Monitoring the Semantic Change of COVID-19-related Expressions Using Dynamic Word Embeddings

Bogdán Asztalos, and Péter Bányász

***Abstract*—In this paper, we investigate how the COVID-19 pandemic has affected the use of language in the online space through measuring the semantic changes of words during the time that includes the outbreak of the pandemic and the months of the lockdown. As a first step, we apply a recent word embedding technique on a time-labelled text corpus collected from social media which represents the semantic relation of words based on their likelihood of co-occurring next to each other. By analyzing different statistical features of the received dynamic embedding, we can identify and quantitatively describe periods where the semantic properties of a chosen word are undergoing significant changes. Since this depends on the context and the usage of these words by the users, we can infer their reaction to the COVID-19-related events and relevant news dated to these periods.**

***Index Terms*—semantic change, COVID-19, social media, word embedding.**

I. INTRODUCTION

COVID-19 had a significant impact on human civilization during the last few years, inducing abrupt changes in society, ranging from the level of personal relationships and the shift towards home office in work management to the significant new challenges with which national healthcare systems had to face worldwide. Presumably, these drastic effects have affected the way we think about many things including interpersonal communication, both on personal and social levels. Reducing the number of personal contacts, online meetings becoming widespread, and several fake news popping up each week have made us change the way we approach new information and express our thoughts. Therefore, the tools of the education system, the government communication, the participants of mass media, and all the content-sharing agents also needed to adapt to these new circumstances, because – independently of their purposes – their messages will reach their goal only if they use new ways [1].

One of the most relevant examples in the topic of COVID-19 is the effectiveness of the epidemiological response of governments, which was primarily influenced by the condition of society, the degree of trust people had in their democratic institutions, the scientific community, and many other factors. The lower confidence people have in these factors, the higher the percentage of people who will refuse to accept epidemiological restrictions or vaccination in the future. However, the passage of time has led to a weakened acceptance of the government's actions. This resulted from the exhaustion of the lockdowns and the global disinformation campaigns that have been spreading on social media for years. The unclear origins of the COVID-19 virus[1] and the often-contradictory messages from different governments have provided an excellent opportunity to spread conspiracy theories and fake news. The scientific community also has a significant role in the defence against the disease, but the infodemic state related to COVID-19 has significantly eroded individuals' trust in the state of science and has reinforced pseudoscientific theories [2].

This tendency has been increased by the nature of the algorithms on social networking sites, which has led to a combination of the most absurd conspiracy theorists, from flat-earthers to anti-vaccinationists who believe that the 5G network causes the coronavirus and that microchips are being implanted in people by vaccination to track them by various actors of the deep state behind the scenes [3]. This problem is even more serious considering that the spread of fake news regarding the coronavirus and the rejection of vaccination will significantly compromise the prevention of epidemics in the future and can result more and more epidemic outbreaks globally [4]. Research has shown that social media users are more likely to believe various COVID-19-related conspiracy theories that reinforce anti-vaccination, and anti-masking, among others [5]. Well illustrates the process that in January 2020, for the first time, a post appeared on social media claiming that the spread of 5G technology was responsible for the spread of the coronavirus, which also led to a rise in anti-China feelings [6]. Then in April 2020, the 5G hoax led to 77 incidents of vandalism of 5G transmitting stations in the UK by individuals afraid of the coronavirus to prevent the further spread of the virus [7].

Therefore, it is important to counterbalance the harmful information flow in the public consciousness, and as COVID-

The research was conducted within the framework of the Network Science Research Group at Ludovika – University of Public Service. The research was supported by the EKÖP-24-3-II-ELTE-213 University Excellence Scholarship Program and the EKÖP-24-4-II-23 University Research Scholarship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

Bogdán Asztalos is with Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary, and with Health Services Management Training Centre, Semmelweis University, Budapest, Hungary (e-mail: abogdan@caesar.elte.hu)

Péter Bányász is with Department of Cybersecurity, University of Public Service Ludovika, Budapest, Hungary (e-mail: banyasz.peter@uni-nke.hu)

DOI: 10.36244/ICJ.2025.2.3

[1]In our study, we cannot and do not take a position on the origin of the coronavirus. We believe that whether it evolved naturally or was released from a laboratory is unimportant in terms of its consequences. The crucial question about the pandemic is how we can 'defeat' it.

19 itself can only be controlled through an inter- and multi-disciplinary approach, this communication-focused problem is also should examined from several directions. Such interdisciplinary cooperation is not unprecedented, similar has happened during the last 25 years, when network-based tools gradually gained relevance in exploring the mechanisms and principles of complex systems [8], including both biological [9] and social systems [10], and network thinking became accepted both by academia [11], industry [12], and governments [13]. The spreading use of artificial intelligence (AI) tools that we see today in all areas of life also highlights that computational techniques are quite helpful and over time they will be indispensable in fully understanding both social, scientific, and governmental problems [14]–[16].

The increasing prevalence of fake news and counterfeit profiles on social media platforms poses a substantial threat to informational integrity and societal trust. Whereas conventional approaches, such as rule-based solutions, metadata analysis, or human fact-checking, have assisted in mitigating the spread of disinformation, they often prove inadequate against the more sophisticated and rapidly disseminating false content [17]. Recent studies indicate that natural language processing (NLP) techniques are more effective in the identification of disinformation than traditional methods relying on fact-checking and source evaluation as they facilitate the representation and analysis of textual data. This enhances the accuracy of detection and fosters a more nuanced understanding of the semantic meanings embedded within the conveyed information [18], [19]. Such methodologies enable meticulous detection and elimination of propaganda while preserving document-level coherence, grammatical integrity, and the content's authenticity in news articles [20]. It is also found that the integration of NLP with machine learning (ML) algorithms can yield effective results in identifying false news and profiles [17], [21], [22].

