Time series analysis on Covid 19 Summarized Twitter data Using Modified TextRank

Ajit Kumar Das[1⋆], Kushagra Chitkara[2], and Apurba Sarkar[3]

[1]Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, West Bengal, 711103
[2]Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, West Bengal, 721302
[3]Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, West Bengal, 711103
writetoajit@yahoo.com, kushagrachitkara37@gmail.com, apurba@cs.iiests.ac.in

Abstract. To interpret people's sentiments using Covid 19 twitter data, we discovered that people's sentiments are distributed across multiple dimensions. Six significant themes were chosen, including administration, disease, healthcare, location, precaution, and citizens, because these topics receive a lot of attention on social media. In this paper We used the modified text rank extractive summarization approach on Covid 19 twitter data to reduce data volume without sacrificing the quality. The keywords are chosen from the pre-processed data set where the frequency of the words exceeds a pre-determined value decided through several trial runs. The goal is to extract the most number of word sets possible from the original tweets. These keywords have been grouped into the six categories listed above. If a keyword belonging to a specific bucket is identified in the summarised file of a given day, the count of that topic is increased by one. The graphs for the counts of all the themes for each day were then plotted. To identify patterns, seven-day moving average graphs are shown for each topic.

Keywords: Covid 19, Twitter data, Time series trend analysis

1 Introduction

Starting from December 2019, there are millions of tweets per day related to covid 19 twitter data. In india the Covid 19 spread has started starting from March 2020. The sentiments analysis talks about positive, negative and neutral sentiment on any subject. However, sentiment on covid 19 data can be analyzed with respect to different dimensions. We have selected six important dimensions namely administration, disease, healthcare, location, precaution and citizens for sentiment analysis as there are great deal of discussion on these topics in the social media. As the volume of the tweets per day is in millions we have utilized

⋆ Corresponding author

graph based modified text rank summarization method to reduce the tweet volume without losing the opinion of the writers. We have collected tweets from April 2020 to September 2020. Thus we get 183 summary files one each per day. These daily summary files are used as input to our method for topic based time series trend analysis. However, to decide the list of keywords for each of the topic we have taken the original tweet files as input. These original datasets contain much irrelevant and redundant information, which are removed, using various text pre-processing techniques. The Twitter dataset contains noisy information, such as URLs, dots, emoticons, symbols & pictographs, transport & map symbols, flags, and so on. The tweets are also cleaned from the hashtags ('#'), retweets ("RT"), '@' symbol used for tagging, etc. In addition to the above symbols, the emoticons and the emojis are removed to get cleaned tweets. The stopwords are removed such that only the words providing the meaningful information are kept in a sentence. After cleaning, the processed file is stored in a separate directory. These processed files are used as input to identify keywords. The target is to select optimum number of word sets from the original tweets. The words for which the frequency is more than a pre-defined value, which is set experimentally by multiple trial runs, are selected as keywords. We have also selected some hindi words written using english alphabets to include the opinion of the hindi speaking people from India. We have then classified these keywords into above six topics. To understand the trend of these topics against time we have used count of keywords as the metric for each topic. On the summarised file of a particular day, if a keyword is found which belongs to a specific bucket or topic, the count of that topic increases by one(1). We have then plotted line graphs against the date to understand the trend of these topics. However, to account for the fact that some day, there may not have enough volume of tweets, we divided the counts obtained in the previous step by the number of tweets for that particular day to normalise it. Better trends were observed in that. For better readability, we even took seven(7) day moving averages for the trends.

2 Literature Survey

Text summarization approaches are categorized into two types extractive summarization [1] and abstractive summarization [2]. Text summarization is also classified as Indicative [3] and informative [4]. Informative summarization technique is used for making Generic [5] and query-oriented [6] summary. Single-document summary [7] takes sentences from the document itself whereas multi-document summary [8] makes a summary by fusing sentences from different documents. Topic-oriented summarization [9] is made based on users topic of interest and the information extracted from the given document related to some specific topic. In this paper we have used the graph based modified text rank summarization and hence will not concentrate on the summarization literature survey. Rather we have concentrated on the twitter literature survey.
The rising popularity of social media platforms such as Facebook, Twitter, Linkedin, and Instagram, each with its unique set of features and applications,

