

Understanding COVID-19 News Coverage using Medical NLP

Ali Emre Varol¹, Veysel Kocaman¹, Hasham Ul Haq¹ and David Talby¹

¹John Snow Labs inc. 16192 Coastal Highway, Lewes, DE 19958, USA

Abstract

Being a global pandemic, the COVID-19 outbreak received global media attention. In this study, we analyze news publications from CNN and The Guardian - two of the world's most influential media organizations. The dataset includes more than 36,000 articles, analyzed using the clinical and biomedical Natural Language Processing (NLP) models from the Spark NLP for Healthcare library, which enables a deeper analysis of medical concepts than previously achieved. The analysis covers key entities and phrases, observed biases, and change over time in news coverage by correlating mined medical symptoms, procedures, drugs, and guidance with commonly mentioned demographic and occupational groups. Another analysis is of extracted Adverse Drug Events about drug and vaccine manufacturers, which when reported by major news outlets has an impact on vaccine hesitancy.

Keywords

NLP, COVID-19, Spark NLP, Natural Language Processing, Vaccine, Adverse Drug Effects.

1. Introduction

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) - commonly known as COVID-19 - was first reported in December 2019. Due to its highly contagious nature, it quickly spread through the world, prompting the World Health Organization to declare the virus outbreak as a global pandemic on March 11, 2020 [1]. As news media reporting is understood to play a central role during national security and health emergencies [2], during the pandemic, media representations of complex, rapidly evolving epidemiological science shape public understandings of the risks, measures to limit disease spread, and associated political and policy discourses [3].

In this study we analyse news coverage from two prominent media outlets: CNN and The Guardian. The dataset includes more than 36,000 articles and is analyzed using the clinical and biomedical NLP models from the Spark NLP for Healthcare library [4]. Spark NLP is an open-source and widely deployed software library, built on top of Apache Spark, that provides production-grade implementations of recent deep learning and transfer learning NLP algorithms and models. It enables combining tasks into unified NLP pipelines in Python, R, Java, or Scala

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): *Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), 10-April-2022*

✉ emre@johnsnowlabs.com (A. E. Varol); veysel@johnsnowlabs.com (V. Kocaman); hasham@johnsnowlabs.com (H. U. Haq); david@johnsnowlabs.com (D. Talby)

🆔 0000-0003-3228-1129 (A. E. Varol); 0000-0002-0065-6478 (V. Kocaman); 0000-0002-8417-3288 (H. U. Haq); 0000-0003-2782-5478 (D. Talby)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

and is the only library that can scale up training and inference on any Spark cluster. Spark NLP for Healthcare provides healthcare-specific algorithms and over 400 pre-trained models that have obtained state-of-the-art accuracy on public academic benchmarks in biomedical named entity recognition [5], clinical assertion status detection [6], medical relation extraction [7], and adverse event detection [8].

COVID-19 has been studied extensively - including the impact of its news coverage. In previous studies, [9] and [10] used topic modelling and sentiment analysis, while [11] analyzed discourse structures to study bias in media reporting to influence government policies and global reporting. Cresswell et al. [12] used social media data to analyse public sentiment towards the pandemic and determined media reporting played a major role in public sentiment. This study is unique in applying healthcare-specific deep learning networks, models, and embeddings - enabling the extraction and correlation of over 100 different types of medical entities with high accuracy. Following are the contributions of this study:

- Using fine-tuned NLP models to find most prevalent covid symptoms, prevention guidelines, research institutions, covid variants, and other entities for analysis.
- Geographical and demographic analysis of news coverage to understand most affected countries, population age groups, and professions.
- Putting the entire covid coverage of 2020 & 2021 on a timeline for each country, to understand temporal variation and correlation of case count and media coverage.
- Comparing the findings from unstructured text with the statistical data reported by WHO and analysing coherency.
- Correlating adverse reactions and drug brands by extracting and linking drug and reaction entities.

2. Analysis

The dataset comprises of 36,354 live blogs and key moments about COVID-19, published in 2020 and 2021. A live blog is a web page where news media outlets offer daily live coverage about an ongoing event. Each live blog consists of news stories and key moments. Journalists manually select the key moment stories from the whole set of news articles. We excluded the key moments from live blogs; hence, we can appropriately compare live blogs with key moments; otherwise, there will be some overlapping parts. To scrape the news articles, we leveraged web scraping algorithms by [13], and updated according to our use-case. Updated code and Spark NLP pipelines can be accessed [14]. As stated in [13], first, we collected the live blogs and then applied some pre-processing steps such as parsing HTML format, converting the dates, and extracting key moments from live blogs. The data is organized by title, text, date, URL, and key moment status fields. After data preparation, the executed NLP pipeline included sentence segmentation, tokenization, calculating embeddings, multiple named entity recognition steps, and relation extraction.

