie. sadly mine is from **13 reasons why** when alex came out to his family. idk **why** but that scene always makes me **feel better** 5 1 10	Upliftment
 **mimo** @mimindayo · Sep 9, 2018 🌻**SIERRA BURGESS** IS A LOSER🌻. I watched it last night and it **made** me cry. In my opinion, this movie is **better** than #ToAllTheBoysIveLoveBefore because it's more emotional and it teaches us some lessons p.s. #Sunflowersong is so beautiful #SierraBurgessIsALoser	Relief

4 Conclusion

A sum of 2566 tweets (n = 2566) on the following content, namely, Sierra Burgess is a loser (n = 512), I Feel Pretty (n = 661), 13 Reasons Why (n = 897), and Little Miss Sunshine (n = 496), were retrieved from the REST-API of twitter and analysed for word frequencies, n-grams of networks, and sentiment analysis. The results indicated that the opinion of the viewers was overall positive with an aggregate compound sentiment score of 0.7360. Tweets which were identified as statements by the Textblob sentiment analyzer were classified into four categories of positive impact indicators (text categorization), namely relatability with the content, content, upliftment, gratitude, and relief. Also, descriptive analysis of data revealed that the prevalence of an unequal and a very high percentage of Tweets from the female users indicates or leads to an assumption that the gender of the characters and characterization might be a key component the observed effect - that is binge-watching mass media content related to body positivity and self-esteem improves the self-esteem levels of the viewers, given that, the users are able to relate with the characters in terms of Gender. Thus content specific binge-watching as a stimuli improves self-esteem levels through triggering either of these factors in the subject namely, relatability, upliftment, gratitude, and relief.

References

1) Duraivel SS. *Understanding vaccine hesitancy with application of Latent Dirichlet Allocation to Reddit Corpora*. 2021. doi:10.21203/rs.3.rs-616664/v1.
2) Nemes L, Kiss A. Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*. 2021;5(1):1–15. Available from: https://dx.doi.org/10.1080/24751839.2020.1790793.
3) Namugera F, Wesonga R, Jehopio P. Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. *Computational Social Networks*. 2019;6(1). Available from: https://dx.doi.org/10.1186/s40649-019-0063-4.
4) Ahmed MS, Aurpa TT, Anwar MM. Detecting sentiment dynamics and clusters of Twitter users for trending topics in COVID-19 pandemic. *PLOS ONE*. 2021;16(8):e0253300–e0253300. Available from: https://dx.doi.org/10.1371/journal.pone.0253300.
5) Schuh G, Reinhart G, Prote JP, Sauermann F, Horsthofer J, Oppolzer F, et al. Data Mining Definitions and Applications for the Management of Production Complexity. *Procedia CIRP*. 2019;81:874–879. Available from: https://dx.doi.org/10.1016/j.procir.2019.03.217.

6) Liu C. Word Frequency Analysis and Intelligent Word Recognition in Chinese Literature Based on Neighborhood Analysis. Springer - Application of Intelligent Systems in Multi-modal Information Analytics. *Springer - Application of Intelligent Systems in Multi-modal Information Analytics*. Available from: https://doi.org/10.1007/978-3-030-51431-0_73.
7) Bérubé N, Sainte-Marie M, Mongeon P, Larivière V. Words by the tail: Assessing lexical diversity in scholarly titles using frequency-rank distribution tail fits. *PLOS ONE*. 2018;13(7):e0197775–e0197775. Available from: https://dx.doi.org/10.1371/journal.pone.0197775.
8) Kruczek J, Kruczek P, Kuta M. Are n-gram Categories Helpful in Text Classification? . *Springer - Computational Science - ICCS 2020*. 2020. Available from: https://doi.org/10.1007/978-3-030-50417-5_39.
9) Elghanam F. Text representation and classification based on bi-gram alphabet. *Journal of King Saud University - Computer and Information Sciences*. 2021;33(2):235–242. Available from: https://doi.org/10.1016/j.jksuci.2019.01.005.