is having a significant impact on our societies. For example, Facebook is a social network in which everyone in the network has a mutual relationship with someone else in the network. On the other hand, not everyone in the Twitter network has a reciprocating relationship with others. Twitter is a social media platform that allows users to post and receive 140-character messages known as "tweets". It was launched in 2006. There are different types of twitter data such as user profile data which is static and tweet messages which is dynamic. Tweets could be textual, images, videos, URL, or spam tweets. Most studies do not, usually, take spam tweets and here we have concentrated on text tweets. Twitter is a large forum for presenting and exchanging various ideas, thoughts, and opinions. People comment, compliment, discuss, fight, and insist regardless of where they come from, what religious beliefs they hold, whether they are rich or poor, educated or ignorant. In contrast to Facebook, where users may restrict the privacy of their profiles, Twitter allows unregistered users to read and monitor the majority of tweets. The huge amount of data offered by Twitter, such as tweet messages, user profile information, and the number of followers in the network, is extremely useful for data analysis[10]. The public timeline, which displays all of the users' tweets from across the world, is a massive real-time information stream with over one million messages per hour. As a result, tweets can be used to search through social media data and locate messages that are related to one another. Opinion mining[11], also known as sentiment analysis[12], is the process of determining the sentiment that the writer wishes to convey in his or her message. The text polarity, i.e. whether the message has a positive, negative, or neutral sentiment, is usually represented by the Sentiment. So, for any topic, Twitter may be used to grasp the sentiment of the population at large. For analysing feelings from text, a machine learning approach can be utilised. Using a Machine Learning method, some sentiment analysis is performed on Twitter posts about electronic devices such as cell phones and computers. It is possible to determine the effect of domain information on sentiment classification by undertaking sentiment analysis in a specific domain. Sentiment analysis offers numerous possibilities for developing a new application. In the industrial sphere, sentiment analysis has a significant impact; for example, government agencies and large corporations want to know what people think about their product and its market value. The goal of sentiment analysis is to extract a person's mood, behaviour, and opinion from text. Sentiment analysis is widely employed in a number of fields, including finance, economics, defence, and politics. Unstructured and structured data can be found on social networking sites. Unstructured data makes up over 80% [13] of the data on the internet. To find out what people think on social media, sentiment analysis techniques are applied. Sentiment analysis consist of four stage named as: tweet retrieval, tweet pre-processing, classification algorithm and evaluation. The general steps for twitter sentiment analysis is shown in figure 1.

Input: The subject matter or issue for which we want to do sentiment analysis is the input. We start by choosing a subject, then gather tweets pertaining to that issue, and finally perform sentiment analysis on those tweets. In this paper

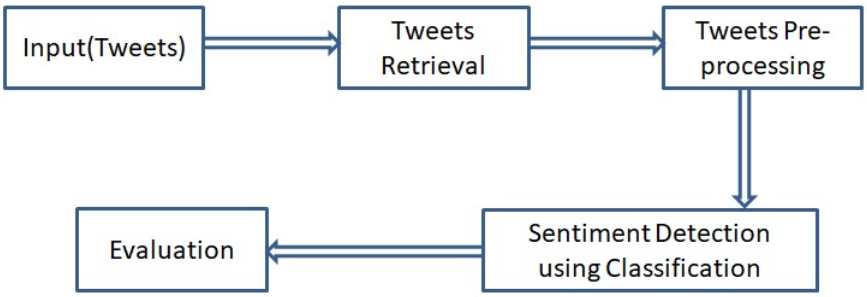

Fig. 1: Twitter sentiment analysis steps

our subject is covid 19.

Tweets Retrieval: Tweets are retrieved in this step, and they can be in any format unstructured, structured, and semi-structured. We can collect tweets using different programming languages such as Python or R (the programming language and the software environment for data analysis) or Java API.

Pre-processing: In this step data is filtered by removing irrelevant, inconsistent, and noisy data. The majority of studies focused on software, such as R, When it comes to processing Twitter data, R has several limitations and is inefficient when working with big amounts of data. A hybrid big data platform, such as Apache Hadoop (an open source Java framework for processing and querying enormous volumes of data on large clusters of commodity hardware), is typically used to solve this challenge. Hadoop can also handle structured and semi-structured data, such as XML/JSON files. Hadoop's strength is in storing and processing vast amounts of data, whereas R's strength is in analysing data that has already been processed.

