Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 177-190 recibido 13-12-2020 revisado 24-01-2021 aceptado 02-02-2021

Discovering topics in Twitter about the COVID-19 outbreak in Spain

Descubriendo temas en Twitter sobre el brote del COVID-19 en España

Marvin M. Agüero-Torales[1], **David Vilares**[2], **Antonio G. López-Herrera**[1]
[1]University of Granada, Spain
[2]Universidade da Coruña, CITIC, Spain
maguero@correo.ugr.es, david.vilares@udc.es, lopez-herrera@decsai.ugr.es

Abstract: In this work, we apply topic modeling to study what users have been discussing in Twitter during the beginning of the COVID-19 pandemic. More particularly, we explore the period of time that includes three differentiated phases of the COVID-19 crisis in Spain: the pre-crisis time, the outbreak, and the beginning of the lockdown. To do so, we first collect a large corpus of Spanish tweets and clean them. Then, we cluster the tweets into topics using a Latent Dirichlet Allocation model, and define generative and discriminative routes to later extract the most relevant keywords and sentences for each topic. Finally, we provide an exhaustive qualitative analysis about how such topics correspond to the situation in Spain at different stages of the crisis.

Keywords: COVID-19, Twitter, social networks, topic modeling.

Resumen: En este trabajo, analizamos lo que los usuarios han estado discutiendo en Twitter durante el comienzo de la pandemia causada por el COVID-19. Concretamente, analizamos tres fases diferenciadas de la crisis del COVID-19 en España: el propio tiempo de pre-crisis, el estallido de la enfermedad y el confinamiento. Para llevar esto a cabo, primero recolectamos una gran cantidad de tuits que son preprocesados. A continuación, agrupamos los tuits en distintas temáticas usando un modelo de Latent Dirichlet Allocation, y definimos estrategias generativas y discri–minativas para extraer las palabras clave y oraciones más representativas para cada tema. Finalmente, incluimos un exhaustivo análisis cualitativo sobre dichos temas, y cómo estos se corresponden con distintas problemáticas surgidas en España en distintos momentos de la crisis.

Palabras clave: COVID-19, Twitter, redes sociales, modelado de temas.

1 *Introduction*

The outbreak of the SARS-CoV-2 virus and the global spread of the COVID-19 disease has encouraged people and organizations to express their opinion, discuss topics and warn about the evolution of the pandemic in social media platforms such as Twitter.

Unlike previous occasions, such as SARS-CoV in 2002 (World Health Organization (WHO), 2020b), where social media still were in an embryonic state and natural language processing (NLP) still had limited practical applications; we are now in a situation where users generate a vast amount of written content, that can be analyzed by automatic tools to discover the topics societies care about, and their sentiment. This has been already the case for some precedent events or catastrophes in recent years, such as the 2016 US political elections (Grover et al., 2019) or some natural disasters, such as the 2011 East Japan Earthquake (Neubig et al., 2011).

In relation to the COVID-19 pandemic, a few specific NLP workshops (Verspoor et al., 2020b; Verspoor et al., 2020a) have already attempted to highlight how NLP can be used to respond to situations like the current one; addressing a number of challenges that include mining scientific literature and social media analysis, among many others (Wang et al., 2020; Kleinberg, van der Vegt, and Mozes, 2020; Afzal et al., 2020). With research purposes, there has been also efforts on releasing NLP datasets discussing COVID-19 topics (Chen, Lerman, and Fer-

ISSN 1989-7553. DOI 10.26342/2021-66-15

rara, 2020; Banda et al., 2020; Kerchner and Wrubel, 2020). In this context, the area of topic modeling has not been a stranger to this problem, and a number of authors have showed the options that clustering online posts such as tweets or Facebook messages can offer to monitor and evaluate the evolution of the pandemic through time (Asgari-Chenaghlu, Nikzad-Khasmakhi, and Minaee, 2020; Yin, Yang, and Li, 2020; Amara, Taieb, and Aouicha, 2020).

Contribution In this work, we also focus on the possibilities of performing effective and representative topic modeling over a large set of Spanish tweets. More particularly, we first collect a few millions tweets about COVID-19, mostly between 1 January to 20 April of 2020. Then, we apply latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003) to compute relevant topics in an unsupervised way, and obtain meaningful keywords and sentences through generative and discriminative routes. Finally, we provide an analysis to shed some light about the quality of the extracted topics, and how faithfully they represent what was happening in the Spanish society at different moments of the pandemic.

2 Related work

In what follows, we review topic modeling and NLP research related to COVID-19.

2.1 Topic modeling

In topic modeling, a topic is often viewed as a pattern of co-occurring words that can be exploited to cluster together documents from a large collection (Barde and Bainwad, 2017). Among methods for topic modeling we can find approaches such as the Vector Space Model (VSM) (Salton, Wong, and Yang, 1975), Latent Semantic Indexing (LSI) (Deerwester, 1988), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) or lda2vec (Moody, 2016). Related to this, one of the most well-known, standardized and widely-used methods is Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). More particularly, LDA is an unsupervised clustering approach where documents can belong to multiple topics, and where each topic is a mix of words, which can be shared among topics too.

The applications of these topic modeling approaches are many and include areas such as tag recommendation (Tuarob, Pouchard, and Giles, 2013), text categorization (Zhou, Li, and Liu, 2009), keyword extraction (Yijun and Tian, 2014), information filtering (Gao, Xu, and Li, 2014), similarity search in the fields of text mining (Pham, Do, and Ta, 2018), and information retrieval (Andrzejewski and Buttler, 2011).