Among the tools of NLP, word embedding methodologies have turned out one of the most instrumental in the identification of disinformation disseminated on social media platforms as they facilitate a deeper comprehension of the context and semantics of text [23], [24]. Word embedding algorithms like GloVe, TF-IDF, Word2Vec, and FastText-based embeddings provide sophisticated text analysis and classification techniques [25], and have proven effective in detecting disinformation within social media contexts [21], [26]. In the domain of deep learning integration, neural network models leverage embeddings to effectively capture contextual and sequential information inherent in textual data, thus markedly enhancing the accuracy of fake news classification, and understanding of context and the structural elements of disinformation [27], [28].

Besides, word embeddings provide a quantitative method to define semantic space and measure semantic distance i.e. the degree of difference between meanings of different words. By using temporal data, word embedding is also capable of measuring the semantic changes of words and gives us an opportunity to observe what phenomena have happened in human language [29], [30]. As we show in this paper, this could be used for finding out how COVID-19 has affected the way people use language, and so indirectly identifying social events and processes that can be intervened with the right communication and decision-making.

Our study aims to engage in the global scientific discourse about the problems that the pandemic created with our humble tools, proposing a new methodology that we believe can be adopted by different disciplines.

In this paper, firstly we briefly overview the necessary notions of distributional semantics and the Word2vec word embedding method, and then we introduce our results about applying diachronic word embedding on textual data from online social data.

II. Distributional semantics and word embedding

Distributional semantics (DS) is a subfield of computational linguistics that quantifies semantic similarities and differences between words and linguistic terms [31]. Since the scope of our study is to conduct a quantitative investigation about how much the meaning of words has changed during the COVID-19 pandemic, we relied heavily on the concepts of DS. The main tool we used is word embedding which represents words mathematically in a high dimension space and hence allows us to consider them as objects in a real geometric space. However, to understand word embedding and the logic behind constructing it, one should be aware of the distributional hypothesis and some of its consequences.

A. Distributional hypothesis and the role of co-occurrence statistics

It is a well-known fact that knowing the dictionary definition of a word and using it in a real-life sentence are not equivalent. The former is not even required for the latter, however, the way people use words highly depends on their lexical meaning, so it is possible to infer the semantic information from the linguistic context [32], [33]. This principle led Harris to the following statement: "a word is characterized by the company it keeps", which later became known as the distributional hypothesis and now is the basis of DS [34], [35].

One of the consequences of the distributional hypothesis is the fact that the statistical properties of linguistic context encode the meaning of words. Hence, a computational analysis of word occurrences in a large text corpus will let us distinguish different word meanings from each other and recognize similarities and differences between them [36]. For example, the word *dog* probably has more common neighbouring words with *cat* than with *house*, because there are more sentences where *dog* and *cat* are interchangeable than *dog* and *house* are. This simply follows from the fact that the meaning of the word *dog* is closer to the meaning of *cat* than to the meaning of *house*.

Hence, to express the semantic content of a word computationally, one needs to collect the words occurring in its context and analyse their statistical distribution. To store such statistics, the simplest solution is to count the co-occurrences of word pairs, so to gain the statistics of one individual word, only the concerning co-occurrence data values need to be accounted for. Therefore, studies in computational linguistics

focus on the co-occurrence statistics of text corpora and languages [37], [38].

Collecting co-occurrence statistics allows us to perform many computational analyses. One of the most apparent techniques is to build co-occurrence networks by linking words occurring in each other's context. Such a network should store all the valuable relational information about linguistic patterns, and by examining it explicitly with the tools of network study, it can reveal the fundamental semantic relationships [39], [40]. Observing the evolution of these networks can unfold the language evolution which is responsible for the network structure behind the scenes. In our study, we have focused on word embedding, which is obtained by a more complex derivation from co-occurrence statistics, but the above close relationship between co-occurrence networks and embeddings can be used both in the fields of computational linguistics and network science [41].

B. Word embedding and Word2vec

Word embedding is a technique that represents the above abstract concept of "co-occurrence statistics" and makes the semantic meaning interpretable for computers. The idea is to assign a set of real numbers (i.e., a multidimensional vector) to words in such a way that the algebraic properties of vectors reflect the same relationships between them as those that exist between the words. For instance, if two words have similar meanings, their vector representation should be close to each other in the multidimensional space, called the embedding space. Therefore, word embedding simply transforms semantic differences into geometric distances [42].

Constructing this vector representation can be done in many ways. The simplest methods, like LSA, take the co-occurrence value of some pre-selected context words and apply some kinds of dimension reduction on them [43], but in the last decade, more sophisticated word embedding techniques were introduced using unsupervised ML algorithms [44].

During our investigation, we used Word2vec's Skip-gram model with negative sampling (SGNS), which is one of the most widely used word embedding methods in recent years. It also uses unsupervised learning: the objective of the model is to estimate a co-occurrence probability using a two-layer neural network and the vector representation of the words is a kind of by-product of the optimization process. The cause of its popularity is that it can process a large amount of data in a short time because of its sampling procedure and also the fact that the constructed embedding returns psycho-linguistical relations more efficiently than other alternatives [45], [46]. The output of such an embedding is visualised in Fig. 1.

