H-TFIDF: What makes areas specific over time in the massive flow of tweets related to the covid pandemic?

Rémy Decoupes, Rodrique Kafando, Mathieu Roche, Maguelonne Teisseire

AGILE: GIScience Series, 2, 2, 2021. https://doi.org/10.5194/agile-giss-2-2-2021
Proceedings of the 24th AGILE Conference on Geographic Information Science, 2021.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras.
This contribution underwent peer review based on a full paper submission.

H-TFIDF: What makes areas specific over time in the massive flow of tweets related to the covid pandemic?

Rémy Decoupes[a](corresponding author), Rodrique Kafando[a], Mathieu Roche[a,b] and Maguelonne Teisseire[a]

remy.decoupes@inrae.fr, rodrique.kafando@inrae.fr, mathieu.roche@cirad.fr, maguelonne.teisseire@inrae.fr

[a] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[b] CIRAD, F-34398 Montpellier, France

Abstract.

Data produced by social networks may contain weak signals of possible epidemic outbreaks. In this paper, we focus on Twitter data during the waiting period before the appearance of COVID-19 first cases outside China. Among the huge flow of tweets that reflects a global growing concern in all countries, we propose to analyze such data with an adaptation of the TF-IDF measure. It allows the users to extract the discriminant vocabularies used across time and space. The results are then discussed to show how the specific spatio-temporal anchoring of the extracted terms make it possible to follow the crisis dynamics on different scales of time and space.

Keywords. TF-IDF, Hierarchical analysis, Pandemic situation, social network

1 Introduction

Social networks have become useful to detect and analyze events, dangers or threats like disease outbreaks, nature hazards, social movements, etc (Tsou and Leitner (2013)). Users of these platforms can be seen as citizen journalists or sensor observations (Nagarajan et al. (2009)). Their reporting cover three dimensions: theme, spatial and temporal. Unlike other social networks, Twitter has the particularity of allowing its users to access, comment and contribute to all topics or threads (Li et al. (2012)).

Analysing such big amount of information is still challenging due to the specificity of the messages (i.e. specific vocabulary, shortness of the message). Most works focus on identifying trends, top-k most significant words over a period of time, the most representative words for a city or for a country, to a certain month, in a year, and etc. Nevertheless, what is crucial for a pandemic analysis is to exactly identify the discriminant vocabulary over space and time by exploring the possible specificities without any a priori. The decision maker needs include a dynamic analysis allowing him to navigate through the spatial and temporal dimensions (Depoux et al. (2020)).

That is the main objective of the work presented in this paper. More precisely, the two following points are detailed: (1) identifying the specific terms and (2) focusing on the spatial and temporal navigation that could improve the tweet analysis. We illustrate these various ways of exploring tweet corpus in the health context of the coronavirus COVID-19 pandemic. By using an adaptive interest measure, our proposal offers a global view of the evolution of words over space and time. Experiments applied on a public dataset underline that some lessons could be learned.

The interest of our proposal is illustrated through the analysis of the top extracted terms for Greece (See Section 4.3 for more details). In the TOP10 terms at the end of February, the words "thessaloniki" and "carnival" appear. The first term refers to the first COVID-19 case in Greece who have been treated in the Thessaloniki hospital. Tweets containing this term report panic buying of antiseptic in this area two days after. Following up "carnival" highlights the Greek government decision to cancel the Xanthi Carnival. Our approach extracts terms mainly used for a certain region and period that report information about the local situation, the crisis management and opinions

of inhabitants. These information would be hard to discovered using methods from the literature (See Section 4.3 for details). Hence, proposing navigation on different levels of a hierarchy could be very helpful for the epidemiological surveillance.

The rest of the paper is organized as follows. Section 2 is devoted to related work. Section 3 details the specific measure for tweet analysis. Some experiments and results are presented in Section 4. Finally, we conclude in Section 5 with some future works.

