Disinformation Research in Bulgarian. Datasets and Tools

Milena Dobreva

GATE Institute, Sofia University St Kliment Ohridski, 125 Tsarigradsko shose, Sofia, 1734 Bulgaria

Abstract

This paper presents the initial results of an ongoing study on developing a digital toolkit for experimentation with social media content in Bulgarian aimed at discovering disinformation. This effort is aimed at providing a basis for systematic studies of Bulgarian social media content. The initial datasets used for experiments are related to the Covid-19 pandemic on samples of posts from Facebook and Twitter. The paper also describes some initial ideas on building an infrastructure for collaboration in social media research in Bulgarian.

Keywords

Disinformation, classification, stakeholders, social media, Facebook, Twitter, CrowdTangle, API, media literacy

1 Introduction

Disinformation is not a new phenomenon, but with the expansion of digital content creation and delivery and of social media use, in particular, it became a serious problem impacting the wellbeing of citizens and the processes of informed decision making in our society.

In this paper, we use the term ‘disinformation’ which is one of the terms which captures the continuum of information that is not correct combined with the presence or absence of the intention of harm. Some of these terms also indicate a political intent (propaganda, astroturfing), genres (satire) and social and media phenomena (misinformation, fake news, rumours, hoax, etc.). Within the context of this paper, we use the term ‘disinformation’ as false information intended to mislead when it was created.

To a great extent, the challenges in progressing research in this domain are two-fold; first, the complexity introduced by the scale of the phenomenon and second, the need to develop collaborations that integrate different points of view, as discussed in Section 1.1.

1.1 Various perspectives

There is a growing body of **research** on various aspects of disinformation aimed at different stages of its lifecycle, such as its creation (by humans or technological tools) or the patterns of spread and the impact on human society. Issues around tracing its spread, quantifying its impacts and developing methods for automated content analysis, which would allow identifying its instances, are at the core of information studies in this area. Artificial intelligence research and methods for supervised learning, in particular, are among the most popular choices for advancing this domain. Several recent EC-supported research projects experimented with different methods of automated verification of information [1].

While this international research advances the methods for the analysis of content in the more popular languages around the globe, the media and social media content in Bulgarian still needs more focused research and tools that would help understand how disinformation spreads on a national level. Over the last decade, there have been a number of studies done mainly by academics engaged in media

Education and Research in the Information Society, September 27–28, 2021, Plovdiv, Bulgaria
EMAIL: Milena.Dobreva@gate-ai.eu
ORCID: 0000-0002-2579-7541

studies, library and information literacy studies and some pilot studies on the application of natural language processing, mostly in fake news detection. Besides the initial research attempts, within the current Bulgarian landscape, there is a number of institutions that engage actively with fact-checking coming primarily from the journalism, media and NGO sectors. However, a larger scale data infrastructure that would combine collecting and curating relevant datasets (e.g. social media content) and developing tools that facilitate the analysis of data in the Bulgarian language from large datasets are currently missing.

There is also poor visibility of the work done in Bulgaria in the international aggregations of fact news. For example, **fact-checking** organisations, mainly originating from the media and NGO sectors, are involved in detailed checks of thousands of statements appearing in the media and then spreading through social media. An aggregator of fact-checking information, the Poynter Institute, currently provides access to over 7,000 fact checks from over 70 countries [2]. Only six of them are related to Bulgaria and were submitted between March and September 2020; the first Bulgarian fact-checker `factcheck.bg` [3] is still not visible in this database.

Media literacy is the primary educational domain most engaged with developing skills for identifying disinformation. For several years, there has been European-wide metrics in media literacy - the media literacy index [4]. In 2021 Bulgaria ranked in 30th place out of 35 European countries. Numerous educational institutions provide educational programmes to improve citizens' ability to assess information, discover relevant sources, and apply critical thinking and rational judgment.