2.2 Text Mining on English COVID-19 related tweets

With the COVID-19 outbreak, different authors have tried to apply topic modeling and text mining techniques to help analyze and monitor the situation of the pandemic, with a great focus on English messages. For instance, Asgari-Chenaghlu, Nikzad-Khasmakhi, and Minaee (2020) analyzed English tweets and detected the trending topics and major concerns of people with respect to COVID-19, by proposing a model based on the Universal Sentence Encoder (Cer et al., 2018). The model first derives a semantic representation and similarity of tweets and, over those similar tweets, it applies text summarization techniques to provide a summary of different clusters. In a related line, Yin, Yang, and Li (2020) proposed a framework to analyze the topic and sentiment changes in society over time due to the COVID-19, using Twitter to collect the source data. More specifically, they used a dynamic LDA for topic modeling over fixed time intervals (Blei and Lafferty, 2006), and VADER for sentiment analysis (Hutto and Gilbert, 2014). Chandrasekaran et al. (2020) examined the key topics among 13.9M English tweets about COVID-19, dealing with areas such as economy and markets, spread and growth in cases, treatment and recovery, impact on the healthcare sector, and governments response. They explored the trends and variations, and how those key topics, and associated sentiments changed over a period of time of 17 weeks, between 1 January 2020 and 9 May 2020. More particularly, they used guided LDA for topic modeling (Jagarlamudi, Daumé III, and Udupa, 2012), an LDA-variant where the model is guided to learn topics that are of specific interest, using priors in the form of seed words, and again VADER for sentiment analysis.

Also, Abd-Alrazaq et al. (2020) use LDA to detect topics such as the origin of the virus and its impact on people and countries, analyzing 2.8M English tweets. In addition, they performed sentiment analysis with `TextBlob` (Loria, 2020) and extracted some social network statistics for each topic, such as the number of followers, the number of likes of tweets, the number of retweets, the user mentions, or the link sharing, calculating the interaction rate per topic. At a smaller scale (100K English tweets) and considering only the pre-crisis lockdown period (from 12 December 2019 to 9 March 2020); Boon-Itt (2020) presented a work to understand public perceptions of the trends of the COVID-19 pre-pandemic time. The analysis included time series, sentiment analysis and emotional tendency using the NRC sentiment lexicon (Mohammad and Turney, 2013), as well as topic modeling using LDA.

2.3 Text Mining on Spanish and Multilingual COVID-19 related tweets

As usual in NLP, most of early efforts to monitor COVID-19 user-generated texts have focused on English. However, some work is already available for the Spanish language. For instance, Yu, Lu, and Muñoz-Justicia (2020) compare the news updates of two of the main Spanish newspapers Twitter accounts, *El País* and *El Mundo*, during the pandemic; applying topic modeling and network analysis methods. They identified eight news frames for each newspaper and split it in three clusters: the pre-crisis period (from 19 February to 14 March of 2020), the lockdown period (from 14 March to 11 May of 2020) and the recovery period (from 11 May to 3 June of 2020). Their goal was to understand how the Spanish news media covered the public health crisis in Twitter.

Besides, Carbonell Gironés (2020) proposed a geographical analysis of the opinion and influence of users in Twitter during the covid health crisis, considering tweets written in English and Spanish, and using LDA topic modeling. The first part of the study was a general approach to the analysis of the topics of US and UK users. The second part was an analysis of the interests of Twitter users in Spain during the confinement period (from 14 March to 22 July of 2020). To geolocate the tweets, they performed a country-level search for the English dataset, and a city or province-level search for the Spanish dataset; looking in both cases for any geographic references, both on the Twitter user location field and their biography.

Ordun, Purushotham, and Raff (2020) studied techniques to assess the distinctiveness of topics, key terms and features, as well as the speed of dissemination of retweets over time. They used pattern matching and topic modeling with LDA on a set of 5.5M of tweets written in multiple languages, resulting in 16 topics for English and one for Spanish, Italian, French and Portuguese, respectively. They also applied Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville, 2018) to identify clusters of distinct topics, which discuss case spread, healthcare workers, and personal protective equipment issues.

Beyond Twitter, Amara, Taieb, and Aouicha (2020) have exploited 22K Facebook posts to track the evolution of COVID-19 related trends, with a multilingual dataset that covers seven languages (English, Arabic, Spanish, Italian, German, French and Japanese). They applied an end-to-end analytic process for discovering language-dependent topics covering the duration of the pre-crisis period and part of lockdown (from 1 January to 15 May of 2020). The experiments showed that the extracted topics corresponded to the chronological development of what has been happening, and the measures that were taken in various countries.

3 Methods

In what follows, we describe the methodology of our work, decomposed into four steps: (i) the collection of the corpus, (ii) the language identification and geolocation of the tweets, (iii) the preprocessing, and (iv) the topic modeling approach and its analysis, clustering tweets into topics and extracting representative keywords and sentences.

3.1 Collection of tweets

We first defined a set of keywords to download relevant tweets: *coronavirus*, *COVID-19*, *COVID19*, *2019-nCoV*, *2019nCoV*. Further, as of March 3th, 2020 we added more keys: *SARS-CoV-2*, *SARSCoV2*, *CoV-19*, *CoV19*, *COVD19*, *COVD 19*, *corona virus*, *corona outbreak*.

More particularly, we collected a multilingual corpus of 32.68M tweets, including Twitter posts from 1 January of 2019[1] to 20

[1] In order to have some preceding context, but ex-

April of 2020; from all over the world. We scraped the tweets using the GetOldTweets-python3 (GOT3) library.[2] The reason to use this tool was that it allowed to retrieve old tweets without time limitation. However, the tool did not permit us to filter the retrieval by language. Besides, the Twitter Official API cannot retrieve tweets more than a week ago with a free subscription mode.[3]

3.2 Language identification and geolocation

The next step is language identification to keep only the Spanish tweets. We used four tools for detecting languages, since with GOT3 we could not obtain the language attribute. Those four tools were: `polyglot`,[4] `langdetect`,[5] `langid.py`,[6] and `fastText`.[7] The language is assigned based on majority voting. In case of a tie, we consider the tweet to be Spanish, except if all tools predicted a different language.