2 Related work

From the first signs of the emergence of the COVID-19 disease, the panic may spreads faster than the virus itself. The weight of public sentiment have sometimes led to disproportionate decisions in regard of the public health needs (Depoux et al. (2020)). The authors call for public health agencies to equip themselves with tools to monitor alerts and concerns and lead spatio-temporal analysis of social media in order to adapt their communication to specific regions. Although this is a rich source of information, tweets are still very noisy and sometimes meaningless (Li et al. (2012)). These difficulties are exacerbated by the discussions generated by the pandemic crisis, on which the whole world focuses its attention (Alshaabi et al. (2021)) and shares similar terms among countries like coronavirus, mask, lockdown. To meet the needs of public health, it is therefore necessary to be able to extract the terms specifically used by regions among the massive flow of common words.

Based on linguistics and statistics approaches, metrics are developed to detect discriminating terms. Twitter-based monitoring or alerting systems must adapt to detect change in the terminology used by users. In (Le Sceller et al., 2017), authors want to follow cyber-security news by enriching their list of keywords used to query Twitter. They track co-occurrences of term with their list to extract candidates which would be selected according to a threshold on TF-IDF score (See Section 3). This method deals with the thematic dimension. Other studies adapt the TF-IDF measure to the time dimension. The authors in (Alsaedi et al., 2016) are facing computational issues when they try to sum-up new tweets by extracting TF-IDF top five scores in a time interval. Indeed, with each new tweet, all TF-IDF score must be updated for the whole corpus. When the computation time get longer, it becomes difficult to follow the real time flux. In (Erra et al., 2015), the authors propose an approximate version of the TF–IDF measure for data stream. Following the hypothesis of a fast response needed and a memory small and limited, an approximative TF-IDF measure is computed showing interested performances compare to the classic one.

Retrospective spatio-temporal analysis of tweets contents need discriminant methods for extracting terms as well. Working on past trending topics, the authors of (Nagarajan et al., 2009) alter the TF-IDF metrics to find new key features to answer to the question: what is a region paying attention during a past event? They defined spatio-temporal subsets of the corpus according to time and space intervals. TF-IDF is then modified to enhance term that are often used during this spatio-temporal set and less used globally. At the end of this step, term scores are improved if they are not present in the previous time interval or in the geographical spaces. This emphasizes even more the discriminating terms of a space-time window. In (Bringay et al., 2011), the authors analyze tweets according to their multidimensional characteristics in a data-warehouse by using a Roll-up operator. Each specific context is defined at its associated spatial hierarchy level.

To the best of our knowledge, there is no approach exploiting both spatial and temporal hierarchical aspects. In this paper, we focus on the hierarchical characteristics of these dimensions to provide an adaptive measure for extracting the more relevant information as well as the specific indicators related to one or more hierarchies. Moreover, to review and analyse our results, we adopt a specific representation model called BERT (Devlin et al., 2019). Based on a very large training set (billion word corpus), this model assigns to each term a list of values (a vector) that reflects the meaning of a word with its context. This representation in vectors allows mathematical computations as semantic distance calculation between terms. Multiple models, based on different training set, are openly available like RoBERTa (Liu et al., 2019), COVID-Twitter-BERT (Müller et al., 2020), or DistilBert (Sanh et al., 2020).

3 Hierarchy-based measure for tweet analysis

In this section, we present the improvement of the *TF-IDF* measure that allows us to take advantage of the available hierarchies. It aim is to provide the end-user with facilities to focus on particular levels. Traditionally, the $TF\text{-}IDF$ measure gives greater weight to the specific words of a document (Salton et al., 1975).

Let $D = d_1, d_2, ..., d_n$ be a collection of documents and t a term in the collection, the term frequency-inverse document measure is defined as follows:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (1)$$

with TF and IDF defined as:

$$TF(t,d) = \frac{freq(t,d)}{|d|}$$
$$IDF(t,D) = log_2 \frac{|D|}{|\{d|t \in d\}|} \quad (2)$$

Our proposal is inspired of the $TF\text{-}IDF_{adaptive}$, which identifies the most significant words according to spatial level hierarchies of data cube in the data-warehouse context (Bringay et al., 2011). In this framework, we want to focus on the desired level of the hierarchy. The hierarchical IDF is thus defined for the elements sharing the same hierarchical path. Starting from the TF-IDF definition, we adapt the measure to the specific levels of the associated dimensions (time and space hierarchies).