In the following sections, we'll report the analysis outcomes that we run over the news coverage by age demographics, news cycle evolution over time and the reporting and impact of adverse drug & vaccine events.



Figure 1: Monthly Distribution of the News from Jan'20 - Dec'21.

2.1. General News Analysis

We tracked the live blogs of CNN from 22nd January 2020 and The Guardian from 24th January 2020; until the end of 2021. The distribution of news by sites and years is explained in Figure 1. Upon inspection, it is apparent that the number of news was high in the periods when COVID-19 first appeared and the first deaths had been reported in and outside Asia in early 2020. News of CNN reached its peak point in March 2020 when COVID-19 reached the shores of United States. Other factors of high news coverage included the massive spread in Europe (especially Italy) and promulgation of the virus as a pandemic by the WHO. Conspicuously, CNN COVID-19 reporting gradually decreased after their peak point and became very seldom in late 2021. With all these, as an overall evaluation, a monthly average of COVID-19 live blogs and key moments is 807, and the median is 365.

If we look at the types of news, it primarily consists of live blogs while the key moments constitute for a minority portion. When we combine them for analysis, we get more than 5.3M words in the corpus to create a story.

2.2. Demographics and Geographical Analysis by Named Entities

In order to analyse the news coverage by certain geographical entities, we used a pre-trained named entity recognition (NER) model called *ner_deid*¹ from Spark NLP for Healthcare library to extract country names mentioned in the news. Using WHO's official daily statistics [15], we gathered the number of actual COVID-19 cases and deaths for every half of 2020 and 2021, and then compared with the top 12 countries extracted from the news by the *ner_deid* model.

Figure 2 shows the normalized metric for each country by dividing its value by the maximum value in the respective six-month period. As it is seen on the chart, the most reported cases are

¹https://nlp.johnsnowlabs.com/2021/09/03/ner_deid_subentity_augmented_en.html

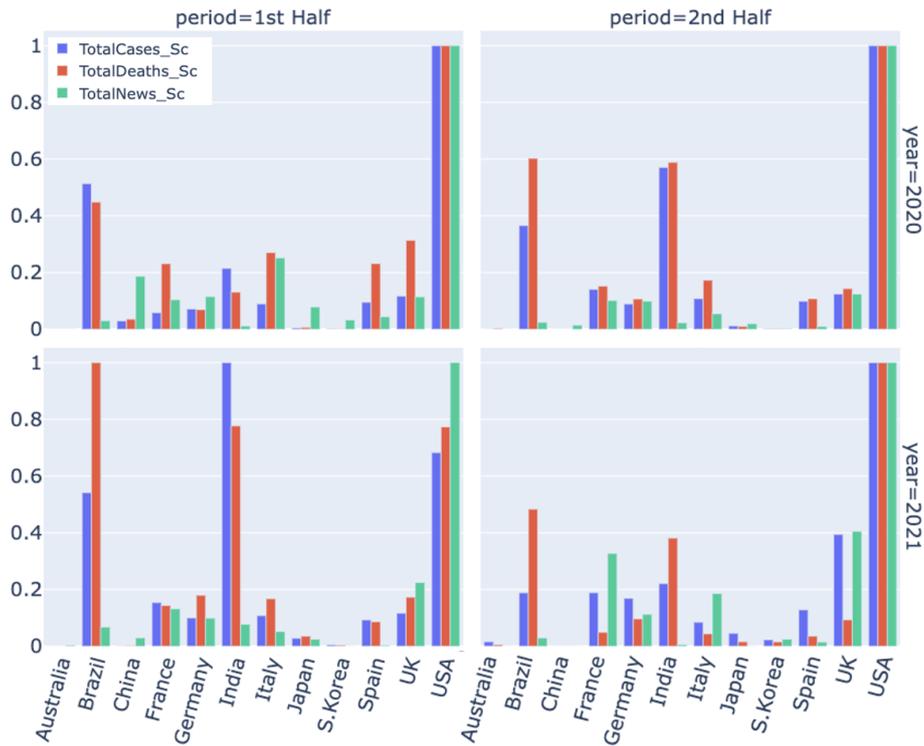


Figure 2: Total number of Cases, Deaths, and News (Scaled by Periods)

seen in the USA, India, and Brazil, respectively. It is also seen that the number of patients who died is also among these three countries, although Brazil ranks second on the death list.

