JCSDA, Vol. 02, No. 02, 1–10
DOI: 10.69660/jcsda.02022501
ISSN 2959-6912

# Artificial Intelligence and Social Media Data as an Alternative Source of Insights During Pandemics: A Case Study of Twitter for COVID-19 in Tanzania

Deogratias Mzurikwao*[,1,2], Asa Kalonga [2,3], Simeon Mayala[1,2] ,Peter Nyanda[4]

[1] *Muhimbili University of Health and Allied Sciences (MUHAS), Department of Biomedical Engineering, Tanzania*

[2] *Emerging Technologies for Health Research & Development Laboratory (ETH-MUHAS), Tanzania*

[3] *XsenseAI company limited Tanzania*

[4] *United Nations Development Programme (UNDP-Tanzania)*
** Corresponding author: dmzurikwao@gmail.com*

Recently, social media has become one of the major sources of information for different sectors, including healthcare. During COVID-19, there was a lot of community engagement on social media than traditional physical engagements. Tanzania, in particular, had a different approach to tackling the COVID-19 pandemic as the country didn't really practice lockdown, and very little information was shared by the authorities in traditional media houses. To understand what really happened during the pandemic, we tried to investigate social media as people were sharing information rather than the traditional media houses, and find the correlation with other sources. In this study, we extracted and analysed four-day periods of Twitter posts of Kinondoni district, Dar es Salaam, Tanzania. The four days were picked intentionally as it was the time when Tanzania started approaching the pandemic, as the rest of the world changed its course. Most of our analysis results significantly correlate with the results reported by the government during the same period, $21^{st}$-$24^{th}$ April 2020. We further performed an analysis of how much the COVID issue was discussed online. As the WHO and many governments around the world have been providing education to people on how to protect themselves and slow down the spread, none have had a way to measure how well people were educated. We analysed 20,421 tweets of Kinondoni district, the most populous district in Dar es Salaam, and where many expats live, and found out how well people were educated about the CORONAVIRUS disease. We further created an Artificial Intelligence algorithm, a Deep learning to be specific, which has been able to classify tweets into COVID and Non-COVID classes with an accuracy of 93%. Our results mean that social media data analysis can be used as a tool for topic modelling to detect the most trending topics like disasters, election events, and epidemics.

*Keywords*: Social media data, COVID-19, pandemic, Artificial Intelligence, Web scraping.

## 1. Introduction

Social media has become one of the major sources of information [1]. It has been a major source for those seeking information for health-related matters, as study shows that around 59% of adult Americans have accessed this type of information

from social media [2], mostly Facebook and Twitter. According to reports, Chinese social-media platforms WeChat [2] and Weibo [3] saw spikes in the terms 'SARS,' 'coronavirus,' and 'shortness of breath,' weeks before the first cases were confirmed. In this study, we have pulled and analysed four (4) days' period Twitter posts of Kinondoni district, Dar es Salaam, Tanzania. Most of our analysis results significantly correlate with the results reported by the government during the same period, $21^{st}$-$24^{th}$ April 2020. We further performed an analysis of how much the COVID issue was discussed online. As the WHO and many governments around the world have been providing education to people on how to protect themselves and slow down the spread, none have had a way to measure how well people were educated. We analysed 20,421 tweets of Kinondoni district, the most populous district in Dar es Salaam, and where many expats live, and found out how well people were educated about the CORONAVIRUS disease. As there is an average of 500 million tweets per day [4], it is impossible to keep up with the analysis of all the tweets, especially in an epidemic period where real-time information is of high value. We created an Artificial Intelligence algorithm, a Deep learning to be specific, which has been able to classify tweets into COVID and Non-COVID classes with an accuracy of 93%. Our results mean that social media data analysis can be used as a tool for topic modelling to detect the most trending topics like disasters, election events, and epidemics.

Social media has become the main source of news online, with more than 2.4 billion internet users, nearly 64.5 per cent receive breaking news from social media sites like Facebook, Twitter, YouTube, Snapchat, and Instagram instead of traditional media. Many people go to traditional media to just verify the news hours after it has been circulated over social media platforms. According to Forbes online magazine, in a recent survey, 50 per cent of Internet users surveyed said that they hear about the latest news via social media before ever hearing about it on a news station. It is estimated that 90.4% of Millennials, 77.5% of Generation X, and 48.2% of Baby Boomers are active social media users. Users spend an average of 3 hours per day on social networks, posting, commenting on different posts on social media.