1.2 Bulgarian stakeholders and disinformation

The domain of disinformation develops with the active contribution of three types of stakeholders:

- **Governmental institutions**, whose interest is in providing trustworthy and timely information to the citizens. Disinformation is detrimental to these efforts; one contemporary example is the low uptake of Covid-19 vaccination in Bulgaria which is partially explained by the active disinformation around vaccines.
- **Organisations of journalists** and NGOs active in the domains of *human freedoms* and *literacy* acting as fact-checkers;
- **Educational institutions** provide training in the domains of information and media literacy. In some cases, libraries are among the active contributors to educating different segments of the population.
- **Academic institutions** and technology companies develop new methods for the discovery of disinformation.

Table 1 summarises some of the key institutions within the specific case of pandemic-related disinformation in the Bulgarian context.

Table 1: Example of stakeholders in the specific case of Covid-19 disinformation in Bulgaria

Type of stakeholder	Examples of institutions and their stakes
Government	• Ministry of Health: responding to misinformation campaigns; informing public policies; planning for campaigns; estimating the impact of disinformation related to the nation's health. • Coordination headquarters for the response to the pandemic: fine-tuned assessments of the disinformation.
Media organisations and NGOs	• Association of European Journalists: the first fact-checker in Bulgaria • Institute for the study of democracy: works on fact-checking, although its main focus is on disinformation related to politics.
Education	• Ministry of education: reaching the teachers in the secondary school and respectively the students • Libraries: providing information sessions and guidance on reference.

Requirements of different stakeholders have not been gathered and analysed on a systematic basis in Bulgaria. This paper is not focusing on a complete analysis of their requirements, but it would help to fine-tune future research efforts to existing needs. The table does not include the academic and technological stakeholders which usually approach the task identifying gaps in technological tools and experimenting with various approaches.

The combination of the needs in providing conditions for more comparable studies in social media analysis in Bulgarian with the need to enhance the visibility of work done in Bulgaria and support the educational efforts in media literacy in our county motivated the GATE Institute to start an effort in gradually developing a digital infrastructure will provide support for comparable research in social media analysis in Bulgarian. This paper is the first effort to summarise the initial work in this direction. It is structured as follows. Section 3 provides a succinct overview of relevant work. It also introduces SoMeChar, a new model to characterise datasets extracted from social media. Such characterisation helps compare datasets that grow rapidly in numbers. Section 3 presents two datasets (from Facebook and Twitter) and some initial observations on the data and on the challenges in human annotation using a model developed within a different cultural setting. Finally, section 4 explores potential directions for future work.

2 Related Work

In their systematic review of disinformation research, Meel and Vishwakarma [5] identify among the areas for future research the domains of datasets and multilinguality:

> *"Datasets: The establishment of convincing gold standard datasets in this field is highly required as most of the research is being done on customised datasets. Because of the lack of publicly available large-scale datasets a benchmark comparison between different algorithms cannot be done.*
>
> *Multilingual platform: Most of the work focuses on linguistic features in English language text. Other popular and regional languages (multilingual platform for fake news detection) are not considered yet."* [5, p. 22]

This paper looks at the ways to create a collection of reusable datasets for Bulgarian complemented by tools that could be used for various types of analysis.

2.1 Datasets and their analysis

Multiple studies aimed to analyse large-scale social media datasets for potential misinformation. Possibly the most ambitious dataset is **Mega-COV** [6]. Featuring 1.5 billion tweets, Mega-COV was designed for studying Covid-19. The dataset consists of tweets from 2007 to 2020 (the initial date precedes the spread of Covid-19 but is motivated by offering content that would allow analysing user behaviour on a larger scale). Mega-COV includes tweets in more than 100 languages. Not all languages could be identified from the original data, and the authors used two language identification tools to attribute languages of contents; applying such tools, however, resulted in the appearance of Latin language in the dataset [6: p. 3406]. It is hard to believe Latin is used for contemporary tweeting, and this is an illustration of how big data approaches need additional quality checks.

Another study used 17,463,220 tweets in English and identified 226 unique COVID-19 false stories for March 2020 using Goggle Factchecker API [7]. It explored the emotional response to different types of disinformation. Rumours and their spread via social media were analysed in [8].

The review of work on datasets shows that multiple developments are, however, difficult to compare and sometimes re-use. In order to capture this variety, we created a model for describing social media datasets that are currently being experimented with in Open Research Knowledge Graph (ORKG) [9]. We called this model SoMeChar.