Taking into account the specificities of the tweets, we adopt the following hypothesis: a term appears only once in a tweet (unlike academic papers or media news, words are rarely repeated inside a tweet), that means TF of terms are low . Thus a tweet cannot be the document unit in the TF-IDF definition. A document d is considered as a triple $(s_i, t_j, \{e_k, occ\})$ where s_i stands for a specific value on the space dimension of the i level of the associated hierarchy, t_j stands for a specific value on the time dimension of the j level of the associated hierarchy, $\{e_k, occ\}$ stands for the set of terms related to the tweets concerned by the value s_i and t_j with its occurrence number. The corpus D is the set of documents considered as the nodes at level i for the space dimension and level j on the time dimension.

Varying the space and time dimensions will modify both TF and IDF contrary to the adaptative measure defined in (Bringay et al., 2011) where only the IDF value definition changes. The Hierarchical *TF-IDF* denoted *H-TF-IDF* is thus defined as: $H-TF-IDF(t, d_{(s_i,t_j)}, D_{(level_i,t_j)}) =$

$$TF(t, d_{(s_i,t_j)}) \times IDF(t, D_{(level_i,t_j)}) \quad (3)$$

with TF and IDF defined as:

$$TF(t, d_{(s_i,t_j)}) = \frac{freq(t, d_{(s_i,t_j)})}{|d_{(s_i,t_j)}|}$$
$$IDF(t, D_{(level_i,t_j)}) = log_2 \frac{|D_{(level_i,t_j)}|}{|\{d'_{(s_i,t_j)}|t \in d'_{(s_i,t_j)}\}|} \quad (4)$$

with $d'_{(s_i,t_j)}$ having the same hierarchical path on the space dimension and the same time dimension value as $d_{(s_i,t_j)}$.

4 Experiments

4.1 Description of the corpus

To evaluate the H-TFIDF efficiency for extracting discriminant terms in spatial and temporal windows, the experiments are conducted on a corpus of tweets related to COVID-19. Several initiatives have been carried out since the beginning of the crisis like (Ofli, 2020), (Balech et al., 2020) and on the machine learning competitions platform Kaggle[1]. As we want to focus on the early beginning of the outbreak, we processed the dataset collected by E.Chen (Chen et al., 2020) which starts the earliest (i.e., January 22, 2020). During this period, common words related to the COVID-19 have been used worldwide (Alshaabi et al., 2021). In this massive flow of similar terminologies, we want to assess the spatio-temporal anchoring of the terms extracted by H-TFIDF. This corpus has been also used by other studies. (Ferrara, 2020) characterizes the impact of Twitter bots on COVID-19 conspiracies and (Jiang et al., 2020) analyses the political polarization in the United States. This corpus have been enhanced as well by (Lopez and Gallemore, 2020) to complete their collection with other keywords and apply state-of-the-art methods on sentiment analysis and named entity recognition.

For our study, we worked on a subset of this corpus. First, we removed retweets because they provide a resonance chamber and increase number of occurrences that can distort our statistical measurement. Then we only focused on the period between January 22 to February 29, 2020 with around 8,7 Millions of tweets, as countries were encountering their first COVID-19 cases. As described in the next section, Processing pipeline, we geocoded around 2,7 Millions of tweets world wild. Finally, we conducted our experiments on European countries (670 000 tweets) in English (270 000 tweets). Fig. 1 shows the distribution of tweets in European countries.

4.2 Processing pipeline

The processing pipeline, available in our repository[2] has four main steps: collect, indexing & geocoding, pre-process and computation of H-TFIDF as described in fig. 2. The first step consists to retrieve tweets, also called hydrate tweets, from E.Chen corpus of tweet IDs[3].

