CLUSTERING ANALYSIS ON NEWS FROM HEALTH OSINT DATA REGARDING CORONAVIRUS-COVID 19

Alexandru Daia[1], Stelian Stancu[2], Alexandru Vladoi[3] and Constantin Ionescu-Tîrgoviște[4]

[1) 2) 3)] *The Bucharest University of Economic Studies, Romania*

[4)] *"N.C. Paulescu" National Institute of Diabetes, Nutrition and Metabolic Diseases, Romania*

E-mail: alexandru130586@yandex.com; E-mail: stelian.stancu@csie.ase.ro;
E-mail: alexvladoi@gmail.com; E-mail: cit@paulescu.ro

Please cite this paper as:

Daia, A., Stancu, S., Vladoi, A. and Ionescu-Tîrgoviște, C., 2020. Clustering Analysis on News From Health OSINT Data Regarding Coronavirus-Covid 19. In: R. Pamfilie, V. Dinu, L. Tăchiciu, D. Pleșea, C. Vasiliu eds. *6th BASIQ International Conference on New Trends in Sustainable Business and Consumption.* Messina, Italy, 4-6 June 2020. Bucharest: ASE, pp. 669-674

Abstract

Our primarly goal was to detect clusters via gensim libraries in news data consisting of information regarding health and threats. We identified clusters for the periods corresponding: i) Jannuary 2006 until the end of 2019, as December 2019 is considered the first month in which information about CORONVIRUS COVID-19 was made public; ii) between the 1st of Jannuary 2019 and 31st December 2019; and iii) between the 31st of December 2019 and the 14th of April 2020. We conducted experiments using natural language on open source intelligence data from a provider specialized in business risk intelligence and cyberthreat awareness.

Keywords

clustering, health data, covid, news, machine learning, osint

JEL Classification

C55, G40

Introduction

As we progress in a constant increasing globalized world, for the field of Economics and behavioural science, news and the impact of news becomes more and more important. News and globalization, as Boyd-Barrett and Rantanen (1998) show, "are at the heart of modern capitalism". Some scholars state that modern globalization has "started before modernity, or modernity just transformed the nature of globalization". Technological breakthroughs, telecommunication and news are present in our daily lives. Regardless of when this process

has started or how it came about, today's world economies are to some extent connected to each other.

Open source data is an essential building block of financial and political affairs both foreign and domestic. We can therefore argue that news is nothing more than a commodity, and to be treated as such. This commodity keeps us informed, shapes our view on the world and helps us make choices and take decisions. It defines the way we respond to questions and which topics we perceive as being important. News can be represented by information we receive from various channels, it can be any number of platforms that we interact on and engage with on a daily basis.

The underlying assumption behind the proposed analysis, is that major global disruption events, skew the news flow towards certain topics, that we define as clusters. We analyse the news flow on the above-mentioned platforms to test our theory and measure the impact of the virus on information flow. News, the propensity and volume of news influences policy, economic behaviour and markets, and at the same time, economic behaviour, markets and events influence the news.

The rapid spread of the virus causes still significant public attention. A pandemic is the worldwide spread of a new disease, as defined by the world health organisation (WHO). COVID-19 was declared by the WHO a pandemic, meaning the virus has spread around the world, and most people do not have immunity against it. The global media and news databanks are flooding with new information on the virus. As governments around the world struggle to identify policies to combat the virus, the communication link to the general population is of extraordinary importance. We can conclude that the media has a significant role in the fighting and crisis managing for the public and government institutions alike. To note is that misleading or incorrect information can have negative effects on the transmission of the virus and its impact on the public. A focus on the transmission of information and activities around information share is necessary, in order for all stakeholders to understand the situation.

Review of the scientific literature

In the last years, significant research has been published using named entity recognition and topic modelling. Hu et. al. (2019) provide a short overview on the method. Topic models are useful tools for understanding large sets of information. Numerous variations of topic models exist, as we will show below, there are a number of applications and variations of models that can be applied to measure correlations between words into a topic model.

Vavliakis et. al. (2013) use the two techniques to investigate important online or real life events from large textual document streams, they propose that an event triggers discussion and comments on the internet and define an efficient methodology for performing event detection from large time-stamped web document streams.

Abinaya and Winster (2014) show that people use social networks, personal information and significant events that occur all over the world. News broadcasters share these posts, while the users discuss those events and post their reviews or repost. They assume that an event is temporal in nature and it changes over time. They use visual text analysis system to provide temporal views of changes that occur on the event. They propose an event identification system to account for the event occurrence and to identify the discussion flow arising from this event. They use topic clustering and named entity recognition to identify the significant events that are available on the web.

Allahyari et al. (2017) use text mining to sort through text data, created in from social networks, patient records, health care data etc. They argue that text data is a good example of unstructured information, which is easily processed and perceived by humans, but is harder for machines to understand.