III. METHODS

A. Gathering linguistic data with SentiOne

To receive data for the linguistic study, we used an AI-powered social listening tool, called SentiOne with which we searched for COVID-19-related words and expressions, and downloaded posts and web pages containing them. SentiOne is a content-based web analytics platform that is dedicated to crawling and analysing content perceived online channels, like social media platforms, news portals, blogs, etc. [47]. It covers and recognizes 70 languages across the globe. The platform monitoring tool currently monitors over 20 000 000 000 mentions and gathers data from 8 different types of sources, namely portals, blogs, Twitter, Facebook, Instagram, video, forums, and review sites.

The mentions are divided into statements and articles, statements being automatically classified as either positive, neutral, or negative with the use of SentiOne's unique, proprietary algorithm. The platform's sentiment analysis is based on research work by John R. Crawford and Julie D. Henry [48]. They analysed the Positive and Negative Affect Schedule (PANAS). Based on their research, SentiOne's developers created algorithms that help determine the author's emotional attitude to the discussed topic. The platform uses proprietary artificial intelligence algorithms to classify the posts' overall sentiment.

The interactive platform is built upon user-provided keywords and key phrases to look for the specific mentions that, either in themselves or within their context, contain those pre-given phrases that interest the user. The system gathers data in almost real-time yet has a memory that can go back up to 3 years. For quantitative research, data is structured by different focus points and research parameters and is visualized interactively. This technology also supports qualitative research, enabling in-depth analysis and categorizing all the indexed web content.

We used three groups of keywords and key expressions to capture social data: some common words (mostly names of professions) that are not related to COVID-19 in an obvious way and can serve as a stable meaning benchmark; the names of the different vaccine types; and an expression aimed at the 5G-related fake news. The used keywords and expressions are listed in Table I. We have searched for each keyword in 30 different languages[2]. In this research, these queries were run because they are language-independent and have the same meaning in each of the given languages.

B. Diachronic word embedding

To measure how much words change their meaning over time, we used a series of word embedding. The steps of the process with which we received these embeddings is illustrated in Fig. 2.

Since the input of an embedding method is the co-occurrence data as explained in Section II-A, it is possible to construct distinct embeddings for each month starting from the collected statistics corresponding to the individual months. The result of a single embedding is a set of word vectors (i.e. points in a high-dimensional space) representing the meaning relations of the words, so if we get different sets for different months, we can compare them, hence inferring what changed in the language during a month. We assume that

[2]Currently, the following languages are available in fully supported form: Polish, German (+ Swiss German, Austrian German), Russian, Ukrainian, English (+ UK, US, Ireland), Dutch (+ Belgian Dutch), French (+ Belgian French, Swiss-French), Slovenian, Slovak, Hungarian, Romanian, Bulgarian, Serbian, Croatian, Bosnian, Montenegrin, Czech, Danish, Finnish, Swedish, Norwegian, Latvian, Lithuanian, Italian, Spanish, Portuguese, and Greek.