The graph can be interpreted in the following way: (1) If the lengths of the bars are in agreement with each other, it can be deduced that the cases reported and deaths reported are in agreement with the number of news broadcasts in CNN and the Guardian. (2) If the blue and red bars are compatible with each other but not with the green bar, there may be two reasons for a mismatch with the news number. The first reason could be that the country is far from the news source while the second reason could be that the total number of cases and deaths in the country are not considered remarkable by the news sources. (3) If the all the bars are inconsistent with each other, it may indicate that the country in question has not published the case and death numbers transparently. (4) Needless to say, the numbers regarding the countries having inconsistent figures (e.g. Brazil, Russia etc.) might also be explained by the fact that the dataset only covers the news articles from the US (CNN) and UK (Guardian), hence these countries are reported through the lens of other countries (US and UK editors).

Consequently, we can count United States, United Kingdom, France, Germany and Australia as the most compatible countries while China and Brazil can be counted as the most incompatible countries. For Spain, the total number of cases and deaths is consistent, but the volume of news articles is inconsistent. This may be explained by Spain being away from the two news sources.

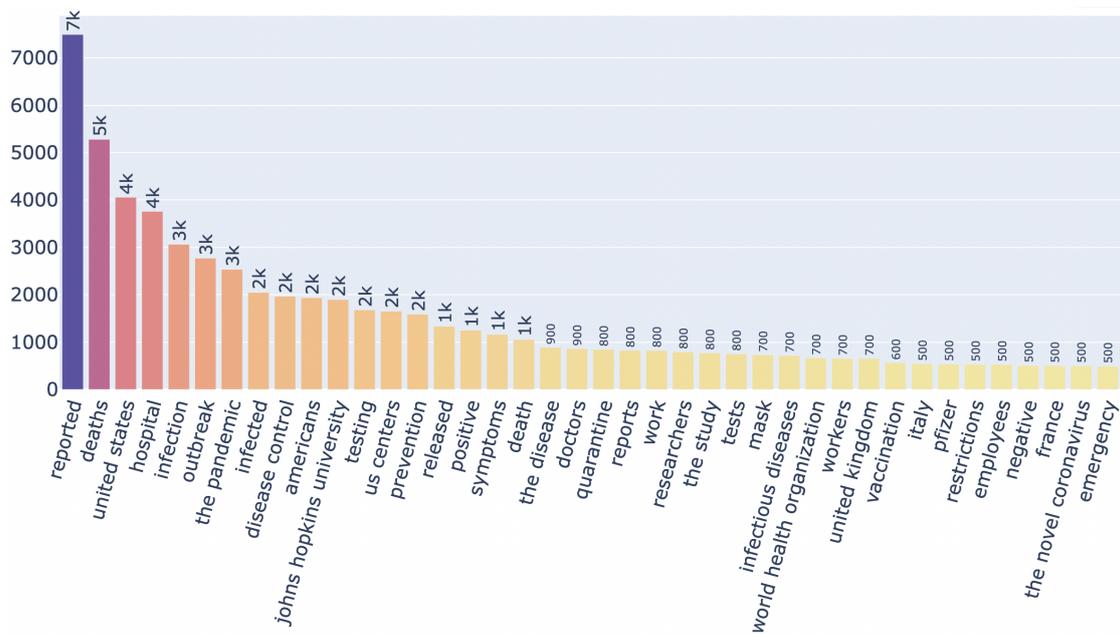


Figure 3: Top 40 Entities in Corpus

2.3. Behavioral and Clinical Analysis by Named Entities

In order to analyse the news with respect to the clinical entities mentioned, we used a pretrained NER model named *ner_events_clinical*² and *ner_jsl*³ to extract major clinical events. These two NER models combined can extract more than one hundred different types of clinical entities at once, with one line of code, enabling a deeper analysis than previous studies of this type.

Figure 3 shows the top 40 entities by frequencies in the corpus. Entities related to coronavirus (deaths, infections, preventive measures, and geographical locations etc) seem to be the most prevalent, followed by entities related to research and development of drugs. When we look at the ranking of countries in this list, the United States holds first place with the most reported number of cases, while the United Kingdom and Italy are in the second and third places - aligning with the WHO stats illustrated in Figure 2. The only university listed is John Hopkins University (JHU), primarily because of its role in organization and coordination of the Coronavirus Resource Center [16]. We also see major vaccine manufacturers in the list due to high coverage of vaccine development and trials.