Liu, Q. et al. (2020) used the same methodology to collect and analyse reports on the COVID-19 virus and how media in China has have delivered health information during the COVID-19 crisis. They showed that the Chinese mass media news lags behind when reporting the major developments in the COVID-19 progress. They show how the situation developed from a local, national problem, to a global issue. They suggest further research in the field of the global space but also exploring the impacts of mass media on the readers through sentiment analysis of news data and the influences of misinformation about COVID-19 delivered through the mass media.

Mendez et. al. (2020) studies the territorial and temporal patterns of EU cohesion policy media coverage. They look at the topic content and tone of news using topic modelling and sentiment analysis techniques for the period between 2010 and 2017 across three territorial levels. They find significant differences in the tone used across territorial levels, with national and transnational levels being more negative than the regional level. They conclude that national and transnational media place more emphasis on EU wide news, while local and subnational media focuses on substantive policy topics corresponding with EU policy objectives. They, also find that news develop significantly over time and reacted to external events, such as the euro and migration crises. They conclude that, the tone of cohesion policy news is positive overall suggesting that the media can, in principle, contribute to public support for the policy and the EU more generally.

Khan et al. (2019) use news extraction from Twitter data to extrapolate about real life events. They develop a technique for analyzing Twitter's raw content. After pre-processing of tweets data and pooling of terms they extract topics using available topic modeling algorithm without modifying its core machinery. Then, they count for each topic the estimated number of tweets per day and correlated top hashtags and construct a time series graph for topic analysis. They identify bursty news detection, topic popularity, and people's way to perceiving an event and the transition over time.

Muppala (2019) uses social media for personality identification of social media users. They identify based on personal "news" community interests and topics using LDA and KATE. Experimental results with Twitter and Facebook data demonstrate that the proposed model has achieved promising results.

Zou et al. (2019) uses LDA to organize topics into a hierarchy automatically. They analyze various real-world datasets and conclude that the hierarchical labels are ambiguous and conflicting in some levels. They introduce an altered topic model that aims to incorporate the hierarchical label information into the topic generation process. They use the model on BBC news and Yahoo! Answers datasets show a significant result in impact analysis on these sets using the adjusted topic modeling with more interpretable hierarchical structure.

We base our work and algorithm on Mikhail Salnikov (2018), who describes how clustering in text works. He explains how the same word in different strings can affect clustering. He presents TF-IDF methods with examples using Python. We use these algorithms for our experiment. We use his method to identify the type of clustering we are analysing as well as the most efficient method to use as explained in the chapter to follow.

Methods and experiments

We have collected data regarding news and threats in health domain from the platform "brica.de", a company specialized in threat analysis or cyber-threats in various domains such as health, military, governments, energy, critical infrastructure, geopolitics and other. Total number of news was pretty large consisting of around 20.000 records.

We divided data into three periods:

- January 2006 to December 2019, particularly the 31st of December 2019 as this was the first month information about the novel corona virus COVID-19 was released to the public.

• 1st January 2019 to 31st of December 2019, as a base for difference between 2019 and 2020.

• 31st December 2019 until the 14th of April 2020 (stop date of this paper), to reflect the impact of the virus on information flow.

We used in our experiments spacy. Spacy is a free open-source library for Natural Language Processing in Python programming language. Its functionalities include clustering and name entity recognition.

Findings

After running the experiment, we identified that the shift in news reported happened abruptly, while before COVID-19 the focus was around digitization of health-care, pharmaceutical drugs development and information (fig.nr.1), after the COVID-19 outbreak the news pivoted around organization and research, in hope of fighting the virus (fig.nr.2). This is a clear shift compared to the full year 2019 that focused on cyber attacks and health-care issues regarding security and protection (fig.nr.3). We identify 4 large topics for each period under review. The global security data provider Brica aggregated in real time the COVID-19 outbreak development. The major themes after December 2019 correctly reflect the focus points around the COVID-19 crisis and the global issue it represents. We recommend that future work should address the impact of COVID-19 on the financial market, and how different medias in different countries reacted to this data.

Fig. no. 1 Results for Topics from 2006 until Covid-19 (Dec'19)

Source: own research

It is interesting to see in the second clustering experiment for the year 2019, captures in topic 3 the keyword "virus". Note that at this point, the majority of people where not aware of virus problems. Also, "government", "prison" and "settlement", where "settlement'' has the biggest size meaning is reported for the period. The most significant importance in the topic indicates that news captured information about what was occurring or what is expected to occur in the near future.