[1] https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april,
https://www.kaggle.com/gpreda/covid19-tweets

[2] https://gitlab.irstea.fr/remy.decoupes/covid19-tweets-mood-tetis

[3] https://github.com/echen102/COVID-19-TweetIDs/

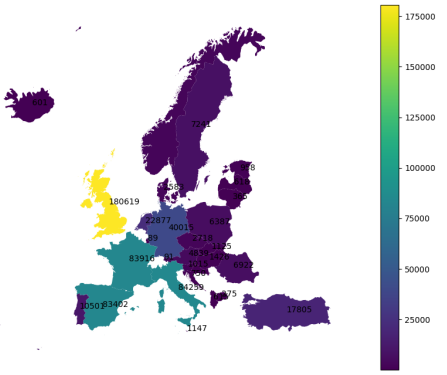


Figure 1. Distribution of geocoded tweets for European countries from 2020-01-22 to 2020-02-29 from E.Chen corpus

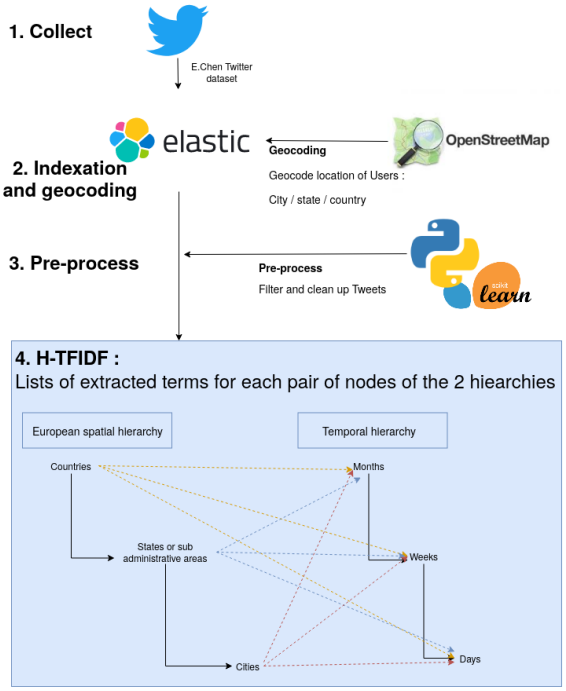


Figure 2. Description of the processing pipeline

Tweets metadata and content are then both indexed into an ElasticSearch[4] database in the second step. ElasticSearch shorten search queries duration which is useful to extract subcorpora, explore data or for pre- and post-processing (Barbaresi and Tinoco (2018)). To increase the number of georeferenced tweets from 2% to almost 20%, locations defined by users (i.e. user.location in tweet metadata) are geocoded (Dredze et al. (2013)), during the indexing, using OpenStreetMap data[5] and the open source geocoder photon hosted by Komoot[6]. Other studies decide to geocode tweets not only with the user location but also with the language used (Alshaabi et al. (2021)) or the content of tweets (Ofli (2020)). However infer location based on the language may lead to mistakes as English, Spanish, Arab or French could be spoken in different countries. Geocoding spatial information in tweet content could add noise, as well, for example in tweets talking about a football match between two cities which could spread the virus among supporters.

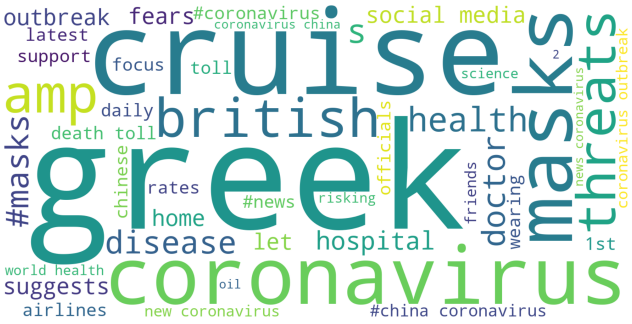
Filtering tweets based on their location and metadata is then processed in the third step. We filter in only original tweets (no re-tweets), in English, from European country.

The fourth task is dedicated to the computation of H-TFIDF. First, tweet contents are aggregated on the smallest spatial and time slot defined, i.e. by city and day. Token occurrences are then calculated using scikit-learn (Pedregosa et al., 2011) on a vocabulary of 25 000 unigram that have at least one character and eventually a "" or "#" to keep name accounts and keywords. Finally we compute multiple TF-IDF on the corpus with multiple document configurations, i.e., aggregated over spatial (city, state, country) and temporal (day, week, month) dimensions to retrieve H-TFIDF.