Recently, the world has been facing the coronavirus outbreak. Although different measures have been taken globally to slow down the spread of the virus, including education on how people can protect themselves, there is no method to measure how effective these measures have been. Different governments, especially in developing countries, cannot measure the impact of different efforts to combat the disease or its impact socially and economically. According to a social media analytics company, Sprinklr, it counted a record nearly 20 million mentions of coronavirus-related terms on March 11, with the US alone. This shows how much social media contains potential data about the outbreak.

This study aimed at extracting and analysing social media data related to coronavirus from social media in Tanzania. By analysing the social media data related

to coronavirus, we expected to derive information like, How much the community is aware of the outbreak, How much people understand about the outbreak, How well people protect themselves, What are the challenges people face in protecting themselves, What are the economic effects of the outbreak, Which regions are more and which regions are less educated about the outbreak and How much the government efforts in combating the situation have been effective.

## 2. Method

This was a pilot project with the aim of seeing and proving a concept of using social media to gain analytical insights into different issues. With a focus on COVID-19, we have pulled five days of English Twitter data from the Kinondoni district. Kinondoni is the most populated district among other districts in Dar es Salaam, with half of the city's population residing within it. It is also home to high-income suburbs and most English-speaking people, hence becomes a perfect target for pulling English tweets.

### 2.1. *Data pre-processing and AI model training*

We have five days tweeter activities in Kinondoni district from $21^{st}$ to $24^{th}$ April, which resulted in a total of 20422 tweets. These dates were specifically chosen as it is one of the peak days of COVID-19 cases in Tanzania, as it was announced by the Ministry of Health of the United Republic of Tanzania. Our dataset contained tweet posts, replies, retweets, and details like user_id, user_name, source of the tweet, profile locations, and coordinates were removed for data protection and privacy issues. Swahili and mix of Swahili and English tweets were removed, remaining with English only English-only tweets. This decision was driven by the availability of annotation resources and the absence of COVID-19-specific Swahili NLP tools at the time. The resulting English corpus formed the basis for all subsequent labelling and model training.

10,000 tweets were manually labelled in categories like COVID (for tweets related to COVID-19) and non-COVID (for tweets related to other stuff but not COVID). The COVID category was further labelled into EDUCATED/Not EDUCATED by manually analysing the tweet if the person who tweeted seems to be educated or not educated with respect to COVID-19. The Not EDUCATED class contained tweets with misinformation, as we assumed that those posting misinformation were not well educated about the pandemic. By using these 10,000 tweets, we trained a deep learning model to automatically classify the remaining 10422 data into the four categories, namely COVID, NON-COVID, EDUCATED, and NOT EDUCATED. The NOT EDUCATED category also contained tweets with misinformation on COVID, as we assumed those who posted misinformation were not well educated about the pandemic, although we should acknowledge that some information was driven by different motives and not exactly about education.

Given the scale of the dataset and the practical constraints on annotation time, a zero-shot classification approach was used to generate an initial seed set of labels before human verification. Two domain-specific keyword dictionaries were constructed: one for COVID-class tweets (e.g., "coronavirus," "COVID," "quarantine," "mask," "pandemic," "lockdown," "PCR," "WHO") and one for NOT EDUCATED-class signals (e.g., "fake," "hoax," "cure," "5G," "conspiracy," "bleach," "microchip"). Tweets were matched against these dictionaries to auto-assign labels across the four categories COVID, NON-COVID, then the COVID category data was labelled to EDUCATED, and NOT EDUCATED, producing an automatically labelled corpus. This labelling process was then reviewed and corrected by two independent annotators working from a shared codebook, focusing verification effort on ambiguous and borderline cases flagged by the keyword rules. This combined approach allowed reliable labelling of 10,000 tweets at a scale that would have been impractical through manual annotation alone. The NOT EDUCATED category included tweets with misinformation, on the premise that those posting inaccurate information were unlikely to be well informed about the pandemic, although we acknowledge that some misinformation may be driven by deliberate intent rather than lack of education.