In total, we identified 5.35M Spanish tweets. In this work, we try to restrict the analysis to the content generated in Spain. For this purpose, we proceeded to filter the tweets in Spanish using the location attribute of the user profile, and look for the name of Spanish cities with more than 50K inhabitants, province names, autonomous regions names, and also any location specified as simply 'Spain'.[8]

After the cleaning process, we obtained ∼1.85M tweets for our topic modeling analysis. It is fair to point out that there is a percentage of tweets with a risk of not being correctly filtered, since the same place name might exist in more than one Spanish speaking country (e.g., 'Guadalajara' for Spain vs. 'Guadalajara' for Mexico). This is a common limitation on Twitter analyses, when it comes to analyze geolocated tweets (see for instance (Vilares and Gómez-Rodríguez, 2018)).

3.3 Preprocessing

We first proceed to lowercase the tweets and remove retweets. We also delete the keywords that were used to collect the tweets (see again §3.1) and other Twitter reserved words such as 'rt', 'fav', 'vía', 'nofollow', 'twitter', 'href' or 'rel'. Moreover, we removed stopwords, non-words (i.e., words compounded with characters that are not alphabet letters), URLs, numbers and punctuation marks. To do this, we used spaCy[9] to tokenize the words, and the Spanish and English stopwords lists from three libraries: NLTK,[10] stop-words,[11] and stopwordsiso.[12] Besides, in order to remove extra noise and cluster more clean topics, we only kept content words (i.e., nouns, verbs, adjectives, and adverbs).

Finally, to reduce word sparsity we used a custom lemmatizer[13] for Spanish, which applies a rule-based lemmatization with spaCy, and relies on Wiktionary,[14] which is a collaborative free-content multilingual dictionary. After the lemmatization step, the tweets whose length is less than three characters were removed. As traditional topic modeling approaches such as LDA, based on bag-of-words, suffer if many outliers are present (which happens in NLP due to the Zipf's law), we ignore terms that have a corpus frequency strictly less than three.

3.4 Topic modeling

For a more clear and comprehensive topic modeling analysis, we cluster the tweets in four weeks per month, except for the year 2019 (for which we collect the few tweets discussing coronavirus topics at that time), and the month of January 2020, which covers the first fortnight and not a week.

More particularly, we cluster the time of analysis into three phases. First, a pre-crisis phase, which includes tweets up to 24 January of 2020; when there was still few cases

pecting just to be able to retrieve a small number of tweets.

[2] `https://github.com/Jefferson-Henrique/GetOldTweets-python`

[3] However, we noted that GOT3 as of 18 September 2020 has been suspended due to the new Twitter policies on tweet payload

[4] `https://polyglot.readthedocs.io/en/latest/Detection.html`

[5] `https://pypi.org/project/langdetect/`

[6] `https://pypi.org/project/langid/`

[7] `https://fasttext.cc/docs/en/language-identification.html`. We used the large model

[8] We obtained the list of place names from the Instituto Nacional de Estadística (INE) `https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=525`

[9] `https://spacy.io/usage/v2-2` and the es_core_news_md language model

[10] `http://www.nltk.org/nltk_data/`

[11] `https://pypi.org/project/stop-words/`

[12] `https://pypi.org/project/stopwordsiso/`

[13] `https://github.com/pablodms/spacy-spanish-lemmatizer`

[14] `https://www.wiktionary.org/`

reported outside China. Second, we consider the outbreak phase, that we will consider to range from 25 January to 14 March of 2020; when the disease started to widely spread across Europe and the rest of the world, but Spain still was not under confinement. This is the period of time where the pandemic information, epidemic back then, was reported but was still not formally considered an alarm by the Spain government. Third, we cover about a month of the official lockdown period of the first wave (from 14 March to 20 April of 2020), when the Spanish government approved a strict social confinement.

As introduced previously, for topic modeling, we will be using *Latent Dirichlet Allocation* (LDA)[15] with collapsed Gibbs sampling inference (Griffiths and Steyvers, 2004); which processes raw text data in an unsupervised fashion to cluster documents that discuss the same topic. We chose LDA because it is standard and has probed robust for many tasks (see also §2.2 and §2.3). For each phase, we will mostly group tweets into weeks,[16] and for each week we will be extracting 10 topics. On the one hand, our goal was to facilitate the comprehension and interpretability. On the other hand, it is worth to note that selecting too few topics would make the clusters very generic and unspecific, while choosing too many could make them too sparse, not representative, and hard to analyze qualitative (Steinskog, Therkelsen, and Gambäck, 2017). Yet, we explored what would be in theory an optimal number of topics for different weeks using three methods: (i) the KL divergence (Arun et al., 2010), (ii) the pairwise cosine distance (Cao et al., 2009), (iii) and the loglikelihood. In all cases the results returned that the ideal number was between 5 and 20 in most of cases.

LDA setup We sampled up to 1500 epochs, and we kept the rest of parameters to the default value in the LDA library we used, i.e., $\alpha : 0.1$, $\eta : 0.01$, where the first corresponds to the Dirichlet parameter for the distribution over topics, and the second to the Dirichlet parameter for the distribution over words.

3.5 Extracting top topic keywords and sentences

To extract the most representative keywords for each topic, we considered both generative (GS) (Equation 1) and discriminative (DS) (Equation 2) approaches:

$$\text{GS(w,z)} = P(w|z) \tag{1}$$

$$\text{DS}(w,z) = P(w|z)/[\max_{z' \neq z} P(w|z')] \tag{2}$$

where w represents a given word and z the topic at hand. In essence, the generative score allows to extract the words that are most representative for each topic independently, in a way that a given word could be relevant for one or more topics, potentially making such topics harder to differentiate among them. On the contrary, the discriminative score allows to represent a topic by a set of keywords that are very representative for such topic, but have little relevance for the remaining ones.