Sentiment Detection: There are different sentiment classification algorithm to find the polarity of a given subject. In supervised learning Naive Bays, SVM and maximum entropy are widely used algorithms. Whereas in unsupervised learning lexicon based, corpus based and dictionary based are used to perform the sentiment analysis. In this paper we have used topic modelling by selecting keywords for each topic as explained in next section.

Evaluation: The output is assessed to see whether we should choose it or not, and the results are then shown as a bar graph, pie chart, or line graph. In this paper we have done time series analysis for each topic and used count of keywords for each topic as the metric for plotting line graph against the date to understand the trend of that topic

3 PROPOSED METHODOLOGY

In this paper we have applied the modified text rank extractive summarization menthod on covid 19 twitter data. Figure 2 represents the process flow diagram of the proposed methodology.

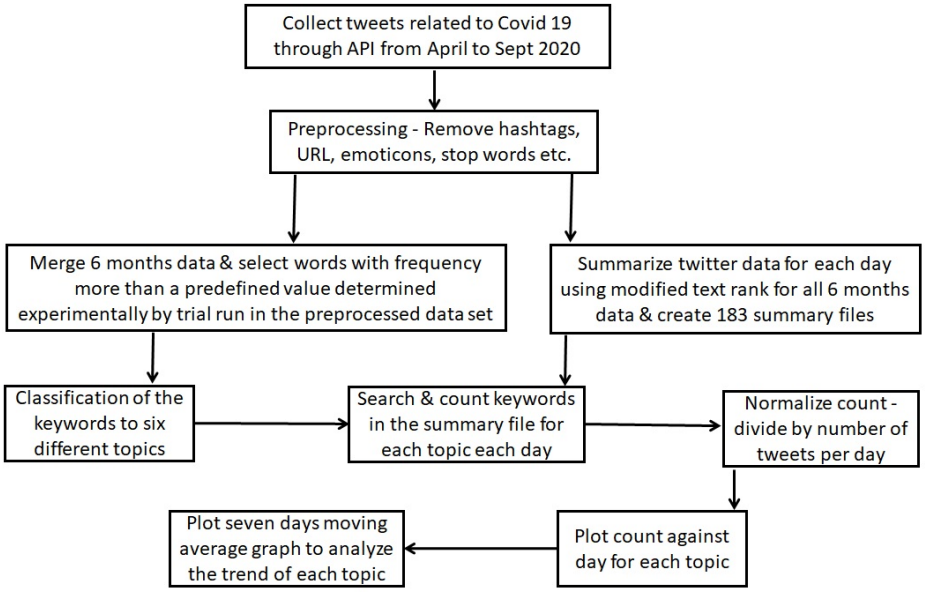

Fig. 2: Process Flow Diagram

3.1 Data collection and preprocessing

The COVID-19 coronavirus articles collected during the period of April 2020 to September 2020 are used to demonstrate the application of the modified text rank summarization method.
The dataset used is the one which is maintained by Panacea Lab [1]. They have maintained data for all covid related tweets for the required time period. We have followed the steps to download the data files as mentioned in the tutorial of the repository. After downloading, and filtering for English language, what we get are the tweet ids. These ids are then concatenated in a text file. This text file is fed to a hydrator tool, DocNow hydrator[2]. After hydrating, what we obtain is

[1] `https://github.com/thepanacealab/covid19_twitter`
[2] `https://github.com/DocNow/hydrator`

a JSONL file with an option to convert that to CSV. An important thing to note is that, the information is lost while converting to CSV. So, we have manually converted the JSONL file to CSV file for extracting necessary information. After getting the relevant information, the tweets are filtered on the basis of if they're geotagged or not. This process is repeated for six(6) months data. In trend analysis we have used global data which covers tweets from all countries.