4.3 Results

H-TFIDF extracts insight terms for each level of the hierarchy, from country per month to city per day. Fig.3 illustrates H-TFIDF results for Greece during February. The size of the term font reflects the rank of the term. We can easily identify terms mainly used in Greece to then explore a specific topic. For example, the analysis of tweets containing the word "cruise" for the week 2020-02-02 reports two topics. The first one is the fear that the virus would spread through boats passengers. The second topic reports concerns about passengers forced to stay on board. Other example, tweets containing the word "Thessaloniki" extracted in week 2020-02-23 report two pieces of information of the local situation in Greece. Thessaloniki hospital is treated the first COVID-19 case in Greece and Twitter

[4]https://www.elastic.co
[5]https://www.openstreetmap.org/
[6]https://photon.komoot.io/



(a) week 2020-02-02

(b) week 2020-02-09

(c) week 2020-02-16

(d) week 2020-02-23

Figure 3. Wordcloud of H-TFIDF terms for Greece by week

users are reporting first panic buying of surgical masks and antiseptic.

4.4 Discussion

In order to evaluate and compare our measure with the state-of-the-art TF-IDF measure (equation 1), we carried out two kinds of experiments. The first one assesses H-TFIDF ability to extract discriminating terms, i.e. terms that can characterise a country / region / city for a period. We thus compute the overlapping between extracted terms from H-TFIDF, TF-IDF and the most frequent terms used by a country. For example, let us focus on United Kingdom (UK) during the last week of January. The table 1 shows the TOP15 terms for each measure. We can observe that H-TFIDF captures terms with a strong anchoring in UK as #brexitday or #coronavirusuk. Other terms like #missamericana or Ighalo refers to movie release or football events that were discussed a lot during this period in UK even if they were not related to the COVID-19 crisis. Finally, global used terms like "coronavirus" or "wuhan" are still present in H-TFIDF but in a lesser amount than TF-IDF. The fig. 4 is another visualisation of the TOP100 terms for each measure. It is a Venn diagram representation (Ho et al., 2021), the intersections of circles give the list of common terms between measures. The intersection between H-TFIDF and the most used terms in UK is greater than TF-IDF and the most used terms. This is confirmed for each country studied. Indeed the fig.5 shows percentage of common words between H-TFIDF or TF-IDF terms in one hand and the most frequent terms used by country in the other hand. The analysis of these overlap ratios reveals that H-TFIDF (in blue) succeed in extracting terms mainly used by local areas better than TF-IDF (in orange). However the overlap between H-TFIDF and the most frequent terms should not be 100% either because we except H-TFIDF to not extract terms also used in other countries.

H-TFIDF	TF-IDF	Frequent terms
#brexitday	#brexit	coronavirus
#coronavirusuk	#china	china
#missamericana	#corona	#coronavirus
#walls	#coronavirus	people
coronavirus	#coronavirusoutbreak	virus
wuhan	#coronavirustruth	corona
#coronarivus	#covid19	outbreak
model	#hongkong	just
virus	#outbreak	news
china	#usa	like
people	#wuhan	chinese
york	#wuhancoronavirus	cases
#coronavirus	#wuhanvirus	world
just	absolutely	death
ighalo	according	wuhan

Table 1. Term ranking - last week of January - UK

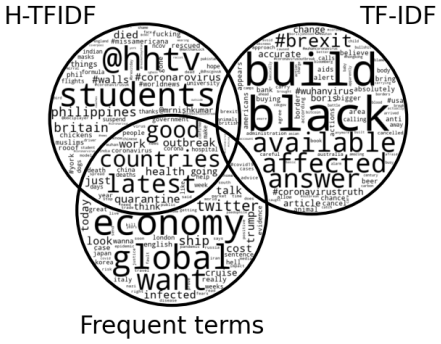
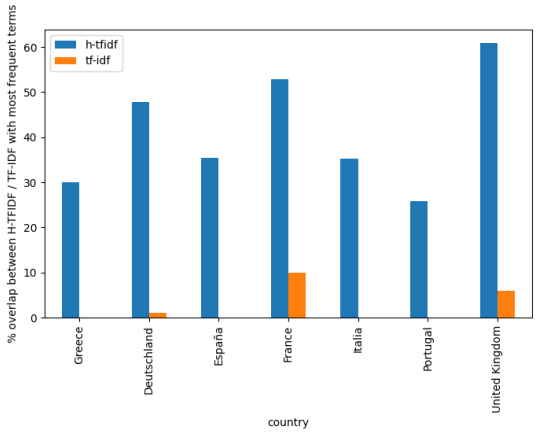