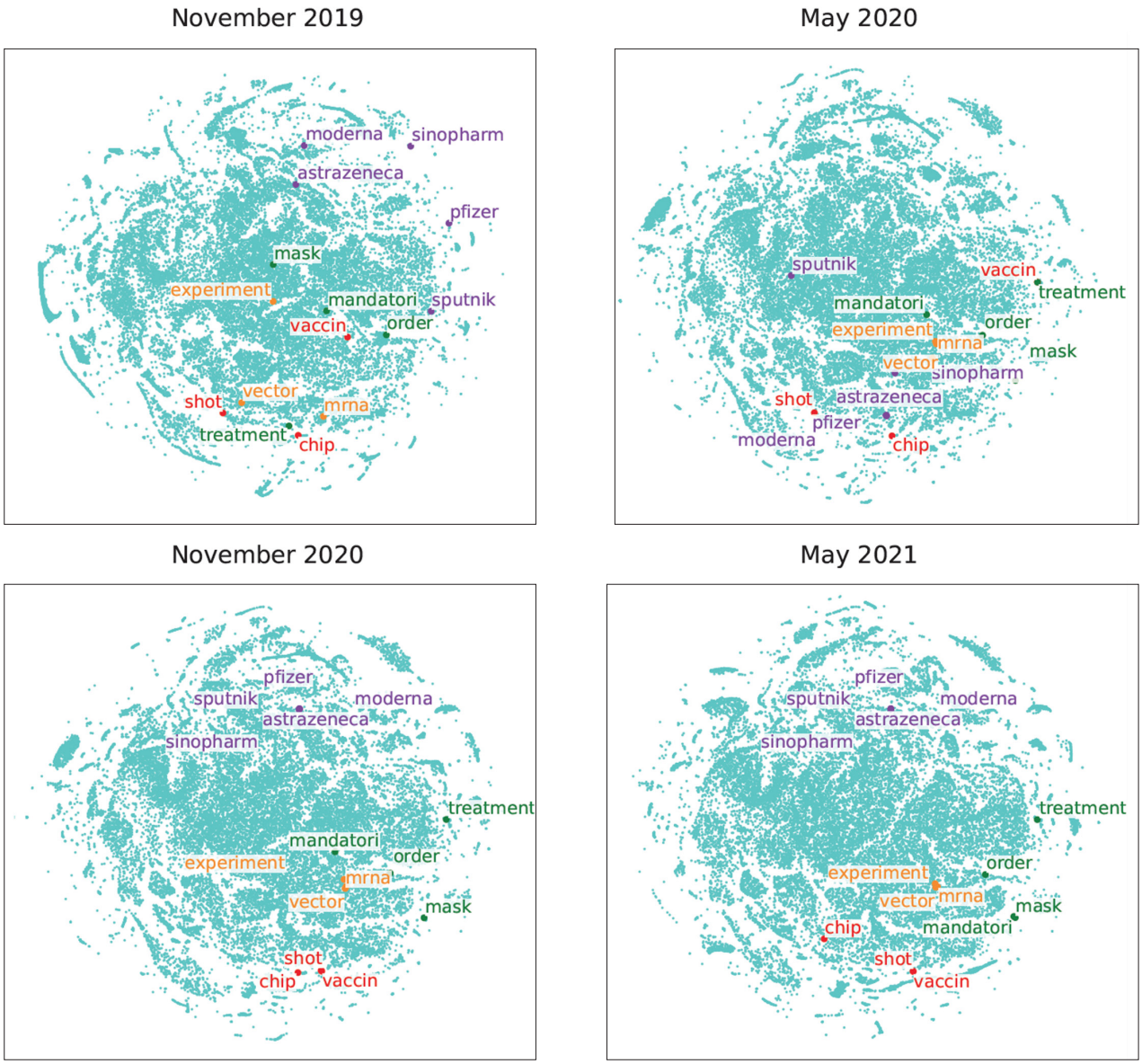

Fig. 1. Visualisation of diachronic word embedding. The coordinates of words can be considered as their position in a snapshot of the semantic space, where words are continuously walking. We plotted four different state during the examined period, where the meaning changes of the highlighted words can be observed. The geometric relation of these pandemic-related terms reflects relatively well the social attitude changes towards these topics.

most words do not change their meaning in a short time, like a month, so comparing consecutive embeddings can reflect the changes that happened so that the concerned words changed their position relative to the others. The obtained results also supported this assumption. This allows us to consider the words in the embedding space as walking particles, and to look at the semantic change in a language as the collective motion of these particles. Also, as previous research has shown, this collective motion of words can be studied quantitatively, and their behaviour can be described as an anomalous diffusion with a significant subdiffusive character which might be derived from the governing non-linear microscopic dynamics of human language evolution [49], [50].

To build such a series of embedding, we have run through all the textual sources in the downloaded raw data and performed the data processing methodology published by Asztalos et al. [49]. The used embedding method was Word2vec's Skip-gram model with negative sampling (SGNS) [45], [46] into 300 dimensions.

Since the dataset contained time-labels for each post about when they were published, we could distinguish different co-occurrence statistics coming from different periods. We applied monthly separation, i.e., handled independently the statistics from the 24 months between July 2019 and June 2021. The received series of embedding can be interpreted as snapshots of how word meanings migrate over the overall cloud of words. Four of these snapshots are illustrated in 2D in Fig. 1.

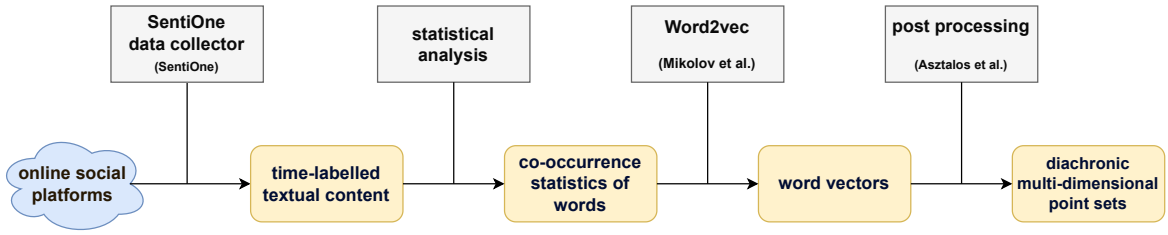

Fig. 2. The steps of our data process flow.

TABLE I
LIST OF KEYWORDS AND EXPRESSIONS WE USED FOR SEARCHING

Common words	Vaccine names	Fake news
"Bread"	"AstraZeneca"	"5G" and "COVID"
"Driver"	"J&J"	
"Engineer"	"Johnson and Johnson"	
"Film director"	"Moderna"	
"Illness"	"Pfizer"	
"Journalist"	"Sinopharm"	
"Lawyer"	"Sinovac"	
"Miner"	"Sputnik V"	
"Principal"		
"Soldier"		

IV. RESULTS

While we have analyzed the spread of fake news based on several keywords, we decided to focus on the trend of spreading COVID-19 and 5G-related fake news, which led to mass destruction in the UK, as it was mentioned in the Introduction. For the search terms "COVID" and "5G", we identified 1,156,535 shares during the period under study, resulting in approximately 4.4 billion views

As can be seen in panel (a) in Fig. 3, the significant majority of shares are found to be neutral, which can be attributed to three factors:

- AAI examines the shared content based on context, evaluating the expressions it contains in terms of positive and negative aspects, and then deciding which sentiment to attach to it based on the results. However, if the number of positive and negative expressions is approximately equal, the algorithm labels it as neutral.
- In many cases, users simply reshare the content without any explanation or comment, so they cannot analyse the new context.
- A more recent emerging new phenomenon is the intentional use of terms and phrases in the content shared for the express purpose of disinformation, overloading or poisoning the meaning of words, and deliberately creating a different emotional context. For example, the hashtag "#Bidenworstpresidentever" might be used as a positive hashtag in trending posts, even though the hashtag's meaning clearly indicates a negative emotional connotation.