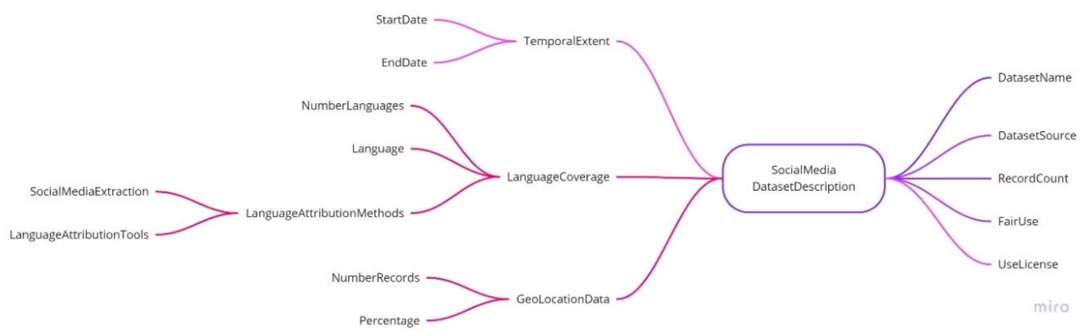

Figure 1: SoMeChar: New model to characterise datasets

2.2 Overview of leading trends in exploring disinformation in Bulgarian

The earliest efforts aimed at studying disinformation phenomena in Bulgarian came from the computer science community. Most studies focused on the disinformation class of fake news, with the earliest one dating back to 2016 [10] and further refinements in [11-14].

One study focused on the specific challenges in forming a dataset that can be processed via natural language processing tools. It concluded that the use of informal language and punctuation confuses parsing tools – this also calls for some amount of preprocessing of a social media dataset in order to make it processable. [15] Several studies of disinformation also came from the media studies community, which explored a range of issues such as the theoretical basis and positioning of disinformation phenomena and strategies to counteract it [16-18]. Last but not least, there is a study that analysed the role of librarians in reducing the impact of disinformation [19].

3 Dipping into Bulgarian Data

Embarking on the task to start developing a technological infrastructure for experimentation with social media analysis in Bulgarian, we started from two experiments to understand better the current state of social media. We took as an example social media coverage of Covid-19 and extracted a Facebook dataset of 300,000 posts from Facebook via Crowdtangle; from this set, we took subsets of some 3,000 most popular posts. We also extracted 5,200 tweets related to Covid-19. Below are some initial observations based on these data.

3.1 Example 1: Annotation of Facebook Posts

The disinformation team at GATE Institute started the practical work on the data infrastructure extracting an experimental dataset based on 52 public group contents from Facebook. The groups were formed around various Covid-19 issues. This topic was chosen due to the currency and the breadth of interest in it, and the multiple impacts of the pandemic on human life and the economies.

The extracted content of some 300,000 posts was sifted for the most popular posts. A subset of 600 posts was processed to produce sets of data with equal weights of the most popular topics identified by a FirstMonday study recently which identified five significant topics within the social media content in English, German and Russian. During summer school in July 2021, we used two content sets – of short (one sentence) and long (several sentences or paragraphs) posts – as an exercise in the annotation. Here we provide some illustrations from the annotation of the set of short texts. The annotation exercise was not preceded by a detailed explanation of how to make decisions on classifications – our team wanted to explore the intuitive responses of the participants. The reasoning behind this approach is that an instruction on how to make decisions in annotation changes the intuitive response and the way of reasoning and brings closer the annotators to the 'experts' who designed the annotation schema.

The sets had 20 posts, five of each of four topics. It was possible to add another value if the annotators thought none of the categories was suitable. The set was annotated by a group of 25 users of four educational and professional types. These preliminary results illustrate that different categories of users have somewhat different annotation preferences; e.g. the librarians were mainly using the proposed classification without adding categories while the undergraduate students made most suggestions for new categories (see Fig. 2). It is also interesting to note that many of the mismatched annotations are applying the class "conspiracy theories" – there seems to be an expectation that much of the content online is serving a conspiracy agenda (see Fig. 3).