The datasets contain many irrelevant and redundant information, which are removed, using various text preprocessing techniques [14]. The Twitter dataset contains noisy information, such as hashtags, URLs, dots, emoticons, symbols & pictographs, transport & map symbols, flags, and so on. We have removed the URLs present in the tweets by using the output of urllib.parse on the tweet. The tweets are also cleaned from the hashtags ('#'), retweets ("RT"), '@' symbol used for tagging, etc. In addition to the above symbols, the emoticons and the emojis are removed to get cleaned tweets. It is achieved by performing string manipulations and using the UNICODE ranges for emojis. After cleaning, the processed article is tokenized into sentences using the NLTK library of Python. The sentences are converted to lowercase, and punctuation are removed from the sentences. The stopwords are removed using nltk.corpus such that only the words providing the meaningful information are kept in a sentence. Next we tokenize the given text into sentences and finally, tokenize each sentence into a collection of words which is used as the input for the modified text rank algorithm.

3.2 TextRank based summarization

In paper graph based text summarization using modified text rank[15], we have considered the sentences in the document which are equivalent to web pages in the PageRank system [16]. The probability of going from sentence A to sentence B is equal to the similarity between two sentences. This modified TextRank uses the intuition behind the PageRank algorithm to rank sentences based on which we can select the most important sentences from the input text document. In PageRank, important web pages are linked with other important web pages. Similarly, in modified text rank algorithm the important sentences are linked (similar) to other important sentences of the input document. Here, isf-modified-cosine similarity takes care of the different level of importance for the corresponding words in the sentences and also consider different length of the sentence in the document. Finally for summarisation top n scored sentences are rearranged as per the input sentence index. This construct the summary of the document.

3.3 Topic wise keyword selection and count

Once the cleaned file was obtained from the pre-processing steps, it was then summarised using modified textrank algorithm. The summary for each day was prepared separately. Thus we got 183 summary files. In the next step, we have selected six topics to understand the trends of these individual topics over time. The topics include, administration, disease, healthcare, location, precaution and

citizens. We chose these topics because they have received a lot of media and public attention. The words that appear more than a predefined value, which is set experimentally by multiple trial runs, in the original tweet set have been selected as a keyword. In our experiment, the predefined value is set to ten(10). We have manually grouped these keywords into the six different topics such that the words that resemble the topic most are considered as part of that topic. The topic wise keywords selected are shown in figure 3

Topic	administration	disease	healthcare	location	precaution	citizens
Keywords	['relief', 'lie', 'pulis', 'srkaar', 'sarkar', 'sarkaar', 'congress', 'police', 'pm', 'chief', 'minister', 'hm', 'members', 'member', 'distributed', 'govt', 'government', 'suppo', 'food', 'judgement', 'distributing', 'diyaa', 'modi', 'scandalous', 'cm', 'kovind', 'mjduur', 'food,', 'presidents', 'nehru', 'dynasty', 'attacks', 'opponents', 'leadership', 'narendra', 'prime', 'bjp','aadesh','amendment','ordinance,','commitment','raashn', 'ration', 'express','scam','niti','president','modis','fund','ruupaay', 'rupee', 'producer','constable','maansiktaa','crore']	['manifests','covid 19', 'covid-19', 'corona', 'coronavirus', 'cases', 'koronaa', 'positive', 'patients', 'covid', 'pandemic', 'crisis', 'koroonnnaa', 'spread', 'dies', 'virus', 'cured', 'donate', 'diseases', 'maut', 'suffering', 'deaths', 'died','epidemic','death','patient','demise','case','dead','coronavirus','maaro', 'tested', 'fighting','symptoms']	['help', 'nivaarnn', 'humanitarian','aid','treatment', 'ilaaj', 'ilaaz', 'metabolics', 'stitched', 'labs', 'blood', 'fight', 'rapid', 'stepped', 'respect', 'healthcare', 'plasma', 'icmr', 'hospitals', 'donated', 'wellbeing','commend','lose','save','lives','hospital','dr','doctor', 'dr.', 'antivenom','treatment','poisonous','medicines', 'donating','medical','testing','giving','gratitude','village doctors']	['india','delhi', 'desh', 'world', 'dillii','dilli', 'mumbai', 'indias', 'tamil', 'country', 'china', 'states', 'state', 'tmilnaaddu', 'tamilnadu', 'manipurdignity','area','maharashtra','chennai','nadu','yuupii', 'up', 'ghaziabad', 'gaajiyaabaad','american', 'amerikn','countries','nizamuddin','bihar,','central','western','karnataka','south']	['proactive','masks', 'test', 'kits', 'face', 'kit', 'kitt', 'walking', 'health', 'measures', 'safety,','mask',',purchase', 'ghr','ghar', 'lockdown', 'home', 'month', 'stay', 'lonkddaaun', 'lockdown']	['office','people', 'workers', 'indian', 'log', 'lady', 'bhuukhe', 'bhukhe', 'everyone', 'youth', 'girl', 'human', 'krodd','crore', 'media', 'migrant', 'thalapathy', '12year', 'kmaane','kamane', 'shri', 'shrii', 'actor', 'worker', 'khaanaa', 'khana', 'lakhs', 'millions', 'appeals','news','everyones', 'everyone','brothers','brother', 'person','protect','journalists','journalist', 'sisters','sister', 'needy','body','private',',sir','community','neighbors','shops','privaar',',parivar', 'parivaar', 'relatives','family']