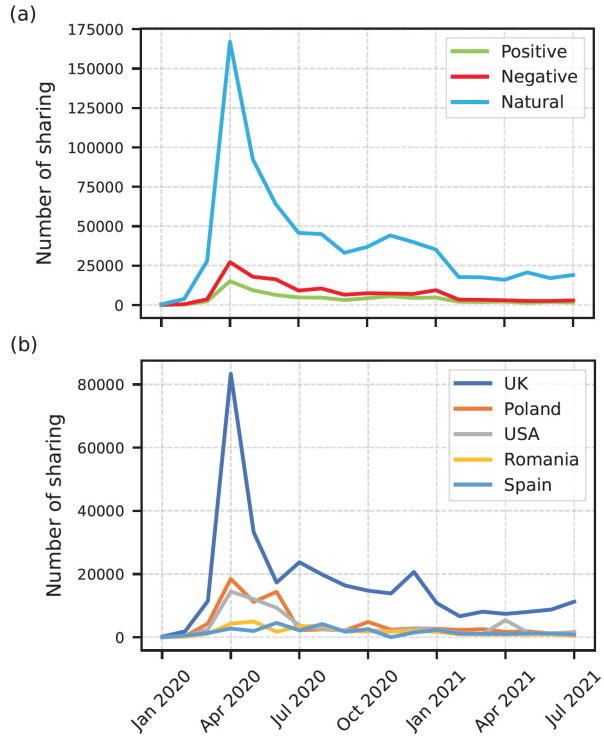

Fig. 3. Number of shares over time found by the keyword "COVID" and "5G". In panel (a) they are grouped by sentiment, while in panel (b) they are grouped by the country.

Fig. 3 shows that the number of shares related to the terms "COVID" and "5G" significantly increased from March 2020, reaching a record 230,000 shares in April, which coincides with the previously cited destruction of 77 5G towers. If we compare this with the territorial spread of the sharing (see Fig. 3, panel (b)), it is not surprising that the dominant location of the sharing is the UK: overall, there are 317,313 shares for the two search terms above, which, narrowed down to April 2020, means more than 80,000 shares.

Concrete examples of how we can visualize linguistic changes with the help of word embeddings can be seen in Fig. 1. Word embedding methods (like Word2vec) place words in a high-dimensional space so that words with similar meanings are closer to each other. Typically, the number of dimensions is several hundred (in our case 300), so the procedure result is relatively abstract. Still, with dimension reduction tools, it is possible to illustrate the structure of the "word cloud" in 2D. Using the t-SNE visualization [51], we plotted four

reaching a record 230,000 shares in April, which coincides with the previously cited destruction of 77 5G towers. If we compare this with the territorial spread of the sharing (see Fig. 3, panel (b)), it is not surprising that the dominant location of the sharing is the UK: overall, there are 317,313 shares for the two search terms above, which, narrowed down to April 2020, means more than 80,000 shares.

Concrete examples of how we can visualize linguistic changes with the help of word embeddings can be seen in Fig. 1. Word embedding methods (like Word2vec) place words in a high-dimensional space so that words with similar meanings are closer to each other. Typically, the number of dimensions is several hundred (in our case 300), so the procedure result is relatively abstract. Still, with dimension reduction tools, it is possible to illustrate the structure of the "word cloud" in 2D. Using the t-SNE visualization [51], we plotted four states of the word cloud in Fig. 1. I.e., these figures are approximations of the high-dimensional objects representing semantical meanings. It can be seen that the overall structure of the cloud (i.e., general meaning relations in the language) is the same (supporting our basic assumption), but some pandemic-related words get closer to each other while some get further. Thus, it can be observed, for instance, how the different pharmaceutical companies manufacturing vaccines grouped together, or how the word *chip* was firstly associated with vaccines and moved away later.

To perform some more descriptive analysis, we collected the 30 words closest to some chosen test words each month and plotted the monthly changes in this set in Fig. 4. Suppose that the spatial distance represents the meaning difference between words. In that case, we can assume that a word changes its closest neighbours because its meaning changes over time, so the number of changed neighbours correlates with its meaning change, hence can be viewed as a "rate of meaning change". In Fig. 4 this value is plotted over time for four everyday words (panel (a)) and for four vaccine names (panel (b)). One can see those words with stable, conventional meanings (e.g., word, friend) have a lower change rate of around 10-15; this can be considered the static value. In contrast, words like mask or president are more agile in the scanned period, probably because of the social events that happened that time. Vaccine words (more precisely the names of the companies producing them) had no fixed neighbour set in the beginning, but they decayed to the static value when the collective social opinion about them reached its final state so that we can identify the acceptance order of different vaccines, consistent with real-life experience.