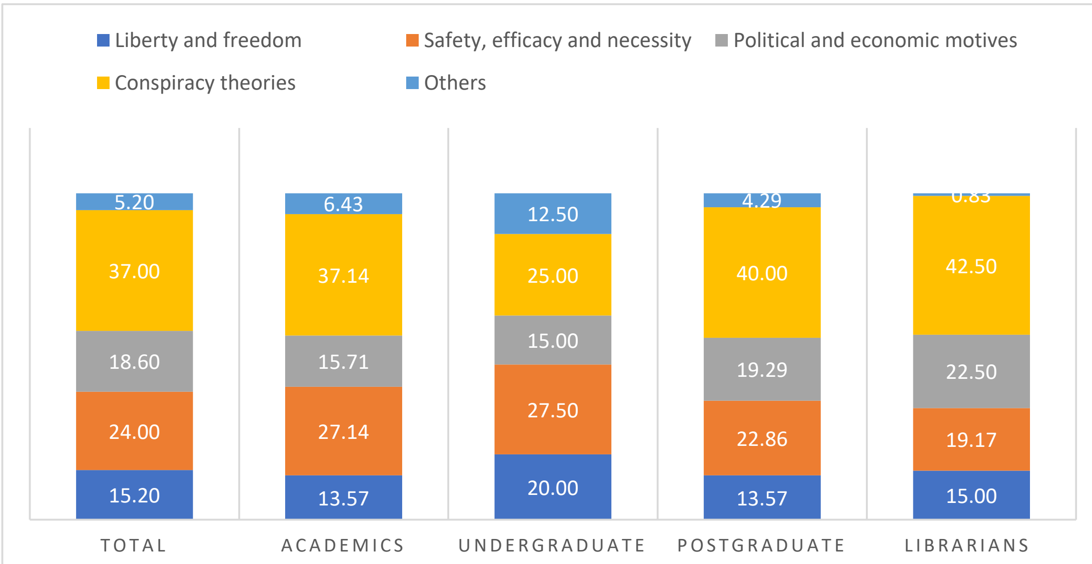

Figure 2: Distribution of the annotations by four groups of users (7 academics, 5 undergraduate, 7 postgraduate students and 6 librarians).

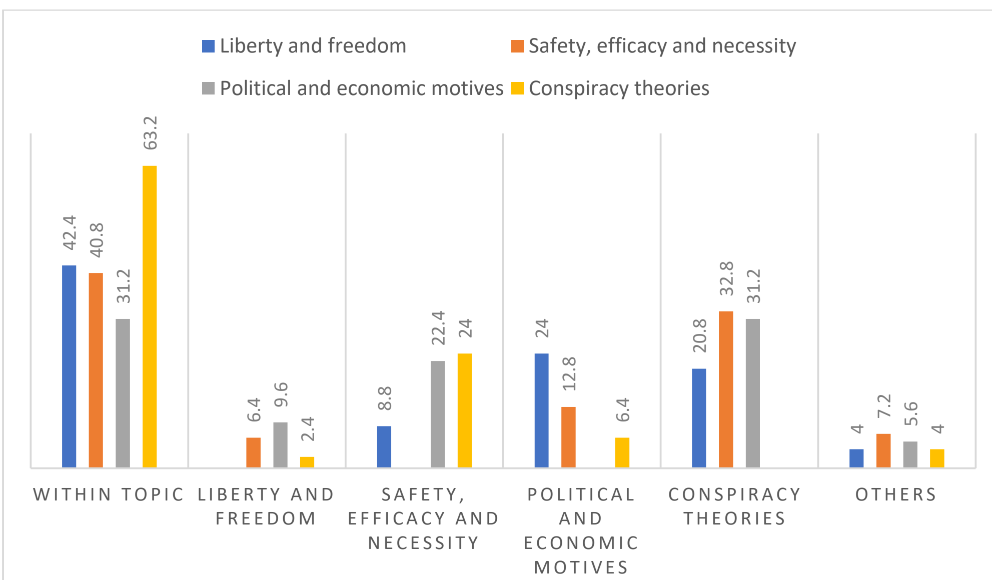

Figure 3: Distribution of the annotations which match the original annotation and mismatched annotations per category.