Fig. 3: Topic and Keywords

To get the trend of these topics, we have searched for those keywords in the summarised file. If a keyword belonging to the Administration bucket was found, then the count of Administration went up by 1 for that particular day. We then plotted the graphs for the counts of all the buckets for each day and observed the trend for the same.

However to account for the fact that some days there may not have enough volume of tweets, we divided the counts by the number of tweets for that particular day to normalise it. Better trends were observed in that. For better readability, we even took seven(7) day moving averages for the trends.

4 Experimental Result

The time series graphs obtained for these six topics are explained in Figure 4 and 5 .

Fig. 4: Topic: Administration, Disease & Healthcare

Topic Administration : For this graph we could clearly observe a small but constant rise in count Topic: Administration until about 1st week of May, after which it meets a steep decline. This could be explained by the fact that most of Europe had seen the worst of the first wave and the cases were in a decline there. The United States was giving a constant number of cases and India too hadn't experienced the effect of the first wave and was still in lockdown. From about 1st of June to mid August, we see a fairly constant line with a spike here and there. This could be because of how India was easing it's lockdown restrictions so a few days of news may have inspired the spike. Further Europe and China were beginning to open their borders and hence administration tweets could be associated with them. After mid August, we suddenly see a spike in the graph. This is because of the fact that the US was giving an all time high cases at that point of time. Further what could have inspired this huge spike is the fact that the US was about to start with their presidential debates and a lot of administration related tweets would have been regarding these debates. The first of these debates was held on 30th of September and the huge spike is a clear indication of the fact that the election inspired a lot of such tweets.

Topic Disease : The disease graph is fairly constant throughout as was expected because after normalising the data, the relevance of covid in itself has not seen any decline. A few talking points are the spike observed around July. Apart from the cases in India being on a constant rise, this could also be associated with the fact that this was when Oxford University had first announced they were actively working on a vaccine. Also Brazil was right in the peak of its first wave. The disease graph then was more or less a straight line.

Topic Healthcare : As we could see in the early part of the graph, the counts are really high. This unusually high number of counts is associated with the fact that early April was the time when Boris Johnson was hospitalised because of Covid. Further this when we were seeing cases of violence against healthcare workers in India that may have inspired these tweets to be in such large numbers. As Covid started getting normalised and restrictions started lifting, the healthcare related tweets started dropping. The next spike we observe is around mid June. The constant rise could be explained by the fact that world leaders like Brazil's president Bolsonaro and soon after Amit Shah tested positive for Covid.

Topic Location : We observe the count to be high earlier on and then facing a dip. This could be because of the fact that Europe and the US were past the peak of the first wave and weren't tweeting regarding the location of Covid. The next spike that we see around early June is when Europe finally starts announcing it's opening up of borders so a lot of tweets might have been regarding travelling in covid. Also this is around the time that Chinese-Indian relations soured with people going as far as calling covid a Chinese lab made virus. A lot of tweets could've been inspired by this. Meanwhile the then president Trump insisted on calling it the 'Chinese Virus' inspiring a lot of hate crimes against Asian-Americans which is reflected in the spike starting June. The next spike we

Fig. 5: Topic: Location, Precaution & Citizens

see is around September when Trump started actively campaigning and many of his supporters went about a similar Asian hate as seen before.