The deep learning classifier was built using a hybrid CNN-LSTM architecture, chosen for its strengths in short-text classification: convolutional layers extract local n-gram features (e.g., "social distancing," "wash hands") while the LSTM layers capture sequential dependencies across the full tweet. Tweet text was tokenised and represented using pre-trained GloVe word embeddings (100-dimensional, trained on Twitter data), which were fine-tuned during training to adapt to the COVID-19 domain. The network consisted of an embedding layer, a 1D convolutional layer (128 filters, kernel size 3, ReLU activation), a max-pooling layer, a bidirectional LSTM layer (64 units), and a fully connected softmax output layer with four nodes corresponding to the four target classes. Dropout (rate = 0.5) was applied after both the convolutional and LSTM layers to mitigate overfitting.

The 10,000 labelled tweets were split into training (80%), validation (10%), and test (10%) sets. The model was trained for 20 epochs using the Adam optimiser (learning rate = 0.001) with categorical cross-entropy loss. Early stopping was applied based on validation loss to prevent overfitting. The trained model was then used to classify the remaining 10,422 unlabelled tweets into the four target categories. Model performance on the held-out test set yielded an overall accuracy of 93%. Per-class precision, recall, and F1-score were computed to account for class imbalance between the COVID and NON-COVID categories. The NOT EDUCATED class, being the smallest, was subject to additional scrutiny and results should be interpreted with this class imbalance in mind.

## 3. Results

From the pulled tweets, we analysed to find if COVID-related words were among the most popular topics on Twitter in five days. It is known that news about the coronavirus started trending on Chinese social media before it was officially reported. This analysis aimed at testing whether social media data can be used to detect an event like epidemics for early intervention.

In Fig. (1), as obtained from the Worldometer [5] From the $21^{st}$ of April 2020, the COVID-19 cases were on the rising phase in Tanzania, with a little slowing down approaching 24th April 2020. Prior to generating the figures below, we performed a stop words process to remove some common English words like "the", "is", "are", "it", "this", and the like, and retained only topic words. This is a common practice in natural language processing, as this word always occurs frequently in English sentences. This trend correlates much with our analysis of tweets extracted on the same duration as Most of the CORONAVIRUS related words were more common on the $21^{st}$, $22^{nd}$, and $23^{rd}$, and significantly slowed down on the $24^{th}$, as can be seen in figures 2A, 2B, 2C, 2D, and 2E below. The overall analysis of the top 20 most commonly discussed topics on Twitter in Figure 2F shows that COVID-19-related words were the most dominant words on Twitter in Kinondoni district. On April $17^{th}$, the president of the United Republic of Tanzania declared three days of national prayer to help defeat the coronavirus. Both of our pulled tweets from $21^{st}$ to $24^{th}$ were dominated by the word "GOD", as can be seen in Fig. (2), A-E. Also, Figure 2E, which plots the overall discussed topics on Twitter, the name "GOD" still dominated the tweets posted and retweeted.

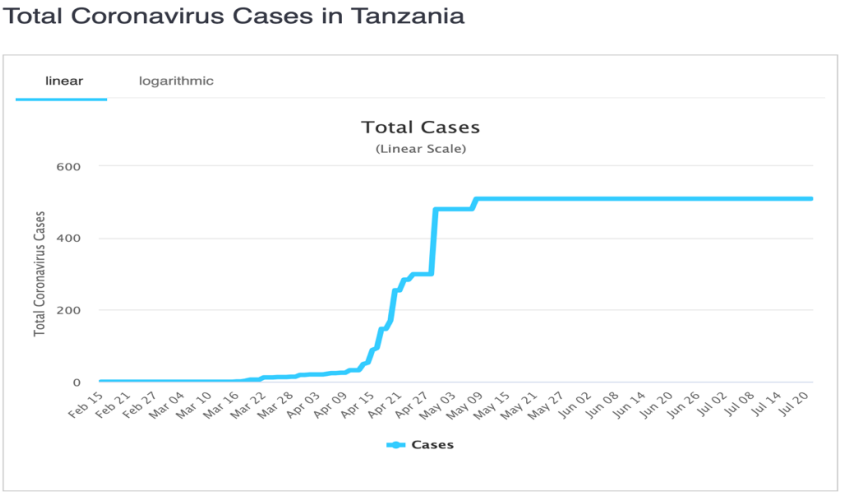


Fig. 1. COVID-19 cases tracking as per Worldometer tracking

It should be known that, "umwalimu", in the official Twitter name of Honourable Ummy Mwalimu, the Minister of Health, Community Development, Gender, Seniors, and Children in the Government of the United Republic of Tanzania, and was one of the Key figures in the COVID response in Tanzania.