An additional aim was to check to what extent the proposed structure is convenient. We allowed participants to enter their suggested categories as ‘Other’. Fig. 4 captures their proposals, which include as most popular proposals “Fake news”, “Spam”, “Opinion” and “Illiteracy”.

Figure 4: A word cloud capturing the potential additional categories (entered under “Other”)

The experimentation with the most popular posts on Facebook demonstrated that annotation in Bulgarian following schemas created for other languages and cultures is likely to give preference to interpretations that state that a message was designed to spread a conspiracy theory.

In addition to this experiment which takes the most popular content from social media, we did a random sampling analysis based on Tweeter data.

3.2 Example 2: Analysis of tweets

In August 2021, GATE extracted 52.000 tweets related to Covid-19 and started a detailed manual analysis of a subset of 1.800 tweets.

The current research practice in social media datasets often deals with huge amounts of content and applies analytical tools to them. Still, the academic literature does not report extensively on the preparation of the sets. Our team spent substantial time processing a randomised subset of 1.800 tweets manually. It identified that the API extracted some tweets in Russian, Macedonians and other languages using the Cyrillic alphabet. In the initial processing, we included a manual language attribution. We also checked the tweets with embedded URLs during this stage and discovered that some 5% of the tweets are clickbaits. They are shaped as titles with a hyperreference to an article; however, the link most frequently leads to a ‘newsmedia-styled’ website without clear attribution of the owner and does not offer the advertised content.

In order to use tools for text analysis, we also added punctuation where a sentence did not have a full stop and unified the spelling of Covid-19. This set was found to have a vocabulary density of 0.495 with an average number of words per sentence 13.3.

The most frequent words included *Covid-19*, *Pfizer*, *vaccine* and *one*.

We also observed a very high use of named entities (names of politicians and famous personalities; institutions; countries). Currently, we are exploring the semantic clusters (negative ones including words such as *death*, *died*, *got critically ill*, etc.).

We also present an initial snapshot of the presence of different geographic entities in the explored subset on Fig. 5.

Figure 5: Geographic entities mentioned in the tweets' subset

4 Summary of results and agenda for future work

This paper provided an initial overview of previous work done on the analysis of social media content in Bulgaria and some experiments which looked at the properties of social media content in Bulgarian in Twitter and Facebook. The work done so far illustrates the importance of data management aspects (the issues of how datasets are extracted and prepared for use in the analysis are not usually discussed in detail in academic publications, but working with a high-quality dataset would influence the quality of the analysis. This is why we focused our work on the aspects of extracting, cleansing and enriching datasets from Bulgarian social media. In addition, we started experimenting with different analytical tools. By all means, this type of work will continue.

The main areas for future research in the domain of identifying disinformation in Bulgarian language will combine:

- **Developing a designated data and tools infrastructure:** this will be the first specialised infrastructure to collect and store datasets from Bulgarian social media which will be used for experimentation in the discovery of disinformation.
- **Capturing requirements of the stakeholders:** there is a clear gap in the understanding what tools would be most helpful for the different types of stakeholders – more work around this would help to identify how to shape analytic environments and what dataset structures are best suited for the tasks at hand.
- **Producing well documented anonymised datasets of social media content:** the datasets will be an essential contributor to open science research through sharing with other interested stakeholders, including other researchers.
- **Providing documented experimentation with existing tools:** since the process of analysing social media is quite complex and there are various types of tools needed, GATE will be collecting systematically and sharing the experience of using various tools. This information is precious for the scholarly community as it saves time. GATE will regularly publish technology watch reports and updates and will play the role of a specialised observatory in the domain of social media and disinformation analysis in Bulgarian.
- **Developing specifications of tools that are needed for the analysis of the content:** in cases when the quality of the existing tools is not sufficient or when no tools support specific analysis, GATE will be leading the development of new tools.
- **Designing educational materials on the use of the infrastructure:** these will support the stakeholder community in the whole spectrum of educational activities.