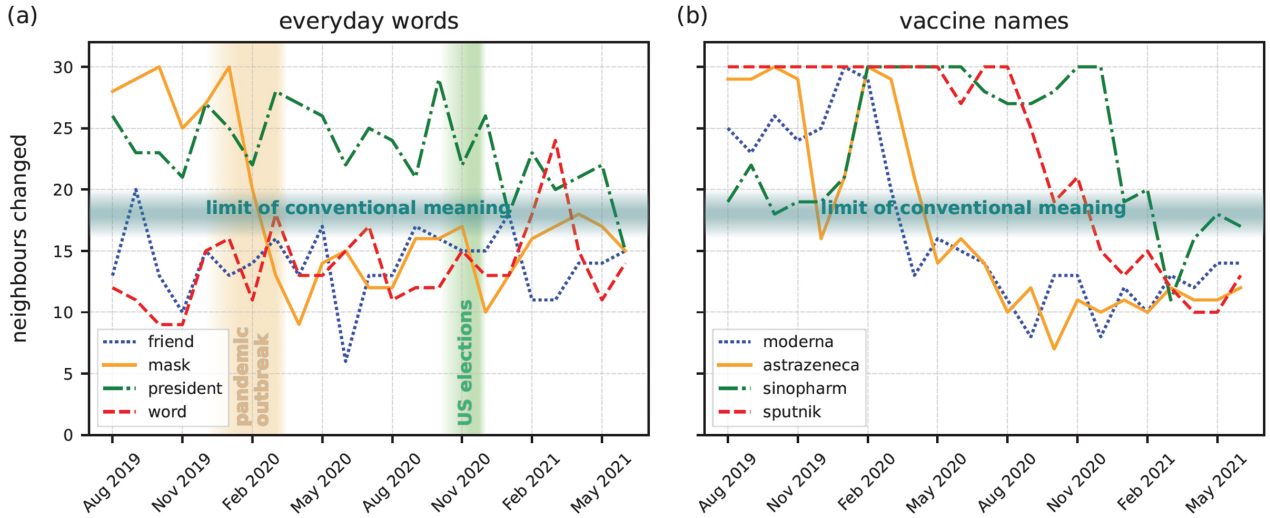

Fig. 4. The number of neighbours (out of 30) of a certain word which change from one month to another. Panel (a) shows four everyday words while panel (b) shows four vaccine words. Meaning of words with lower value than 16-20 can be viewed as static. On panel (a), the effect of social events like the outbreak of COVID or the US election campaign can be identified, while on panel (b), the gradual acceptance of vaccines can be observed.

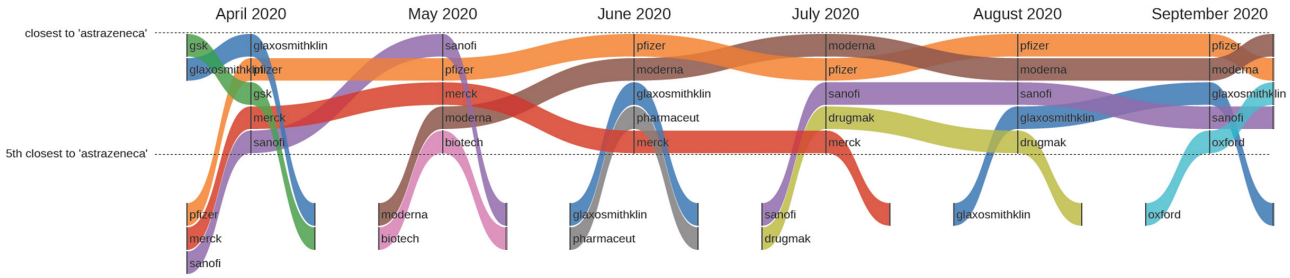

Fig. 5. The five closest words to AstraZeneca by time. The upmost stripe represents the closest word, while the lowest represents the 5th closest. The social change in what terms is associated with AstraZeneca over time is well observable.

To study in more detail what happens in the studied period, we can look for the closest few words to a certain word each month. Fig. 5 presents the case of AstraZeneca for its 5 closest neighbours. This Sankey figure shows the change of these words and their order over time. The following change in social attitude over time can observed: Pfizer and Moderna became really close when they started to develop the COVID-19 vaccine, but Merck and Glaxo Smith Kline (which are also pharmaceutical companies but did not have their own vaccine) left the immediate neighborhood. This also underlines the fact that public opinion about AstraZeneca (especially on social media platforms) was influenced primarily by the vaccine-manufacturing activity of the company in the examined period.

V. CONCLUSION

The results of this study hold significant implications for both scientific inquiry and practical strategies aimed at combating disinformation, particularly in relation to the evolving nature of fake news and linguistic shifts observed during the COVID-19 pandemic. The implementation of diachronic word embedding technologies has established a novel methodological framework for tracing the temporal development of linguistic meanings, thereby enabling an analysis of the transformation of social narratives and emotional patterns.

A key innovative aspect of the methodology is the application of linguistic representations within an interdisciplinary framework, facilitating not only textual analyses but also a comprehensive examination of the dynamic phenomena associated with social change. The findings can inform the formulation of targeted strategies to combat disinformation and enhance the efficacy of political and public health communications. Investigating the evolving context surrounding words linked to fake news will illuminate particularly virulent topics and phrases circulating on social media, thus allowing for more effective deployment of early intervention measures.

Furthermore, the visualization of linguistic embeddings facilitates both qualitative and quantitative assessments of shifts in social narratives. This dual methodological framework strengthens interdisciplinary research by integrating insights from network theory, NLP, and emotional analysis. The incorporation of new techniques and methodologies is poised to address pressing global challenges, such as epidemics and political conflicts, at both theoretical and practical levels.

ACKNOWLEDGMENT

The authors thank Péter Pollner and Gergely Palla for useful discussions at various stages of the project, and for the help with many technical details.