Figure 4 shows the Top 10 treatments, symptoms, drugs, and procedures in the order of their frequencies. It is clear that concepts like vaccine, wearing masks, oxygen support, and ventilators appear often in treatments. These entities along with isolation, quarantine, social distancing, and self-isolation, which are at different ranks in the list, have a significant place in disease control.

On the other hand, the symptoms presented in Figure 4 are widespread. People who observe

²https://nlp.johnsnowlabs.com/2021/03/31/ner_events_clinical_en.html

³https://nlp.johnsnowlabs.com/2020/04/22/ner_jsl_en.html

treatments	treatment occurrence	symptoms	symptom occurrence	drugs	drug occurrence	procedures	procedure occurrence
wear masks	1842	cough	243	pfizer and biontech	594	intubation	67
isolation	777	blood clot	214	astrazeneca	275	inoculations	33
quarantine	658	sneeze	97	hydroxychloroquine	228	dialysis	18
vaccination	563	confusion	90	moderna	131	resuscitation	12
oxygen support	482	fatigue	86	chloroquine	57	bypass	10
ventilators	445	shortness of breath	63	novavax	57	autopsies	9
social distancing	342	headache	57	sanofi	49	organ transplants	8
intensive care	248	flu-like symptoms	56	tamoxifen	49	cpr	7
the pfizer/biontech vaccine	161	unwell	50	sinovac	40	hand washing	4
self-isolate	136	anger	49	hand sanitizer	31	gtr	3

Figure 4: Top 10 Treatments, Symptoms, Drugs, and Procedures

any of these symptoms turn to hospitals or health institutions to find out if they have COVID-19. Coughing, blood clotting, sneezing, fatigue, shortness of breath, and headache are the prevailing COVID-19 symptoms. It is worth noting that even a headache alone was enough to arouse the suspicion of coronavirus during this period. However, very specific symptoms related to loss of smell and taste are also observed, although their frequency is lower.

When analyzing named entities related to drugs, the companies producing vaccines stand out. Note that drug names also include drug brand names, drug manufacturers, and drug ingredients. The data suggests that the manufacturer's name is more prominent compared to the drug's scientific name in news coverage. Hydroxychloroquine and chloroquine were granted an emergency use authorization (EUA) on March 28, 2020 for treating COVID-19 cases. However, on June 15, 2020, FDA revoked its use [17], limiting it only to hospitalized patients under heart monitoring.

The most commonly mentioned medical procedures feature intubation and inoculation in the top two places. Intubation has been the most frequently applied procedure for patients admitted to the intensive care unit during this period.

The pie chart in Figure 5 shows that the most mentioned age group is 60 years and above, followed by adolescents. This is probably due to the fact that COVID-19 has higher severity in older people. According to scientific observations, WHO declared that COVID-19 is often more severe in people who are older than 60 years or who have health conditions like lung or heart disease, diabetes, or conditions that affect their immune system [18].

Comparing the most prevalent entities in 2020 versus 2021 - the coronavirus, which was in the first place in 2020, fell to second place in 2021, ceding the first place to the COVID-19 vaccine. The word 'reported' fell from the second most encountered entity in 2020 to the fourth in 2021. This decline can be attributed to less willingness to reporting of COVID-19 cases since according the [19] the trend of reported cases is not declining. On the contrary, there is an increase in both the number of new cases and the number of deaths.

In addition to healthcare-specific named entity recognition and resolution models, Spark NLP also provides unsupervised key phrase extraction from free text. Results indicate that the entity counts follow the general trend where COVID-19 infection is the most prevalent term, followed by entities related to death, geographical locations, media outlets, vaccine development, and vaccination reports.

COVID-19 also highlighted specific professions namely the medical doctors and nurses. Our NER analysis show that researchers, workers, teachers and government officials are the most

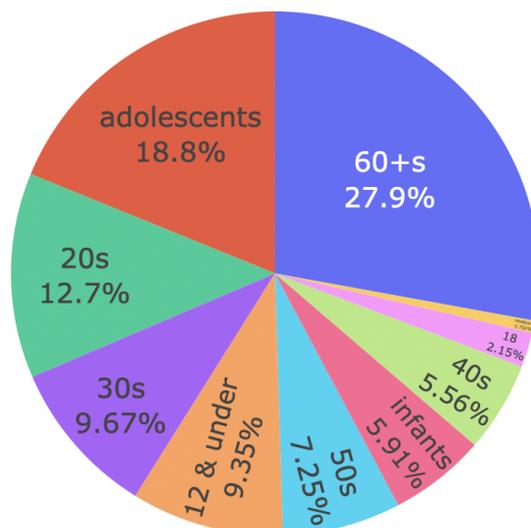


Figure 5: Age groups mentioned in news

cited professions in the news. Understandably, the most commonly mentioned people in the news during this period are healthcare professionals.