Although the top keywords for each topic are useful, they might provide a limited view of what is actually being discussed. To counteract this, we also defined a generative (Equation 3) and discriminative (Equation 4) routes to extract the most representative sentences (tweets) for each topic, ideally being able to determine the topic by simply reading a few documents. The motivation to define these two different routes is the same than the one we made to extract the top keywords.

$$\text{GS}_{\text{sent}}(s,z) = \sum_{w \in s} \text{GS}(w,z)/\text{Length}(s) \tag{3}$$

$$\text{DS}_{\text{sent}}(s,z) = \sum_{w \in s} \text{DS}(w,z)/\text{Length}(s) \tag{4}$$

where s is the input document, for which we consider its length, in order not to only select the longest documents; although in the case of Twitter this is less of an issue than in other topic modeling approaches that must deal with actual long documents.

The full code is available.[17]

Limitations Sociolinguistic studies that collect data from social media such as Twitter can suffer from biases that can be hard to measure, identify or correct. For instance, it is well-known that a small percentage of

[15] In particular, we rely on the `https://github.com/lda-project/lda` implementation

[16] As introduced before, we use week here in an informal sense, referring to periods of time of 7 days, but not necessarily from Monday to Sunday.

[17] `https://github.com/mmaguero/twitter-analysis`

Topic	Discriminative Keywords	Generative Keywords
‘W1’ (from January to December of 2019)		
2	respiratorio, enfermedad, gripe	respiratorio, gripe, enfermedad
Magnifica guia para diferenciar los sintomas que causa la gripe y otros virus respiratorios. Junto con la gripe siguen circulando rinovirus, virus respiratorio sincitial y coronavirus, entre otros. <URL>		
1	enfermedad, gripe, respiratorio	enfermedad, respiratorio, gripe
@user informa de 27 casos de neumonia atipica, probablemente virica, en Wuhan (Hubei, China) en fecha 31/12/2019. El SARS (coronavirus) se inicio asi en 2003. Habra que seguir evolucion y esperar el diagnostico. <URL>		
W2-3 (from 1 to 16 January of 2020)		
8	alerta, hospital, poner, red, oms, china, mundial, mundo	china, oms, alerta, hospital, poner, mundial, mundo, red
UN NUEVO CORONAVIRUS PONE EN ALERTA A CHINA <URL>vía @user		
5	confirmar, japón, chino, infección, caso, china, animal, aparición	caso, confirmar, japón, china, infección, chino, ciudad, identificar
Japón confirma el primer caso de coronavirus vía @user <URL>		
W4 (from 17 to 24 January of 2020)		
9	emergencia, declaración, declarar, organización, reunión, convocar, decisión, determinar	oms, emergencia, internacional, declarar, mundial, alerta, salud, china
La OMS no declaró la emergencia por el coronavirus <URL>		
1	millón, cuarentena, habitante, frenar, ampliar, pekín, transporte, aislar	china, ciudad, wuhan, millón, cuarentena, persona, cerrar, brote
Más de once millones de chinos, en cuarentena por el coronavirus <URL>		

Table 1: Some representative topics for the weeks corresponding to the **pre-crisis** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

Twitter users generate the majority of content (Wojcik and Hughes, 2019). In this line, we believe that many of the collected tweets have its origin in newspapers and journalists accounts, that condition how other users tweet about this topic on Twitter, and therefore the detected topics can be heavily dependent on how national media decide to spread the news. Yet, this is the natural behaviour of this network, and in this particular work we decided not to control for this variable.

4 *Results*

We consider sixteen sets of tweets (mostly grouped in a weekly basis), extracting the ten most representative topics for each one according to LDA. To refer the topics, we will represent them with the top eight keywords and the most salient tweets. For clarity, and due to the large amount of weeks and topics, we will just illustrate and analyze some relevant topics extracted by our approach for different weeks, and try not to repeat common topics that span through the whole period. Usernames and urls are cut due to anonymity and space reasons, respectively.

4.1 Pre-crisis time

During this pre-crisis time, it is possible to see how the model captures that the COVID-19 was still not a concern for the Spanish society, which perceived the disease as an external problem, as reflected in many of the extracted topics. For clarity, Table 1 illustrates some relevant topics with top keywords and tweets, but we briefly discuss the content of the table below. To assess the relevance of the topics, we will be matching those against news from the newspapers that were published at the time in different Spanish media.

‘W1’ (from January to December of 2019) For the year 2019, we only could extract a total of seven topics, since the corresponding subset of tweets related to COVID-19 or coronavirus was still tiny (a total of 43 tweets after preprocessing). Still, we believe the results are interesting, since we observed that at this time most of Spanish tweets dealing with coronavirus still had to do with veterinarian diseases or even the zoonosis of coronavirus (i.e., how it is transmitted between animals and humans through the air), Yet, we found a few relevant tweets about COVID-19 that started to show up. We illustrate this as part of Table 1.

W2-3 (from 1 to 16 January of 2020) This time can be considered as the start of the emergency (Agencia EFE, 2020). In this line, we observed how our model started to identify this situation as well, clustering tweets about the World Health Organization (WHO) alerts to hospitals about symptoms, procedures, etc., and also about the increase in the number of cases in China.

W4 (from 17 to 24 January 2020) The crisis started to expand and from our model we see how the topics differ from previous weeks (see the third group of rows in Table 1). For instance, it shows how China started to apply restrictions in many locations of its territory (e.g., Wuhan) (El Boletín, 2020).

4.2 Outbreak time

In this phase, we see how the LDA approach reflects emergency declarations, the first cancellations of massive events in Spain, as well as the first suspicious cases; causing in consequence an increase of the concern among the Spanish society, which started to look and ask for sanitary products. This is also the phase where the approach captures a transition from international to national concerns. We will breakdown this more in detail in the next paragraphs, matching again the topics against news from the newspapers to qualitatively verify the quality of the extracted topics. Table 2 illustrates such topics with the top keywords and tweets from the model.