REFERENCES

[1] T. Bat-Erdene, N. Zayed Yazan, X. Qiu, I. Shakoor, A. Mekni, P. A. Kara, L. Bokor, and A. Simon, "On the quality of experience of content sharing in online education and online meetings," *Infocommunications Journal*, vol. 14, no. 2, pp. 73–84, 2022. **DOI**: 10.36244/ICJ.2022.2.8

[2] J. Hua and R. Shaw, "Corona virus (covid-19) "infodemic" and emerging issues through a data lens: The case of china," *International journal of environmental research and public health*, vol. 17, no. 7, p. 2309, 2020. **DOI**: 10.3390/ijerph17072309

[3] D. Allington, B. Duffy, S. Wessely, N. Dhavan, and J. Rubin, "Healthprotective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency," *Psychological medicine*, vol. 51, no. 10, pp. 1763–1769, 2021. **DOI**: 10.1017/S003329172000224X

[4] I. Ullah, K. S. Khan, M. J. Tahir, A. Ahmed, and H. Harapan, "Myths and conspiracy theories on vaccines and covid-19: Potential effect on global vaccine refusals," *Vacunas*, vol. 22, no. 2, pp. 93–97, 2021. **DOI**: 10.1016/j.vacun.2021.01.001

[5] D. Romer and K. H. Jamieson, "Patterns of media use, strength of belief in covid-19 conspiracy theories, and the prevention of covid-19 from march to july 2020 in the united states: survey study," *Journal of medical Internet research*, vol. 23, no. 4, p. e25215, 2021. **DOI**: 10.2196/25215

[6] M. Rich, "As coronavirus spreads, so does anti-chinese sentiment," *The New York Times*, Jan 2020, available at: https://www.nytimes.com/2020/01/30/world/asia/coronavirus-chinese-racism.html, accessed: 12 Jun 2023.

[7] C. Reichert, "5g coronavirus conspiracy theory leads to 77 mobile towers burned in uk, report says," https://www.cnet.com/health/5g-coronavirus-conspiracy-theory-sees-77-mobile-towers-burned-report-says/, 2020, accessed: 12 Jun 2023.

[8] M. Mitchell, "Complex systems: Network thinking," *Artificial intelligence*, vol. 170, no. 18, pp. 1194–1212, 2006. **DOI**: 10.1016/j.artint.2006.10.002

[9] E. Alm and A. P. Arkin, "Biological networks," *Current opinion in structural biology*, vol. 13, no. 2, pp. 193–202, 2003. **DOI**: 10.1016/S0959-440X(03)00031-9

[10] J. Scott, *Social networks: Critical concepts in sociology*. Taylor & Francis, 2002, vol. 4.

[11] D. J.Watts, "The "new" science of networks," *Annu. Rev. Sociol.*, vol. 30, no. 1, pp. 243–270, 2004. **DOI**: 10.1146/annurev.soc.30.020404.104342

[12] B. Axelsson and G. Easton, *Industrial networks (routledge revivals): A new view of reality*. Routledge, 1992.

[13] S. J. Ball and C. Junemann, Networks, new governance and education. *Policy Press Bristol*, 2012.

[14] S. Makridakis, "The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms," *Futures*, vol. 90, pp. 46–60, 2017. **DOI**: 10.1016/j.futures.2017.03.006

[15] M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam et al., "On scientific understanding with artificial intelligence," *Nature Reviews Physics*, vol. 4, no. 12, pp. 761–769, 2022. **DOI**: 10.1038/s42254-022-00518-3

[16] V. J. Straub, D. Morgan, J. Bright, and H. Margetts, "Artificial intelligence in government: Concepts, standards, and a unified framework," *Government Information Quarterly*, vol. 40, no. 4, p. 101 881, 2023. **DOI**: 10.1016/j.giq.2023.101881

[17] E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur, "A survey on the use of large language models (llms) in fake news," *Future Internet*, vol. 16, no. 8, p. 298, 2024. **DOI**: 10.3390/fi16080298

[18] R. Kozik, G. Ka ̨tek, M. Gackowska, S. Kula, J. Komorniczak, P. Ksieniewicz, A. Pawlicka, M. Pawlicki, and M. Chorás, "Towards explainable fake news detection and automated content credibility assessment: Polish internet and digital media use-case," *Neurocomputing*, vol. 608, p. 128 450, 2024. **DOI**: 10.1016/j.neucom.2024.128450

[19] S. E. V. S. Pillai, "Enhancing misinformation detection through semantic analysis and knowledge graphs," in *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*. IEEE, 2024, pp. 1–5. **DOI**: 10.1109/ICDECS59733.2023.10503553

[20] A. Bhagat, F. Mallick, N. Karia, and A. Kaushal, "Indeprop: Information-preserving de-propagandization of news articles (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 915–12 916. **DOI**: 10.1609/aaai.v36i11.21594

[21] H. Kaur, "Fake news detection using semantic analysis and machine learning techniques," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023, pp. 1–6. DOI: 10.1109/ICCCNT56998.2023.10307799

[22] A. A. Ali, S. Latif, S. A. Ghauri, O.-Y. Song, A. A. Abbasi, and A. J. Malik, "Linguistic features and bi-lstm for identification of fake news," *Electronics*, vol. 12, no. 13, p. 2942, 2023. DOI: 10.3390/electronics12132942

[23] P. Mitra and L. Jacob, "Fake news detection and classify the category," in *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*. IEEE, 2022, pp. 1–7. DOI: 10.1109/TQCEBT54229.2022.10041596

[24] C.-O. Truiča and E.-S. Apostol, "It's all in the embedding! fake news detection using document embeddings," *Mathematics*, vol. 11, no. 3, p. 508, 2023. DOI: 10.3390/math11030508

[25] A. Neelima and S. Mehrotra, "A comprehensive review on word embedding techniques," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*. IEEE, 2023, pp. 538–543. DOI: 10.1109/ICISCoIS56541.2023.10100347

[26] M. A. Al-Tarawneh, O. Al-irr, K. S. Al-Maaitah, H. Kanj, and W. H. F. Aly, "Enhancing fake news detection with word embedding: A machine learning and deep learning approach," *Computers*, vol. 13, no. 9, p. 239, 2024. DOI: 10.3390/computers13090239