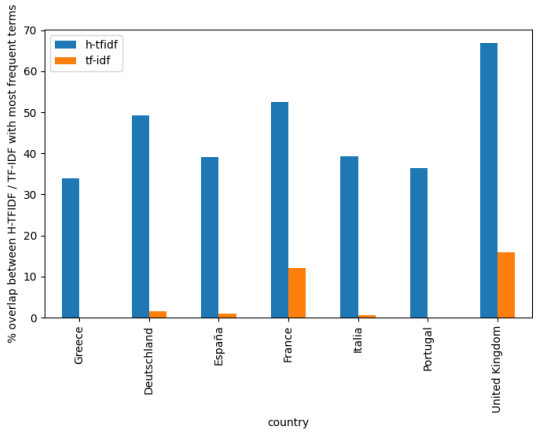
Figure 4. Common terms between H-TFIDF, TF-IDF and most frequent terms for United Kingdom in the last week of February

For the next evaluation, we highlight and compare semantic diversity of H-TFIDF selected terms compare to TF-IDF selected terms. We thus encode each term using the language model DistilBERT (Sanh et al., 2020) as introduced in the related work section. To visualize this comparison, we apply a dimension reduction of the DistilBert representation of our terms with t-SNE algorithm, i.e., we reduce the 766 dimensions to 2 and then visualize it in a scatter plot for H-TFIDF terms in fig.6a and for TF-IDF terms in fig. 6b. Both figures have the same scale. Each term is a dot and is defined by two coordinates in this semantic space. The axes therefore have no unit. The distance between two dots (or two terms) is anti-proportional to their semantic similarity. For instance "coronavirus" and "corona" will be closer than "coronavirusuk". For this study, we focus on the semantic extent of H-TFIDF compared to TF-IDF. We can observe that H-TFIDF term projections have a taller extent than TF-IDF. The analysis of term clusters (set of nearby terms) could be part of further works related to topic modelling. This reflects the greater semantic diversity of H-TFIDF terms. This diversity of terms is interesting as it could be used as a proxy to monitor a local situation and could help to understand what makes this local situation special compared with other places during a crisis or an emergency.

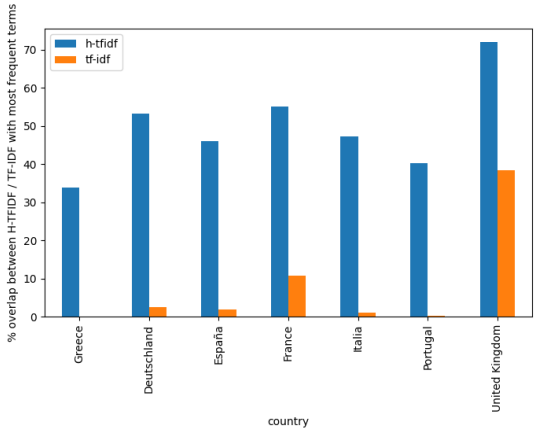
To sum-up, H-TFIDF has three advantages. These evaluations show that H-TFIDF captures terms with a strong anchoring for a region at a period. Moreover, these terms reflect local concerns since there are widely and specifically used in their spatio-temporal windows. Finally, we pointed out that these terms bring also a greater semantic richness.