Topic Precaution : Popularity of precaution related tweets is seen to be constantly declining. This is because we started getting accustomed to the new

normal. However a steep spike is seen about early June. This is when Europe had seen through the first wave so as people were stepping out more, they were constantly warned to keep precautions in check. Even sports like Formula 1 and Football were starting again so precaution measure tweets were gaining popularity. But the most important factor here would be the rise of the 'Black Lives Matter' protests occurring all around the world after the shooting of George Floyd. As more and more people wanted to participate in the protests and marches, they were reminded to keep precautions in check and hence a lot of precaution related tweets would be because of this. Lastly, after this decline, we see another spike in mid September. This is when India was giving the highest single day cases seen and the Prime Minister was urging everyone to follow precaution measures. Even in the US, presidential debates were starting to begin and the rallying leading up to these events would have inspired precaution tweets.

Topic Citizens : By just taking a quick look, we can conclude that this graph looks very similar to the precaution and location graphs. And much for the same reasons. The June spike is associated with people travelling and protests occurring in different parts of the globe. A slow and steady decline follows and ultimately a huge spike around September because of the extremely high number of cases in India and the simultaneously occurring elections in the US.

5 Conclusion and Future Work

In this paper we have studied different literature for twitter data sentiment analysis. We have used modified text rank algorithm to summarize covid 19 twitter data of each day starting from April 2020 to September 2020. Using keyword based topic modelling for six predefined topics which are of great importance with respect to covid 19, we have done a trend analysis of the topic. The behaviour of the global twitter user could be explained scientifically from the trends of the graph. This also explains the effectiveness of the summary to reduce data set as well as the bucketing of the keywords into defined topics. As a future work we plan to do time series cluster analysis on the global covid 19 twitter data to cluster the countries. This will provide us the similar trends for countries in the global scenario.

References

1. Wong, K.F., Wu, M., Li, W.: Extractive summarization using supervised and semi-supervised learning. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics (2008) 985–992
2. Khan, A., Salim, N.: A review on abstractive summarization methods. Journal of Theoretical and Applied Information Technology **59**(1) (2014) 64–72

3. Kan, M.Y., McKeown, K.R., Klavans, J.L.: Applying natural language generation to indicative summarization. In: Proceedings of the 8th European workshop on Natural Language Generation-Volume 8, Association for Computational Linguistics (2001) 1–9
4. Saggion, H., Lapalme, G.: Generating indicative-informative summaries with sumum. Computational linguistics **28**(4) (2002) 497–526
5. Gong, Y., Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2001) 19–25
6. Tang, J., Yao, L., Chen, D.: Multi-topic based query-oriented summarization. In: Proceedings of the 2009 SIAM International Conference on Data Mining, SIAM (2009) 1148–1159
7. Litvak, M., Last, M.: Graph-based keyword extraction for single-document summarization. In: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, Association for Computational Linguistics (2008) 17–24
8. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document summarization by sentence extraction. In: Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4, Association for Computational Linguistics (2000) 40–48
9. Harabagiu, S., Lacatusu, F.: Topic themes for multi-document summarization. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2005) 202–209
10. Anber, H., Salah, A., Abd El-Aziz, A.: A literature review on twitter data analysis. International Journal of Computer and Electrical Engineering **8**(3) (2016) 241
11. Adarsh, M., Ravikumar, P.: Survey: Twitter data analysis using opinion mining. International Journal of Computer Applications **128**(5) (2015)
12. Patel, A.P., Patel, A.V., Butani, S.G., Sawant, P.B.: Literature survey on sentiment analysis of twitter data using machine learning approaches. IJIRST-International journal for Innovative Reasearch in Scinece & Technology **3**(10) (2017)
13. Mittal, A., Patidar, S.: Sentiment analysis on twitter data: A survey. In: Proceedings of the 2019 7th International Conference on Computer and Communications Management. (2019) 91–95
14. Vijayarani, S., Ilamathi, M.J., Nithya, M., et al.: Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks **5**(1) (2015) 7–16
15. Mallick, C., Das, A.K., Dutta, M., Das, A.K., Sarkar, A.: Graph-based text summarization using modified textrank. In: Soft computing in data analytics. Springer (2019) 137–146
16. Mihalcea, R.: Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics (2004) 20