W5 (from 25 to 31 January of 2020) During this week, the approach kept identifying online discussions about the WHO emergency declarations, considering COVID-19 as a global coronavirus threat (Pérez, 2020). Also, the approach extracted topics related to international restrictions, such as the airplane company Iberia suspending flights to Shanghai (CatalunyaPress.es, 2020), at the same time that Russia closed its frontiers with China (Ellyatt, 2020).

W6 (from 1 to 7 February of 2020) Following the trend of announcing emergency declarations, the model started to identify international issues, such as the infection and posterior death of Li Wenliang (BBC News, 2020), a Chinese doctor that alerted about the first cases of COVID-19 in December 2019, but also national ones; such as the confirmation of the first case of coronavirus in Spain, in the Canary Island of La Gomera (Linde, 2020). This matches the time where the number of cases seemed to start to spread (still slowly) all around the world.

W7 (from 8 to 14 February of 2020) During this week, the coronavirus started to have an important economic effect in Spain, which is reflected by the model, discovering topics that showed how users discussed the potential (finally confirmed during this week too) cancellation of the 2020 Mobile World Congress (MWC 2020), which usually takes place in Barcelona (Pardeiro, 2020). On the healthcare side, additional (few) cases started to be reported in Spain, such as in Mallorca, where it was reported the second Spanish case of COVID-19 (Bohórquez and Güell, 2020). During this and next weeks, we started to observe how there is a slow transition from international to national topics.

W8 (from 15 to 21 February of 2020) During this week, the topics where in line with those discussed in the previous weeks, such as the cancellation of the MWC 2020 (see Table 2) and its repercussion. This 'last-long' topics made sense at the time, since the cancellation of the MWC 2020 was the first massive event cancelled in Spain, with important economic consequences. Other international issues such as the sustained increase of cases in China or in the cruise ship Diamond Princess (Almoguera, 2020) seemed to occupy Twitter users during this time, too.

W9 (from 22 to 29 February of 2020) These are the final days before the lockdown period, and in retrospective, it is easy to see how some of the topics extracted reflected the immediate seriousness of the situation. We see how the model captures that the WHO advised to the public (World Health Organization (WHO), 2020a) to wash hands frequently. It is interesting to see in Table 2 how 'farmacia' (pharmacy) appears together with 'gel' (gel), 'lavarse' (to wash), 'mano' (hand) and 'alcohol' (alcohol), 'agotar' (to run out of) among the top keywords for the corresponding topic. In this context, it is well-known that these products were scarce in pharmacies and stores, and actually this problem lasted for long during the lockdown period. Also, related to the immediate seriousness of the situation, the model captured how despite of not being confined, the world economy started to suffer with the stocks set for the worst week since 2008 (Sano, 2020).

W10 (from 1 to 8 March of 2020) Just before the lockdown, we observe how among the topics extracted there are topics that we see everyday in the current pandemic life. For instance, as shown in Table 2, we kept seeing the importance of washing hands and keep a good hygiene with the use of soap (World Health Organization (WHO), 2020a). Also,

Topic	Discriminative Keywords	Generative Keywords
W5 (from 25 to 31 January of 2020)		
9	oms, emergencia, declarar, declaración, sanitaria, organización, comité, convocar	oms, internacional, emergencia, salud, declarar, alerta, mundial, china
Declara OMS emergencia por coronavirus - Vía @user <URL>		
1	vuelo, cerrar, suspender, frontera, kong, hong, rusia, aerolínea	china, vuelo, cerrar, suspender, brote, frontera, evitar, kong
Iberia suspende los vuelos a Shanghái por el coronavirus <URL>...		
W6 (from 1 to 7 February of 2020)		
4	alertar, acusar, news, silenciar, intentar, bbc, difundir, confusión	médico, china, chino, morir, alertar, wuhan, muerte, wenliang
Por favor lean. Porque esto no lo va a contar ningún medio que alerte sobre el coronavirus . <URL>...		
1	gomera, alemán, ingresado, contacto, jalisco, vitoria, ecuador, isla	caso, españa, gomera, paciente, sospechoso, hospital, salud, síntoma
En España ya tenemos un caso de coronavirus ,en La Gomera ,un alemán.		
W7 (from 8 to 14 February of 2020)		
6	mallorca, negativo, británico, ingresado, palma, princess, diamond, gomera	caso, mallorca, españa, crucero, paciente, confirmar, sospechoso, salud
Confirman un caso de coronavirus en Palma de Mallorca <URL>... <URL>		
3	sony, amazon, gsma, bajas, lg, nvidia, ericsson, intel	mobile, congress, barcelona, mwc, empresa, cancela, cancelar, sony
Tras las bajas de LG, Ericsson, NVidia, Amazon y Sony #coronavirus #MWC2020 <URL>...		
W8 (from 15 to 21 February of 2020)		
5	crucero, diamond, princess, pasajero, colombiano, camboya, evacuado, ucrania	crucero, cuarentena, japón, caso, diamond, princess, pasajero, wuhan
NUEVOS CASOS DE CORONAVIRUS EN CRUCERO DIAMOND <URL>... <URL>		
6	mobile, barcelona, cancelación, cancelar, maratón, evento, mwc, congress	mobile, china, tokio, barcelona, cancelar, cancelación, maratón, guerra
Suspenden el Mobile World Congress de Barcelona por el coronavirus <URL>... <URL>		
W9 (from 22 to 29 February of 2020)		
6	mano, farmacia, lavarse, gel, desinfectante, alcohol, agotar, carne	mascarillas, mano, gente, mascarilla, evitar, comprar, miedo, hospital
Cómo prevenir el #coronavirus . Lávate las manos, lávate las manos, lávate las manos.................. lávate las manos. <URL>... <URL>		
1	bolsa, economía, mercado, caída, ibex, pérdida, crecimiento, wall	china, bolsa, economía, mercado, crisis, mundial, impacto, económico
'Esto es mercado. Esto me pone' @user #bolsa #COVID19 <URL>...		
W10 (from 1 to 8 March of 2020)		
7	mano, metro, lavarse, higiene, agua, gel, jabón, lavar	mano, evitar, medido, contagio, mascarillas, covid, persona, tomar
me voy a lavar las manos que no quiero el coronavirus		
4	patología, contagioso, anciano, letalidad, diferencia, estacional, comparación, hambre	gripe, persona, año, morir, mortalidad, gente, matar, enfermedad
Se llama Virus Corona Patologías Previas		

Table 2: Some representative topics for the weeks corresponding to the **outbreak** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

'metro' (underground) is a top keyword of such topic, since at that time there was a discussion about the chances of getting infected (e.g., in the public transport) (CNN, 2020). In a different topic, we see what it seems to be a discussion comparing the flu and covid, and how they affect to the population, which was a popular comparison at the time.