Knowing how much people discuss a certain topic can be a measure of how much people care about or are impacted by the particular issue [5]. Fig. (3) below shows the trend of the percentage of COVID-related tweets against other tweets in a period of four days, April $21^{st}$ to $24^{th}$, 2020. It can be seen that, although people continued to tweet about other stuff, COVID-related tweets contributed an average of approximately 18% to the total daily tweet volume across the four-day period. The number of retweets and likes COVID-related tweets received further tells the extent how much COVID occupied online conversations.

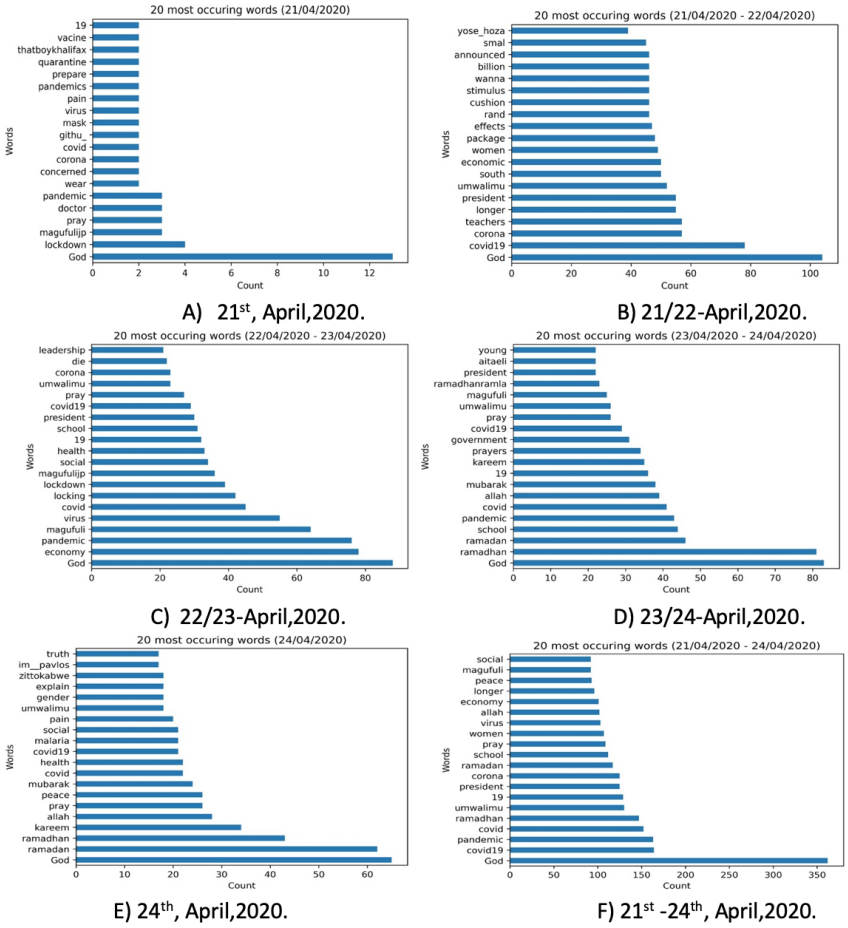


Fig. 2. The top 20 most common spoken words on Twitter, $21^{st}$-$24^{th}$ April 2020.

Many news agencies have been employing people to manually track trends of different topics on social media. As we have pulled more than 20,000 tweets in just 4 days, within one district, this shows humans can't track the trends. Globally, there are more than 500 million tweets in a single day [6]. This makes it impossible to keep track of the activities happening online in real time. To automate this while keeping up to date with the online data, we have built and trained a deep learning algorithm that can automatically classify the tweets into four different categories with an accuracy of 93%. This model was trained with 10,000 tweets out of the more than 20,000 tweets we collected.

In our analysis of tweets originating within Kinondoni district, we found that more than 18% were discussing COVID, as shown in Figure 3A. The trending of COVID-related tweets picked a sharp rise between $22^{nd}$ to $23^{rd}$, April of 2020, as shown in Figure 3B. This correlates with the graph of Worldometer reporting the statistics of COVID-19 cases in Tanzania, as shown in Fig. (1) above.