2.4. Impact of Adverse Drug and Vaccine Events

Adverse drug reactions/events (ADR/ADE) have a major impact on patient outcomes and healthcare costs. An analysis of extracted Adverse Drug Events versus drug and vaccine manufacturers reported by major news outlets has an impact on vaccine hesitancy as well as people’s reactions to medications and public health measurements dictated by regulatory agencies. In order to analyse ADEs as well the adverse events of vaccines mentioned in the news, we used a pre-trained NLP pipeline named *explain_clinical_doc_ade*⁴ that comes with the Spark NLP for Healthcare library. This NLP pipeline is the first scalable end-to-end solution for mining ADE’s from unstructured text, including Document Classification, Named Entity Recognition, and Relation Extraction Models within a unified NLP pipeline [8].

Figure 6 is a heatmap of correlation between vaccines and adverse reactions. When we explore the news dataset, we see that some of the vaccines are listed as their manufacturer’s name, but some are only with generic titles such as COVID-19 vaccine or mRNA vaccines. Since the heavy use of such general entities dilutes the results, Figure 7 shows the correlation while eliminating general vaccine names.

One of the most prominent observation the entanglement of clotting with AstraZeneca and Johnson & Johnson. This phenomenon was widely reported by the media, suspending the use of the drugs by these companies [20]. General allergic reactions are more common for vaccines from Pfizer & BioNTech and Moderna [21].

⁴https://nlp.johnsnowlabs.com/2021/07/15/explain_clinical_doc_ade_en.html

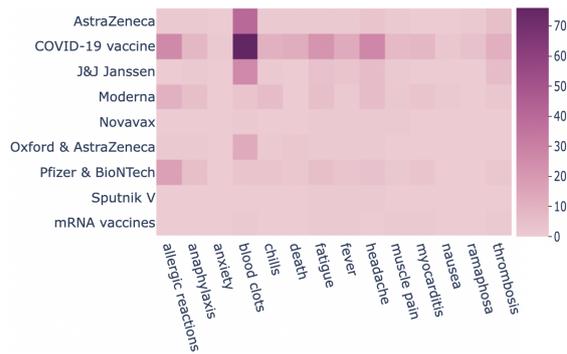


Figure 6: ADE of Vaccines

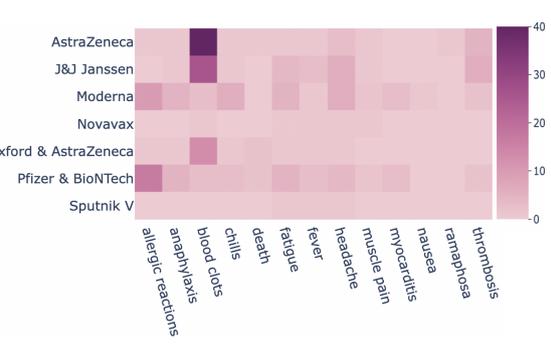


Figure 7: ADE of Vaccines without Generic Titles

3. Conclusion

As a global pandemic, COVID-19 has caused millions of deaths and impacted most people on earth. The media has played a major role in shaping public awareness, education, and opinion. Due to the growing number of media channels, the constant barrage of news, the prevalence of social media in sharing and shaping the news, and frequent shifts in regulations dictated by governments to combat the pandemic, analysing such a sheer volume of information objectively in a short amount of time can be one of the most practical current applications of NLP. In this study, we analyse news coverage from two prominent media outlets by using the clinical and biomedical NLP models of Spark NLP for Healthcare.

The analysis shows that the total number of cases and deaths reported by the news are mostly consistent with the numbers shared by WHO officially. The pre-trained NLP models perform well on extracting the most relevant terms and entities that have been widely used in the news. One tangible outcome of this study can be the automated mining of the adverse reactions to drugs and vaccines that are used to combat the virus. The analysis shows the correlations between the prominent vaccine manufacturers and reported adverse events just by relying on the news articles used for this study. Future research can broaden this automated analysis to include additional news sources, news in languages other than English, dealing with fake news, or to correlating changes in news coverage to resulting changes in public opinion and behavior.