4.3 Lockdown time

During the lockdown phase (until April), we can observe in Table 3 how the topics discussed mostly focused on the worst consequences of the pandemic, such as the big economic crisis, the large number of deaths per day, and also some collective actions such as thanking the healthcare workers. Again, we give a brief explanation below these lines, and match the topics against news in the media.

W11 (from 9 to 16 March of 2020) Here we consider the week where the Spanish society stopped to have free movement. More particularly, the government approved strict social confinement on 14 March of 2020 (Cué, 2020). Besides, the model found topics about the acknowledgement to the healthcare workers and the solidarity applause (La Razón, 2020), which was very popular in Spain dur-

Topic	Discriminative Keywords	Generative Keywords
W11 (from 9 to 16 March of 2020)		
3	aplauso, frenalacurva, aplausosanitario, cuarentenaya, yoelijoserresponsable, felizlunes, arena, agradecimiento	covid, yomequedoencasa, casa, quedateencasa, cuarentena, coronavirusesp, cuarentenacoronavirus, responsabilidad
Aplausos para que suenen más que los truenos que hoy hay en Madrid. Hoy mis aplausos para todos. Para \Saldremos de esta / #COVID19 <URL>		
8	ocasionado, aprobar, pymes, paliar, erte, fiscal, boe, hipoteca	covid, medido, crisis, gobierno, alarma, situación, sanitaria, empresa
#RealDecreto 463/2020 #estadodealarma #COVID19 <URL>#pymes #Autonomo #Cordoba @user <URL>		
W12 (from 17 to 24 March of 2020)		
6	respirador, fabricar, ifema, impresora, envío, coronavirius, epis, todosobremovil	covid, hospital, sanitario, mascarillas, madrid, personal, estevirusloparamosunidos, quedateencasa
#ElonMusk puede que empiece a fabricar respiradores #COVID19 <URL>		
1	higiene, jabón, distanciamiento, acatar, lavado, fanb, geacam, comerciales	covid, medido, evitar, contagio, prevención, propagación, salud, tomar
Entre más higiene se tenga, mayor es la protección ante los patógenos como el #COVID19 <URL>...		
W13 (from 25 to 31 March of 2020)		
1	erte, pago, despido, prestación, contrato, fiscal, ertes, alquiler	crisis, medido, gobierno, empresa, económico, trabajador, autónomo, sanitaria
Información para los afectados por ERTE debido al COVID19 . #ERTE #Coronavirus <URL ><URL>		
3	civil, guardia, desinfección, desinfectar, higiene, cumplimiento, jabón, estación	medido, persona, evitar, contagio, salud, seguridad, prevención, casa
Unos 400 guardias civiles con coronavirus en #CLM , según la @user @user -. Vía @user <URL>... <URL>		
W14 (from 1 to 7 April of 2020)		
1	animal, respiratorio, tigre, gato, mascota, zoo, bronx, contaminación	persona, paciente, enfermedad, síntoma, casa, contagio, evitar, matar
Si los tigres se contagian de coronavirus , ¡ojito los que tenéis gato!		
8	confirmado, cifra, elevar, defunción, diarios, ascender, activos, descender	caso, fallecido, españa, muerte, muerto, dato, número, país
637 muertes por coronavirus en un día, la cifra más baja en 13 días <URL>		
W15 (from 8 to 14 April of 2020)		
2	cifra, curado, reino, récord, ascender, contabilizar, diagnosticado, acumular	caso, fallecido, españa, muerto, muerte, dato, número, persona
Las 510 muertes por COVID-19 en un día, la cifra más baja desde el 23 de marzo <URL>		
5	johnson, intensivo, boris, testimonio, alta, universitario, clmpressdigital, sosprisiones	hospital, médico, paciente, sanitario, madrid, personal, persona, profesional
Coronavirus : Boris Johnson fue dado de alta. <URL>		
W16 (from 15 to 20 April of 2020)		
5	luis, sepúlveda, escritor, homenaje, chileno, fútbol, club, dep	año, morir, hospital, luis, fallecer, quedateencasa, yomequedoencasa, historia
Luis Sepúlveda muere por coronavirus <URL>... <URL>		
2	distanciamiento, prohibidorendirse, enestafamilianadieluchasolo, yonosoyungastosuperfluo, bicicleta, espandemia, comunidadvalenciana, saltarse	confinamiento, quedateencasa, yomequedoencasa, cuarentena, medido, casa, evitar, alarma
¿El distanciamiento social podría ir incluso más allá de 2021? #COVID19 #coronavirus <URL>		

Table 3: Some representative topics for the weeks corresponding to the **lockdown** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

ing the lockdown period. In a related line, topics like this one also captured the feeling of the importance of staying at home to prevent becoming infected and reduce the workload of these workers.

W12 (from 17 to 24 March of 2020) For this week, the model extracted topics discussing the personal hygiene measures to combat the COVID-19. The topics also reflect the lack of equipment in the hospitals, which was a problem at the beginning of the pandemic. More particularly, the model was able to identify as a topic the lack of ventilators in Spain, and also the rest of the world, as reflected by the most salient discriminative tweet. This matches the news at the time, which discussed the use of 3D printers to provide such ventilators (Polo, 2020), or hacking some objects to adapt them for medical use (Cristian Fracassi, 2020).