[27] S. Pande, S. Rathod, R. Joshi, G. Chvan, D. Jadhav, P. Phutane, S. Gonge, and K. Kadam, "Fake news identification using regression analysis and web scraping," *IJSSE*, vol. 12, pp. 311–318, 2022. DOI: 10.18280/ijsse.120305

[28] S. A. Althubiti, F. Alenezi, and R. F. Mansour, "Natural language processing with optimal deep learning based fake news classification." *Computers, Materials & Continua*, vol. 73, no. 2, 2022. DOI: 10.32604/cmc.2022.028981

[29] P. J. Worth, "Word embeddings and semantic spaces in natural language processing," *International journal of intelligence science*, vol. 13, no. 1, pp. 1–21, 2023. DOI: 10.4236/ijis.2023.131001

[30] K. Durrheim, M. Schuld, M. Mafunda, and S. Mazibuko, "Using word embeddings to investigate cultural biases," *British Journal of Social Psychology*, vol. 62, no. 1, pp. 617–629, 2023. DOI: 10.1111/bjso.12560

[31] G. Boleda, "Distributional semantics and linguistic theory," *Annual Review of Linguistics*, vol. 6, no. 1, pp. 213–234, 2020. DOI: 10.1146/annurev-linguistics-011619-030303

[32] W. E. Nagy, P. A. Herman, and R. C. Anderson, "Learning words from context," *Reading research quarterly*, pp. 233–253, 1985. DOI: 10.2307/747758

[33] C. Van Petten, "Words and sentences: Event-related brain potential measures," *Psychophysiology*, vol. 32, no. 6, pp. 511–525, 1995. DOI: 10.1111/j.1469-8986.1995.tb01228.x

[34] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954. DOI: 10.1080/00437956.1954.11659520

[35] A. Lenci et al., "Distributional semantics in linguistic and cognitive research," *Italian journal of linguistics*, vol. 20, no. 1, pp. 1–31, 2008.

[36] C. P. Cook, *Exploiting linguistic knowledge to infer properties of neologisms*. University of Toronto, 2010.

[37] P. Edmonds, "Choosing the word most typical in context using a lexical co-occurrence network," *arXiv preprint cs/9811009*, 1998. DOI: 10.48550/arXiv.cs/9811009

[38] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior research methods*, vol. 39, pp. 510–526, 2007. DOI: 10.3758/BF03193020

[39] K. Stuart and A. Botella, "Corpus linguistics, network analysis and cooccurrence matrices," *International Journal of English Studies*, vol. 9, no. 3, pp. 1–20, 2009.

[40] J. Cong and H. Liu, "Linguistic emergence from a networks approach: The case of modern chinese two-character words," *Plos one*, vol. 16, no. 11, p. e0259818, 2021. DOI: 10.1371/journal.pone.0259818

[41] G. Budel, Y. Jin, P. Van Mieghem, and M. Kitsak, "Topological properties and organizing principles of semantic networks," *Scientific Reports*, vol. 13, no. 1, p. 11728, 2023. DOI: 10.1038/s41598-023-37294-8

[42] S. S. Birunda and R. K. Devi, "A review on word embedding techniques for text classification," in *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*. Springer, 2021, pp. 267–281. DOI: 10.1007/978-981-15-9651-3_23

[43] S. T. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology (ARIST)*, vol. 38, pp. 189–230, 2004. DOI: 10.1002/aris.1440380105

[44] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 298–307. DOI: 10.18653/v1/D15-1036

[45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. DOI: 10.48550/arXiv.1301.3781

[47] SentiOne, "Social listening 101," https://sentione.com/resources/sociallistening, 2023, accessed: 1 May 2023.

[48] J. R. Crawford and J. D. Henry, "The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample," *British journal of clinical psychology*, vol. 43, no. 3, pp. 245–265, 2004. DOI: 10.1348/0144665031752934

[49] B. Asztalos, G. Palla, and D. Czégel, "Anomalous diffusion analysis of semantic evolution in major indo-european languages," *Plos one*, vol. 19, no. 3, p. e0298650, 2024. DOI: 10.1371/journal.pone.0298650

[50] D. Czégel, S. G. Balogh, P. Pollner, and G. Palla, "Phase space volume scaling of generalized entropies and anomalous diffusion scaling governed by corresponding non-linear fokker-planck equations," *Scientific reports*, vol. 8, no. 1, p. 1883, 2018. DOI: 10.1038/s41598-018-20202-w

[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Bogdán Asztalos graduated from the Eötvös Loránd University, Faculty of Science as a Physicist (M.S. degree), and currently is a Phd student in Physics. His main research interest is the mathematical background of theoretical statistical physics, but also works on social network analysis, and looks for interdisciplinary applications of the tools of statistical physics. His supervisor is Péter Pollner.

Péter Bányász graduated in political science from the Faculty of Law and Political Science of the Eötvös Loránd University. He then obtained his doctorate at the Military Engineering Doctoral School of the National University of Public Service. His dissertation is entitled "Opportunities and Risks of Social Media in the Defence Sector." His research interests include the human aspect of cyber security, network theories of psychological operations, and the relationship between privacy and surveillance. He is an assistant professor at the Faculty of Public Governance and International Studies of the University of Public Service and a researcher at the Institute for Cyber Security Research. He is also a senior mentor teacher at the University's Advanced College for National Security and Tivadar Puskás Technical College for Advanced Studies. He is also an active member of several scientific societies.