References

- [1] W. H. Organization, et al., The corona virus disease 2019 (covid-19) (2020).
- [2] A. Laing, The h1n1 crisis: roles played by government communicators, the public and the media, *Journal of Professional Communication* (2012).
- [3] K. J. Mach, R. Salas Reyes, B. Pentz, J. Taylor, C. A. Costa, S. G. Cruz, K. E. Thomas, J. C. Arnott, R. Donald, K. Jagannathan, et al., News media coverage of covid-19 public health and policy information, *Humanities and Social Sciences Communications* 8 (2021) 1–11.
- [4] V. Kocaman, D. Talby, Spark nlp: Natural language understanding at scale, *Software Impacts* 8 (2021) 100058. doi:<https://doi.org/10.1016/j.simpa.2021.100058>.

- [5] V. Kocaman, D. Talby, Biomedical named entity recognition at scale, in: International Conference on Pattern Recognition, Springer, 2021, pp. 635–646.
- [6] V. Kocaman, D. Talby, Improving clinical document understanding on covid-19 research with spark nlp, in: SDU (Scientific Document Understanding) workshop at AAAI 2021, CEUR Workshop Proceedings, 2020. arXiv:2012.04005.
- [7] H. U. Haq, V. Kocaman, D. Talby, Deeper clinical document understanding using relation extraction, in: SDU (Scientific Document Understanding) workshop at AAAI 2022, CEUR Workshop Proceedings, 2021. arXiv:2112.13259.
- [8] H. U. Haq, V. Kocaman, D. Talby, Mining adverse drug reactions from unstructured mediums at scale, in: W3PHIAI workshop at AAAI-22, 2022. arXiv:2201.01405.
- [9] K. Krawczyk, T. Chelkowski, D. J. Laydon, S. Mishra, D. Xifara, B. Gibert, S. Flaxman, T. Mellan, V. Schwämmle, R. Röttger, J. T. Hadsund, S. Bhatt, Quantifying online news media coverage of the covid-19 pandemic: Text mining study and resource, *J Med Internet Res* 23 (2021) e28253. doi:10.2196/28253.
- [10] P. Ghasiya, K. Okamura, Investigating covid-19 news across four nations: A topic modeling and sentiment analysis approach, *IEEE Access* 9 (2021) 36645–36656. doi:10.1109/ACCESS.2021.3062875.
- [11] A. H. Abbas, Politicizing the pandemic: A schemata analysis of covid-19 news in two selected newspapers, *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique* (2020). URL: <https://doi.org/10.1007/s11196-020-09745-2>. doi:10.1007/s11196-020-09745-2.
- [12] K. Cresswell, A. Tahir, Z. Sheikh, Z. Hussain, A. Domínguez Hernández, E. Harrison, R. Williams, A. Sheikh, A. Hussain, Understanding public perceptions of covid-19 contact tracing apps: Artificial intelligence-enabled social media analysis, *J Med Internet Res* 23 (2021) e26618. doi:10.2196/26618.
- [13] A. Pasquali, R. Campos, A. Ribeiro, B. Santana, A. Jorge, A. Jatowt, Tls-covid19: A new annotated corpus for timeline summarization, in: European Conference on Information Retrieval, Springer, 2021, pp. 497–512.
- [14] https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/academic/Text2Story_CEUR_Workshop_2022_April.ipynb, 2022.
- [15] W. H. Organisation, World health organisation, 2022. <https://covid19.who.int/table>.
- [16] J. H. University, John hopkins university coronavirus center, 2022. <https://coronavirus.jhu.edu/map.html>.
- [17] FDA, Fda, 2022. <https://www.fda.gov/drugs/drug-safety-and-availability/fda-cautions-against-use-hydroxychloroquine-or-chloroquine-covid-19-outside-hospital-setting-or>.
- [18] W. H. Organisation, World health organisation, 2022. <https://www.who.int/westernpacific/emergencies/covid-19/information/high-risk-groups>.
- [19] W. H. Organisation, Who coronavirus (covid-19) dashboard, 2022. <https://covid19.who.int/>.
- [20] J. Wise, Covid-19: European countries suspend use of oxford-astrazeneca vaccine after reports of blood clots, 2021.
- [21] S. Meo, I. Bukhari, J. Akram, A. Meo, D. C. Klonoff, Covid-19 vaccines: comparison of biological, pharmacological characteristics and adverse effects of pfizer/biontech and moderna vaccines., *Eur Rev Med Pharmacol Sci* (2021) 1663–1669.