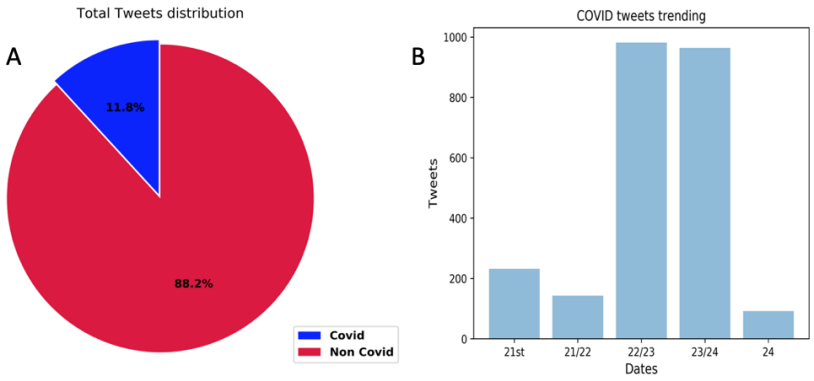


Fig. 3. A: Four days of tweets distribution, B: Four days of trending tweets related to COVID-19, in Kinondoni district.

As mentioned earlier, each day, there are more than 500 million tweets globally [7], it worth noticing how many likes and retweets a tweet has received to know how viral the information in that tweet has spread. According to research published by Microsoft researchers [8], retweet is one of the most important ways to achieve value on Twitter and is often poorly used. A total of 10185 were counted on different tweets posted between $21^{st}$ to $24^{th}$, April 2020 within the Kinondoni district alone. Out of all likes, 2750 were on COVID-related tweets. In retweets, a total of 26453 tweeters were retweeted, 8143 of them were COVID-related. Figure 4A and 4B show the distribution of the likes and retweets, respectively. More than a quarter of likes and retweets were on COVID-related tweets.

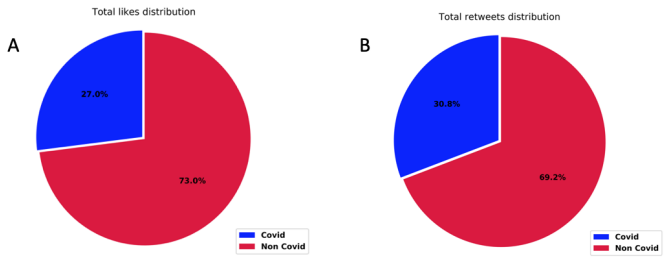


Fig. 4. Likes and retweets distributions.

Unlike other news media, most of the social media posts are not regulated. This has been leading lot of misinformation. Most of the social media accounts generating misinformation do not share accurate details about themselves [9], this makes it harder to really track them down and take action. In addition, although different governments and other organizations were actively providing education on how people can protect themselves about CORONAVIRUS in order to slow down the spread, there is no mechanism to know how well or which parts of the country were not educated for early interventions. In this regard, we checked how well people are educated about the coronavirus by performing sentiment analysis on the posted tweets. We found a relatively higher percentage (12.4%) of all tweets related to COVID seemed not educated about the pandemic out of all 2413 COVID-related tweets we pulled, shown in Fig. (5).

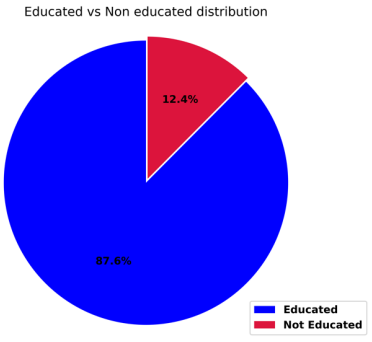


Fig. 5. COVID related educational analysis.

## 4. Discussion

As reported at the beginning of 2020, Active social media users have passed the 3.8 billion mark, with this number increasing by more than 9 percent (321 million new users) every year [10]. These users produce over 2.5 quintillion bytes of data

every day [11], leaving digital foot print. Analysis of these data can yield significant results to understand, analyze, and predict different aspects. Our analysis of Twitter data from Kinondoni district, Dar es Salaam, correlates strongly with the actual situation reported during the same period, demonstrating proof of concept that social media data can serve as a real-time surveillance tool during public health emergencies. The CNN-LSTM classifier's ability to perform automated four-class sentiment analysis at 93% accuracy establishes that this pipeline can scale beyond manual review, which is critical given the volume of tweets generated globally each day[10].