W13 (from 25 to 31 March of 2020) This week covers the last days of March 2020.

Due to the strict confinement, topics concerning job losses and the measures taken by the government to counteract the situation (e.g. the so-called ERTEs) started to arise (RTVE.es, 2020; Gestiona.es, 2020). Among the rest of the topics of this week, we also would like to remark the massive infection of public workers, such as the Guardia Civil officers in Castilla La-Mancha (EFE/CMM, 2020). The infection of public workers during this time of the pandemic was also widely discussed in the news (Requeijo, 2020).

W14 (from 1 to 7 April of 2020) On the national side, some topics reflected the number of casualties per day. More particularly, the beginning of April corresponded to the peak of the first wave, and the beginning of the decreasing trend in the number of infections and deaths per day (Justo, 2020). A bit on a different line, we found topics discussing more diverse aspects of COVID-19, such as the infection in the Zoo of Bronx (New York, USA) of tigers and lions (M.R.M., 2020).

W15 (from 8 to 14 April of 2020) Here, we would like to remark a topic related to an international breaking news, and more particularly about Boris Johnson (the UK Prime Minister) being infected by the coronavirus, together with his evolution, when he even entered the ICU (La Vanguardia, 2020). On the national side, the models kept detecting topics related to the number of deaths in Spain, which was still high and dynamic during that time, but reached some local minima these days (Soteras, 2020).

W16 (from 15 to 20 April of 2020) For the last days of our study, the model found relevant topics too, such as the death of the Chilean writer Luis Sepúlveda (Safont Plumed, 2020) due to COVID-19, or topics related to the need of keeping social distancing, maybe even for months (elEconomista.es, 2020), as reflected by some of the most representative tweets.

4.4 Quantitative evaluation

We performed a small human evaluation to quantitatively estimate the quality of the extracted topics. We took 20 topics randomly from all periods. Then, two annotators were in charge of: (i) determining if given the top 8 keywords and 3 top sentences made possible to infer a topic, (ii) determining if for each top topic word (according to the discriminative score) they belonged to the inferred topic, and (iii) the same as in (ii), but for the 3 most representative sentences. We calculated the percentage of times both annotators positively labelled a sample, obtaining scores of 80%, 56.88% and 71.66% for (i), (ii), and (iii), respectively. In addition, we calculated (ii) but taking into account only the first 3 top keywords of the topic, yielding a score of 75% of positive samples.

5 Conclusion

This paper used a topic modeling approach to shed some light about the topics discussed in Spain during the early stages of the COVID-19 pandemic, including a period of pre-crisis, the outbreak of the disease, and the beginning of the confinement. We collected a large amount of tweets using keywords and cleaned them to keep only Spanish tweets that were written in Spain. After that, we used a Latent Dirichlet Allocation model that learned to cluster such tweets according to the topic they discuss. To represent the topics, we used generative and discriminative routes to extract the most salient keywords and sentences. To verify the quality of the extracted topics, we performed a qualitative analysis matching the topics against relevant news in the newspapers at the same period of time, and a small quantitative evaluation. Overall, the topics show that during the pre-crisis period, users focused on the international panorama than the local situation, while during the outbreak and lockdown phases they focused the most on the Spanish emergency, considering health and economic problems.

Acknowledgements

MMAT has been partially funded by Barcelona Supercomputing Center (BSC) through the Spanish Plan for advancement of Language Technologies ‘Plan TL’ and the Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA). DV is supported by MINECO (TIN2017-85160-C2-1-R), by Xunta de Galicia (ED431C 2020/11), by Centro de Investigación de Galicia ‘CITIC’ (European Regional Development Fund-Galicia 2014-2020 Program, ED431G 2019/01), and by a 2020 Leonardo Grant for Researchers and Cultural Creators from the BBVA Foundation.

References

Abd-Alrazaq, A., D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah. 2020. Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *Journal of medical Internet research*, 22(4):e19016.

Afzal, Z., V. Yadav, O. Fedorova, V. Kandala, J. van de Loo, S. A. Akhondi, P. Coupet, and G. Tsatsaronis. 2020. CORA: A deep active learning covid-19 relevancy algorithm to identify core scientific articles. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Agencia EFE. 2020. La OMS pone en alerta a la red mundial de hospitales por un nuevo coronavirus en China. *www.efe.com*, January.

Almoguera, P. 2020. El coronavirus pone en jaque ahora a Japón y Corea del Sur. *El País*, February.

Amara, A., M. A. H. Taieb, and M. B. Aouicha. 2020. Multilingual topic modelling for tracking covid-19 trends based on facebook data analysis.

Andrzejewski, D. and D. Buttler. 2011. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 600–608.

Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer.

Asgari-Chenaghlu, M., N. Nikzad-Khasmakhi, and S. Minaee. 2020. Covid-transformer: Detecting trending topics on twitter using universal sentence encoder. *arXiv preprint arXiv:2009.03947*.

Banda, J. M., R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, K. Artemova, E. Tutubalina, and G. Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration, August.

Barde, B. V. and A. M. Bainwad. 2017. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 745–750.

BBC News. 2020. Li Wenliang: Coronavirus kills Chinese whistleblower doctor. *BBC News*, February.

Blei, D. M. and J. D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA. Association for Computing Machinery.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bohórquez, L. and O. Güell. 2020. El segundo caso de coronavirus en España es un británico que se contagió en los Alpes. *El País*, February.

Boon-Itt, S. 2020. A text-mining analysis of public perceptions and topic modeling during the covid-19 pandemic using twitter data. *JMIR public health and surveillance, JMIR Preprints. 30/06/2020:21978*.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang. 2009. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781.