- **Experimenting impact assessment methodologies:** GATE will survey continuously what methodologies are being proposed in the impact assessment in the domain of disinformation. They will be applied in the work of the disinformation project and also will be shared with the stakeholders.
- **Events – summer schools, fake news discovery event:** GATE will regularly offer opportunities for spreading the knowledge on the topics of disinformation and social media analysis. The feedback from the first summer school delivered in July 2021 was overwhelmingly positive and emphasised the need to offer more educational programmes to different stakeholders.

4.1 Towards a platform

The tools which could help in the analysis of social media datasets are of different types and are sketched in Table 2. Not all tools are needed for each study – typical studies and the tools used will be summarised and explained as part of the educational activities. For the time being, we are listing the most popular types of tools. In the future, we will create a registry of existing tools and initiate in-house development where appropriate.

Table 2: Types of tools for the SoMePlatform

Type of tool	Examples
Data collection	• Bot detection tools for identifying non-human accounts. • Scrapers, APIs, specialised tools to extract the content (e.g. CrowdTangle)
Data cleaning	• Ad-hoc transformation including deduplication, removal of contents in other languages and anonymisation • Generic data cleaning tools
Data analysis	• Verification tools confirming that photos, videos are authentic. • Language analysis tools: • Syntax analysis (types of structures used) • Tools for Machine learning analysis • Incl. human and automated Classification • Credibility scoring tools assigning a verdict on a scale of accuracy/trustworthiness.
Visualisation	• Various tools visualising the analysis results.
Impact tools	• Tools that help to tract the itinerary of misinformation/disinformation or valorising the impact (damages on individuals/the society).
Educational tools	• Teaching content, games and other tools for individual use or teacher/librarian use.
Repositories of credible sources	• Aggregated lists of credible sources • Aggregated checked facts

The language analysis tools form a big group of potential instruments. These tools can support a variety of analytical tasks, for example:

1. Determining the volume of text (number of posts, posts' length, average length; text or a mixture of formats, number of words/characters). Capturing this for the dataset description.
2. Determining which language(s) is/are used (for Bulgarian confirming that all posts are in Bulgarian).

3. Conducting grapheme analysis: use of uppercase and lower-case letters (in particular, what parts of the messages are emphasised using capitals).
4. Punctuation analysis: any signs of excessive or sloppy punctuation.
5. Analysis of semiotic elements such as a) emoticons, b) abbreviations
6. Bag of words: distribution of words. Clustering of forms of the same word together. Nouns, verbs and other parts of speech. Stopwords, emotionally charged words.

For the time being, we have used some open-source tools, and this work will expand into systematic technology watch reports which compare various options for analysis.

The main ambition of these initial steps taken by GATE institute in the domain of disinformation research is the gradual development of an environment that would allow more academics, technologists and analysts to engage with analysis of Bulgarian social media content. This environment should offer datasets for experimentation, help in identifying what tools can support specific processing and, in a nutshell, would provide better data management in the disinformation research domain.

Acknowledgements

This research work has been supported by GATE project, funded by Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01.

The SoMeChar model was inspired by the work of Open Research Knowledge Graph (ORKG), https://www.orkg.org/orkg/.

I would like to thank Silvia Gargova, Ivo Dzhumerov and Hristiana Nikolaeva, interns of GATE Institute, for their involvement in analysing social media content.

References

[1] EC, Funded projects in the fight against disinformation, n.d. URL: https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/funded-projects-fight-against-disinformation_en

[2] The Poynter Institute, The CoronaVirusFacts, n.d. URL: https://www.poynter.org/ifcn-covid-19-misinformation/

[3] Association of European Journalists, FactCheck.bg, n.d. URL. https://factcheck.bg/

[4] M. Lessenski. Media Literacy Index 2021. Double Trouble: Resilience to Fake News at the Time of Covid-19 Infodemic. 14 pp.. 2021, URL: https://osis.bg/wp-content/uploads/2021/03/MediaLiteracyIndex2021_ENG.pdf

[5] P. Meel, D. K. Vishwakarma, Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities, Expert Systems with Applications, 153(2020), URL: https://doi.org/10.1016/j.eswa.2019.112986.

[6] M. Abdul-Mageed, A. Elmanaby, E. Nagoudi, D. Pabbi, K. Verma, R. Lin, Mega-COV: A Billion-Scale Dataset of 100+ Languages for COVID-19. Proc. of the 16th EACL (2021), pp. 3402-3420.