(a) TOP 100 first terms

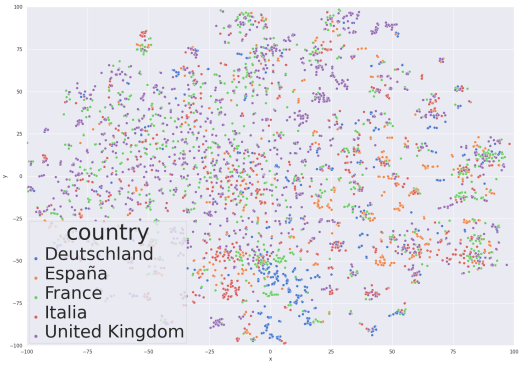


(b) TOP 200 first terms

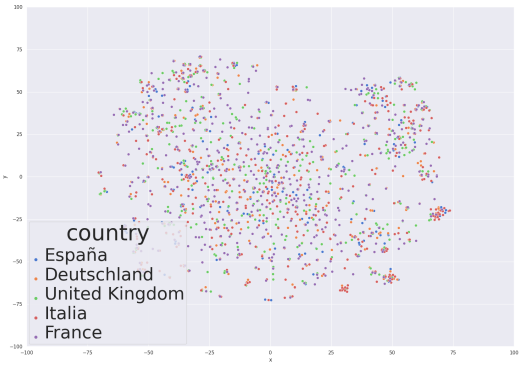


(c) TOP 500 first terms

Figure 5. Comparison of percentage of common words between H-TFIDF or TF-IDF with the most frequent terms per country



(a) H-TFIDF



(b) TF-IDF on subcorpora by country

Figure 6. Projection of H-TFIDF and TF-IDF DistilBert representation in a t-SNE space

4.5 Data and Software Availability

The whole workflow is available at https://gitlab.irstea.fr/remy.decoupes/covid19-tweets-mood-tetis. This repository provides a detailed set of instructions for downloading the data and executing the workflow, including data ingestion to ElasticSearch, computing H-TFIDF and creating figures as presented in this paper.

The workflow underlying this paper was partially reproduced using a data subset by an independent reviewer during the AGILE reproducibility review and a reproducibility report was published at https://doi.org/10.17605/osf.io/rdnyu. The data subset and the associated configuration of the workflow are available at https://doi.org/10.5281/zenodo.4742151.

5 Conclusion and Future Work

This paper proposes a new method for term extraction from tweet data based on spatio-temporal criteria. Browsing through spatial and temporal hierarchies (from country to city and month to day) provides insight of local concerns at different scales. This method, applied on COVID-19 related tweets, illustrates how public health agencies could have new sources of information about local situations.

Our future work will focus on the automatic exploitation of H-TFIDF results in order to monitor the evolution and the dynamic of crisis. To do so, we will enhance H-TFIDF thank to language models such as BERT or GPT-3. These unsupervised models will be used to transliterate the meaning or the semantic of the extracted terms. This semantic representation will be leveraged with specialised controlled vocabularies in disaster management and public health to drive the crisis monitoring.

In addition, an effort will also be made to take into account multilingualism. Indeed, in this current work, we decided to focus only on English tweets as the corpus was collected using only English keywords. This aspect will reduce the representativeness bias of Twitter for non English native speaker and will more effectively reflect the situation in European countries.

Acknowledgements. This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD012. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. This work was also partially supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004, #DigitAg.

References

Alsaedi, N., Burnap, P., and Rana, O.: Temporal TF-IDF: A High Performance Approach for Event Summarization in Twitter, in: 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 515–521, 2016.

Alshaabi, T., Arnold, M. V., Minot, J. R., Adams, J. L., Dewhurst, D. R., Reagan, A. J., Muhamad, R., Danforth, C. M., and Dodds, P. S.: How the world's collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter, PLOS ONE, 16, e0244 476, https://doi.org/10.1371/journal.pone.0244476, https://dx.plos.org/10.1371/journal.pone.0244476, 2021.

Balech, S., Benavent, C., and Calciu, M.: The First French COVID19 Lockdown Twitter Dataset, 2020.

Barbaresi, A. and Tinoco, A. R.: Using Elasticsearch for linguistic analysis of tweets in time and space, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), edited by Banski, P., Kupietz, M., Barbaresi, A., Biber, H., Breiteneder, E., Clematide, S., and Witt, A., European Language Resources Association (ELRA), Paris, France, 2018.