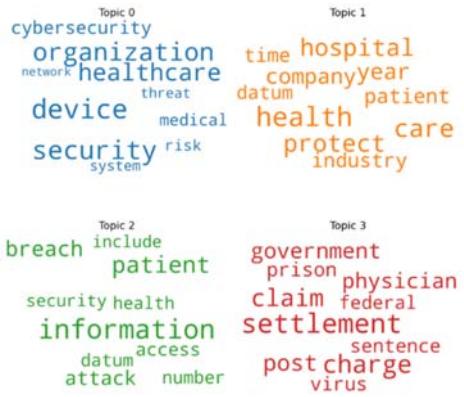

Fig. no. 2 Results for Topics in 2019
Source: own research

Clustering the osint information for the period 31st of December until the 14th of April 2020 (fig.nr.3), we notice that the first topic, Topic 0, is dominated by words like "drug'', "physician'' and "ambulance". Note the size of the words in the above figures reflect the importance of the word for the topic. It is interesting that these are followed by words such as cancer, patient and health, and hence, no mention of covid-19. We have manually inspected news and we have concluded that there could be two reasons: either patients with cancer face death more often when contaminated with covid-19. Other reason could be that effort now is centered in fighting covid-19 and cancer patients are not helped as before covid-19 pandemics, some reports showed that are even abandoned.

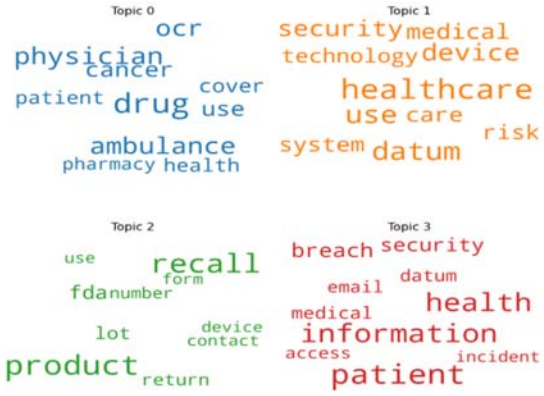

Fig. no. 3 Results for Topics after the COVID-19 outbreak
Source: own research

Conclusions

As we used a large amount of data, consisting of nearly 20.000 news, we can conclude that the clustering method on text data brings us a true big picture on what is happening in the health-care news field.

For the period 2006 to 31 December 2019 we can observe an important distribution in Topic number 0 towards keywords such as ,,prescription'' ,, letter'' ,, drug'', which hold the largest size, and therefore importance. One can argue that over the last 14 years, bureaucracy dominated the healthcare industry, as medical doctors were focused on topics such as

bureaucracy and administration. It is notable that cancer is also reflected for the entire period, as this was a major issue for the year at stake, considering the global fight against this disease. In 2019, we not the presence of topics surrounding “virus” reports, despite the fact that brica only recorded COVID-19 reports starting 31 December 2020. “Government” is also reflected as an important key word, as it shows the importance of the role that national institutions play in health care news.

It is important to note that clustering accurately reflected the news situations of the periods.

References

Abinaya, G., and Winster, S. G., 2014. Event identification in social media through latent dirichlet allocation and named entity recognition. In *Proceedings of IEEE International Conference on Computer Communication and Systems ICCCS14,* pp.142-146.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K., 2017. *A brief survey of text mining: Classification, clustering and extraction techniques*. [pdf] Available at: <https://arxiv.org/pdf/1707.02919.pdf> [Accessed at 19 February 2020].

Boyd-Barrett, O. and Rantanen, T. eds., 1998. *The globalization of news*. London; Thousand Oaks: Sage Publications.

Brica, n.d. *Database*, [online] Available at: <https://brica.de/> [Accessed at 13 April 2020].

Hu, Y., Boyd-Graber, J., Satinoff, B. and Smith, A. 2014. Interactive topic modeling. *Machine learning*, 95(3), pp.423-469.

Khan, M.H.U.R., Wakabayashi, K. and Fukuyama, S., 2019. Events Insights Extraction from Twitter Using LDA and Day-Hashtag Pooling. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications and Services,* pp.240-244.

Mendez, C., Mendez, F., Triga, V. and Carrascosa, J.M., 2020. EU Cohesion Policy under the Media Spotlight: Exploring Territorial and Temporal Patterns in News Coverage and Tone. *JCMS: Journal of Common Market Studies*, 13016.

Salnikov., 2018. *Text clustering with K-means and tf-idf*, [online] Available at: <https://medium.com/@MSalnikov/text-clustering-with-k-means-and-tf-idf-f099bcf95183> [Accessed 13 April 2020].

The Guardian, 2020. *How did coronavirus start and where did it come from? Was it really Wuhan's animal market*? [online] Available at: <https://www.theguardian.com/world/2020/apr/15/how-did-the-coronavirus-start-where-did-it-come-from-how-did-it-spread-humans-was-it-really-bats-pangolins-wuhan-animal-market> [Accessed at 9 February 2020].

Vavliakis, K.N., Symeonidis, A.L. and Mitkas, P.A., 2013. Event identification in web social media through named entity recognition and topic modeling. *Data and Knowledge Engineering*, 88, pp.1-24.

Zou, X., Zhu, Y., Feng, J., Lu, J. and Li, X., 2019. A Novel Hierarchical Topic Model for Horizontal Topic Expansion With Observed Label Information. *IEEE Access*, 7, 184242-184253.