[7] M. Charquero-Ballester, J. Walter, I.A. Nissen, and A. Bechmann, Different Types of COVID-19 Misinformation Have Different Emotional Valence on Twitter (2021). Available at URL: http://dx.doi.org/10.2139/ssrn.3776140

[8] X. Song, J. Petrak, Y. Jiang, I. Singh, D. Maynard, and K. Bontcheva. Classification aware neural topic model for COVID-19 disinformation categorisation. *PLoS ONE,* 2021. 16(2). URL: https://doi.org/10.1371/journal.pone.0247086

[9] The Open Research Knowledge Graph, n.d. URL: https://www.orkg.org/orkg/

[10] M. Hardalov, I. Koychev, and P. Nakov. In search of credible news. Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications. Varna, Bulgaria, AIMSA'16 (2016) pp. 172-180.

[11] [11] G. Karadzhov, P. Atanasova, P. Nakov, I. Koychev. 2018. We built a fake news & click bait filter: what happened next will blow your mind! *Proceedings of Recent Advances in Natural Language Processing, pages 334–343, Varna, Bulgaria, Sep 4–6 2017 (2018) pp.* 334-343. URL: https://doi.org/10.26615/978-954-452-049-6_045

[12] [12] S. Shaar, F. Alam, G. Da San Martino, A. Nikolov, W. Zaghouani, P. Nakov, and A. Feldman. Findings of the NLP4IF 2021 Shared Tasks on Fighting the COVID 19 Infodemic and Censorship Detection. *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda* (2021), pp. 82-92.

[13] [13] T. Mladenova, I. Valova. Analysis of the KNN Classifier Distance Metrics for Bulgarian Fake News Detection. *3rd International Congress on Human Computer Interaction, Optimisation and Robotic Applications (HORA)*: 1-4. IEEE. (2021). URL: https://www.mediapool.bg/hakeri razrabotiha modeli za razkrivane na falshivi novininews264198.html

[14] [14] Y. Dinkov, I. Koychev, and P. Nakov. Detecting toxicity in news articles: Application to Bulgarian. arXiv preprint arXiv:1908.09785. 2019.

[15] [15] P. Psenova. Linguistic features of rumours. The social processes and their impact on language. In: Problemi na sociolingvistikata 13: 2018, pp. 225-231 (in Bulgarian).// П. Осенова. Лингвистични особености на слуховете. Обществените процеси и тяхното отражение в езика. Проблеми на социолингвистиката 13, Международно социолингвистическо дружество: 2018. стр. 225-231, ISSN 1314 5401. (in Bulgarian).

[16] [16] I. Vylchanov. A taxonomic point of view on fake news. In: Medii I komunikacii na 21 wek, 2018. 2(1): pp. 34-30. // И. Вълчанов. Таксономичен поглед към фалшивите новини. Медии и комуникации на 21. век, 2018. 2 (1):24-30. (in Bulgarian).

[17] [17] I. Mavrodieva. Fake news: theoretical dilemmas, methodological aspects and manifestations in times of crysis communication. In: Medii I kpomunikacc na 21 wek, 2020, 2020, 4(1), pp. 47-55. // И. Мавродиева. Фалшиви новини: теоретични дилеми, методически аспекти и проявления в кризисна комуникация. *Медии и комуникации на 21. век*, 2020. 4 (1): 47-55. (in Bulgarian).

[18] [18] I. Indzhov. Compabating fake news using the legislative instruments and the self-regulation in Bulgaria. In: Medii I komunikacii na 21 wek, 2018, 2(1), pp. 48-60. // И. Инджов. Борба с фалшивите новини с инструментариума на закона и на саморегулацията българският случай. *Медии и комуникации на 21. век,* 2018. 2 (1):48 60. (in Bulgarian).

[19] [19] M. Tsvetkova. Librarian as a shield against fake news. In; BBIA online (2018), pp. 16-21. М. Цветкова. In: *BBIA online: The Journal of the Bulgarian Library and Information Association*, 2018. 8 (6): 16-21. (in Bulgarian).