Carbonell Gironés, L. 2020. Geographical analysis of the opinion and influence of users on twitter during the coronavirus health crisis. Final project/degree, Escola Tècnica Superior d'Enginyeria Informàtica, Universitat Politècnica de València.

CatalunyaPress.es. 2020. Iberia suspende los vuelos a Shanghái por el coronavirus.

Cer, D., Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. 2018. Universal sentence encoder.

Chandrasekaran, R., V. Mehta, T. Valkunde, and E. Moustakas. 2020. Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. *Journal of Medical Internet Research*, 22(10):e22624.

Chen, E., K. Lerman, and E. Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

CNN. 2020. Medidas globales por el coronavirus: mantener distancia de un metro, cierre de escuelas y museos, evitar los besos y otras, March.

Cristian Fracassi. 2020. Charlotte valve, March.

Cué, C. E. 2020. El Gobierno informa de que es la única autoridad en toda España, limita los desplazamientos y cierra comercios, March.

Deerwester, S. 1988. Improving information retrieval with latent semantic indexing.

EFE/CMM. 2020. 400 guardias civiles de Castilla-La Mancha tienen Covid-19, según la AUGC.

El Boletín. 2020. China pone en cuarentena a más de 30 millones de personas por el coronavirus. January.

elEconomista.es. 2020. Las medidas de distanciamiento social podrían extenderse hasta 2022 de manera intermitente - elEconomista.es.

Ellyatt, H. 2020. Russia closes border with China to prevent spread of the coronavirus, January.

Gao, Y., Y. Xu, and Y. Li. 2014. Pattern-based topics for document modelling in information filtering. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1629–1642.

Gestiona.es. 2020. Información para los afectados por ERTE debido al COVID19, March.

Griffiths, T. L. and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Grover, P., A. K. Kar, Y. K. Dwivedi, and M. Janssen. 2019. Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145:438–460.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hutto, C. and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, volume 81, page 82.

Jagarlamudi, J., H. Daumé III, and R. Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

Justo, D. 2020. España sigue la tendencia a la baja: 4.273 nuevos contagios por coronavirus y 637 muertes, April.

Kerchner, D. and L. Wrubel. 2020. Coronavirus Tweet Ids.

Kleinberg, B., I. van der Vegt, and M. Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

La Razón. 2020. Emotivo reconocimiento a los sanitarios en forma de aplausos desde los balcones, March.

La Vanguardia. 2020. Boris Johnson recibe el alta y continuará recuperándose de la Covid-19 en su casa, April.

Linde, P. 2020. Sanidad confirma en La Gomera el primer caso de coronavirus en España. *El País*, February.

Loria, S. 2020. textblob documentation. *Release 0.16*, 2.

McInnes, L., J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February.

Mohammad, S. M. and P. D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Moody, C. E. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec.

M.R.M. 2020. Un tigre del zoo de Nueva York tiene coronavirus, April.

Neubig, G., Y. Matsubayashi, M. Hagiwara, and K. Murakami. 2011. Safety information mining—what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973.

Ordun, C., S. Purushotham, and E. Raff. 2020. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.

Pardeiro, M. 2020. El fracaso político del MWC: ”No se va a suspender”. ”No cuelga de un hilo”.

Pham, P., P. Do, and C. D. Ta. 2018. W-pathsim: novel approach of weighted similarity measure in content-based heterogeneous information networks by applying lda topic modeling. In *Asian conference on intelligent information and database systems*, pages 539–549. Springer.

Polo, J. 2020. Coronavirus: La Zona Franca fabricará 100 respiradores diarios con impresoras 3D, March.

Pérez, B. 2020. La OMS rectifica y declara la emergencia global por el coronavirus, January.

Requeijo, A. 2020. La Policía y la Guardia Civil suman ya más de 400 positivos por coronavirus, March.

RTVE.es. 2020. Los ERTE por la crisis del coronavirus suman más de 240.000, March.

Safont Plumed, J. 2020. Muere el escritor chileno Luis Sepúlveda, a causa del coronavirus.

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November.

Sano, H. 2020. GLOBAL MARKETS-World stocks set for worst week since 2008 as virus fears grip markets. *Reuters*, February.

Soteras, A. 2020. COVID-19: 510 muertes en un día, la cifra más baja desde el 23 de marzo.

Steinskog, A., J. Therkelsen, and B. Gambäck. 2017. Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st nordic conference on computational linguistics*, pages 77–86.

Tuarob, S., L. C. Pouchard, and C. L. Giles. 2013. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 239–248.

Verspoor, K., K. B. Cohen, M. Conway, B. de Bruijn, M. Dredze, R. Mihalcea, and B. Wallace, editors. 2020a. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Verspoor, K., K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, and B. Wallace, editors. 2020b. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Vilares, D. and C. Gómez-Rodríguez. 2018. Grounding the semantics of part-of-day nouns worldwide using twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 123–128.

Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Wojcik, S. and A. Hughes. 2019. Sizing up twitter users. *PEW research center*, 24.

World Health Organization (WHO). 2020a. Advice for the public on COVID-19 – World Health Organization.

World Health Organization (WHO). 2020b. WHO statement regarding cluster of pneumonia cases in Wuhan, China. January. Accessed: 2020-08-28.

Yijun, G. and X. Tian. 2014. Study on keyword extraction with lda and textrank combination. *Data Analysis and Knowledge Discovery*, 30(7):41–47.

Yin, H., S. Yang, and J. Li. 2020. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. *arXiv preprint arXiv:2007.02304*.

Yu, J., Y. Lu, and J. Muñoz-Justicia. 2020. Analyzing spanish news frames on twitter during covid-19—a network study of el país and el mundo. *International Journal of Environmental Research and Public Health*, 17(15):5414.

Zhou, S., K. Li, and Y. Liu. 2009. Text categorization based on topic model. *International Journal of Computational Intelligence Systems*, 2(4):398–409.