The finding that 12.4% of COVID-related tweets in Kinondoni district reflected poor awareness of the pandemic carries important practical implications. Studies of COVID-19 misinformation on Twitter have reported that up to 25% of highly shared tweets in some regions contained misleading health information [12], suggesting our observed proportion, while lower, is not negligible. This sentiment classification approach could be operationalised to identify geographic hotspots of misinformation and trigger targeted health messaging campaigns. Rather than relying on broad national communication strategies, health authorities could use real-time social media monitoring to direct accurate information to specific communities where the "not educated" signal is elevated. The dominance of the word "GOD" across all four days linked to the presidential declaration of three days of national prayer on $17^{th}$ April 2020 further illustrates how social media captures socio-cultural dimensions of pandemic response that are invisible to conventional epidemiological surveillance.

Future work should extend this framework to Swahili-language tweets, which would substantially broaden the reach and equity of the surveillance approach. Expanding geographic scope beyond Kinondoni to include rural districts is essential for equity-focused applications. Longitudinal studies spanning the full scope of the pandemic would further enable analysis of how public sentiment and misinformation evolved over time in relation to government policy shifts, providing evidence to inform communication strategies in future public health emergencies across sub-Saharan Africa

## 5. Conclusion

This study demonstrates that social media data, combined with AI-driven natural language processing, can serve as a viable complementary source of public health intelligence during pandemics. In the context of Tanzania's COVID-19 response characterised by limited official data transparency and no formal lockdown Twitter data from Kinondoni district yielded insights consistent with independently reported case trends. The CNN-LSTM classifier achieved 93% accuracy in classifying over 10,000 tweets into COVID-related and awareness categories. Furthermore, the ability to distinguish educated from non-educated responses form covid related tweets offers an innovative mechanism for monitoring public health literacy in near

real-time, which could directly inform communication strategies in future outbreaks. These findings have direct implications for pandemic preparedness planning across sub-Saharan Africa, where social media adoption is growing rapidly and traditional surveillance infrastructure often lags public health need.

**Acknowledgement**: Funding acknowledgement will be presented here.

## References

1. D. Westerman, P. R. Spence, and B. Van Der Heide. Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, 19(2):171–183, 2014. doi: 10.1111/jcc4.12041.
2. Y. Lu and L. Zhang. Social media wechat infers the development trend of covid-19. *Journal of Infection*, Jul 2020. doi: 10.1016/j.jinf.2020.03.050.
3. J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey. Data mining and content analysis of the chinese social media platform weibo during the early covid-19 outbreak: Retrospective observational infoveillance study. *JMIR Public Health and Surveillance*, 6(2):e18700, Apr 2020. doi: 10.2196/18700.
4. D. Sayce. The number of tweets per day in 2020, 2020. URL https://www.davidsayce.co.uk/. Accessed: Apr. 2020.
5. A. Banerjee and S. Chaudhury. Statistics without tears: Populations and samples. *Industrial Psychiatry Journal*, 19(1):60–65, 2010. doi: 10.4103/0972-6748.77642.
6. D. Antonakaki, P. Fragopoulou, and S. Ioannidis. A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 164:114006, 2021. doi: 10.1016/j.eswa.2020.114006.
7. M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala. The covid-19 social media infodemic. *Scientific Reports*, 10:16598, 2020. doi: 10.1038/s41598-020-73510-5.
8. S. Golder. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. Working Paper/Manuscript, 2010.
9. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective, 2017. URL http://arxiv.org/abs/1708.01967.
10. S. Kemp. Digital 2020: Global digital overview, Jan 2020. URL https://datareportal.com/reports/digital-2020-global-digital-overview. Accessed: Apr. 2020.
11. R. H. Bajaj and P. L. Ramteke. Big data-the new era of data. *International Journal of Computer Science and Information Technologies (IJCSIT)*. URL http://www.ijcsit.com.
12. G. Eysenbach. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research*, 11 (1):e11, 2009. doi: 10.2196/jmir.1157.