Bringay, S., Béchet, N., Bouillot, F., Poncelet, P., Roche, M., and Teisseire, M.: Towards an On-Line Analysis of Tweets Processing, in: Database and Expert Systems Applications - 22nd International Conference, DEXA 2011, Toulouse, France, August 29 - September 2, 2011, Proceedings, Part II, pp. 154–161, https://doi.org/10.1007/978-3-642-23091-2_15, 2011.

Chen, E., Lerman, K., and Ferrara, E.: Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set, JMIR Public Health and Surveillance, 6, e19 273, https://doi.org/10.2196/19273, http://publichealth.jmir.org/2020/2/e19273/, 2020.

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., and Larson, H.: The pandemic of social media panic travels faster than the COVID-19 outbreak, Journal of Travel Medicine, 27, taaa031, https://doi.org/10.1093/jtm/taaa031, https://academic.oup.com/jtm/article/doi/10.1093/jtm/taaa031/5775501, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

Dredze, M., Paul, M. J., Bergsma, S., and Tran, H.: Carmen: A Twitter Geolocation System with Applications to Public Health, in: AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI), 2013.

Erra, U., Senatore, S., Minnella, F., and Caggianese, G.: Approximate TF–IDF based on topic extraction from massive message stream using the GPU, Information Sciences, 292, 143 – 161, 2015.

Ferrara, E.: What types of COVID-19 conspiracies are populated by Twitter bots?, First Monday, https://doi.org/10.5210/fm.v25i6.10633, https://journals.uic.edu/ojs/index.php/fm/article/view/10633, 2020.

Ho, S. Y., Tan, S., Sze, C. C., Wong, L., and Goh, W. W. B.: What can Venn diagrams teach us about doing data science better?, International Journal of Data Science and Analytics, 11, 1–10, https://doi.org/10.1007/s41060-020-00230-4, http://link.springer.com/10.1007/s41060-020-00230-4, 2021.

Jiang, J., Chen, E., Yan, S., Lerman, K., and Ferrara, E.: Political polarization drives online conversations about COVID -19 in the United States, Human Behavior and Emerging Technologies, 2, 200–211, https://doi.org/10.1002/hbe2.202, https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.202, 2020.

Le Sceller, Q., Karbab, E. B., Debbabi, M., and Iqbal, F.: SONAR: Automatic Detection of Cyber Security Events over the Twitter Stream, in: Proceedings of the 12th International Conference on Availability, Reliability and Security, ARES '17, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3098954.3098992, https://doi.org/10.1145/3098954.3098992, 2017.

Li, C., Sun, A., and Datta, A.: Twevent: Segment-Based Event Detection from Tweets, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, p. 155–164, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2396761.2396785, https://doi.org/10.1145/2396761.2396785, 2012.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

Lopez, C. E. and Gallemore, C.: An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic, preprint, In Review, https://doi.org/10.21203/rs.3.rs-95721/v1, https://www.researchsquare.com/article/rs-95721/v1, 2020.

Müller, M., Salathé, M., and Kummervold, P. E.: COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter, arXiv:2005.07503 [cs], http://arxiv.org/abs/2005.07503, arXiv: 2005.07503, 2020.

Nagarajan, M., Gomadam, K., Sheth, A. P., Ranabahu, A., Mutharaju, R., and Jadhav, A.: Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences, in: Web Information Systems Engineering - WISE 2009, edited by Vossen, G., Long, D. D. E., and Yu, J. X., pp. 539–553, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Ofli, U. Q. M. I. F.: GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information, IEEE Dataport, https://doi.org/10.21227/et8d-w881, http://dx.doi.org/10.21227/et8d-w881, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

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

Sanh, V., Debut, L., Chaumond, J., and Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020.

Tsou, M.-H. and Leitner, M.: Visualization of social media: seeing a mirage or a message?, Cartography and Geographic Information Science, 40, 55–60, https://doi.org/10.1080/15230406.2013.776754, https://doi.org/10.1080/15230406.2013.776754, 2013.