sciendo
DOI 10.2478/jazcas-2025-0025

ANNOTATED SLOVAK DATASETS FOR TOXICITY, HATE SPEECH, AND SENTIMENT ANALYSIS

ZUZANA SOKOLOVÁ[1] – MAROŠ HARAHUS[2] – DANIEL HLÁDEK[3] – JÁN STAŠ[4]
[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-2337-8749)
[2]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-1756-123X)
[3]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-1148-3194)
[4]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0001-7403-0235)

SOKOLOVÁ, Zuzana – HARAHUS, Maroš – HLÁDEK, Daniel – STAŠ, Ján: Annotated Slovak Datasets for Toxicity, Hate Speech, and Sentiment Analysis. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 279 – 289.

Abstract: The rise of social media has led to an increase in toxic language, hate speech, and offensive content. While extensive research exists for widely spoken languages like English, Slovak remains underrepresented due to the lack of high-quality datasets. This gap limits the development of effective models for toxicity detection and sentiment analysis in Slovak. To address this, we introduce three new annotated Slovak datasets focused on toxic language, offensive language, hate speech detection, and sentiment analysis. These native datasets provide a more reliable foundation for automated moderation compared to machine-translated alternatives. Our research also highlights the real-world impact of online toxicity, including social polarization and psychological distress, emphasizing the need for proactive detection systems on social media platforms. This paper reviews existing Slovak datasets, presents our newly developed resources, and provides a comparative analysis. Finally, we outline key contributions and suggest future directions for improving toxic language detection in Slovak.

Keywords: datasets, hate speech, natural language processing, sentiment analysis, Slovak language, toxic language

1 INTRODUCTION

Natural language processing (NLP) is gaining popularity, driven by the increasing online presence of people and their active use of social media to discuss various topics like politics, the climate crisis, celebrity manners, movie reviews and

the like. Unfortunately, these platforms, instead of fostering constructive discussions, are becoming toxic environments filled with hate speech and hostility.

Rising rivalry, arrogance, and resentment among users contribute to social polarization. Social media comments often turn into arguments, insults, and attempts to prove superiority. Detecting toxicity is crucial to mitigating these negative effects and promoting a healthier online space.

Negative online behaviour can have serious consequences, including mental health issues, self-harm, and substance abuse (Chen 2023; Park 2024; Stroińska 2020). Addressing this issue is essential to reducing harmful speech and creating a safer digital environment. Our research focuses on detecting hate speech and offensive language on social media, aiming to foster respectful and constructive discussions online primarily in the Slovak language.

The rapid expansion of online communication and social media has led to a surge in toxic language, hate speech, and offensive expressions. While numerous studies have focused on detecting such language in widely spoken languages like English, research on Slovak remains scarce. The absence of high-quality Slovak datasets significantly limits the development and evaluation of models for detecting toxicity, offensive speech, and hate speech, as well as sentiment analysis in this language.

Social media platforms, such as Facebook and X, primarily rely on user-reported content to handle harmful language, rather than proactively addressing the issue. However, with access to vast amounts of textual data, these platforms have the potential to implement more effective automated detection systems. Detecting harmful language is essential not only for reducing online toxicity but also for mitigating its real-world consequences, including social polarization, psychological distress, and hate-driven violence. At the same time, we focus on creating native Slovak datasets because machine-translated datasets still do not achieve the same level of effectiveness, as discussed in Sokolová et al. (2023).

To address this gap, we introduce three new annotated Slovak datasets focused on toxic language, offensive language, and hate speech detection, as well as sentiment analysis. These datasets are designed to support the development of robust models tailored to the Slovak language, enabling more effective moderation and analysis of online discourse. Our paper contributes to the growing need for multilingual NLP resources and aims to foster a healthier and safer online environment.

In Section 1 we briefly outline the motivation for creating datasets in the Slovak language and emphasized why machine translation of datasets remains inefficient. Section 2 focuses on existing publicly available Slovak datasets related to toxic language, hate speech, offensive language, and sentiment analysis. As part of study Sokolová (2024), two annotated datasets were created—one for toxic language and another for sentiment analysis. Additionally, annotated a hate speech dataset was

developed as part of a bachelor thesis (Ferko 2024). All datasets are introduced and compared in detail in Section 3, while Section 4 summarizes the scientific contributions of this paper and suggests future directions in detecting toxic language, hate speech, offensive language, and sentiment analysis.

2 RELATED WORK

2.1 Comparison of global datasets

Datasets of textual data worldwide focus on multiple categories of hate speech. In Tab. 1, we present the most well-known, widely used, and verified datasets intended for hate speech detection. However, examples of hate speech in some datasets are not entirely clear, such as the text dataset by Waseem and Hovy (2016) or hierarchical datasets. Moreover, these datasets are of low quality because they are not regularly updated, even though X users adopt new phrases or abbreviations. Additionally, approximately 60% of dataset creators found agreement among annotators (Poletto et al. 2021). Therefore, a useful predictive detection model for hate speech requires relevant and up-to-date datasets. The maturity of datasets is considered a unique challenge for top-quality systems.

According to Kocoń et al. (2021), the separation of annotator groups has a significant impact on the performance of hate detection systems. They also stated that group consensus affects recognition quality. It has been demonstrated that the identity of people who publish tweets introduces bias into the dataset, making it difficult to compile and ensure the quality of negative data. This means that implicit hate speech is therefore difficult to measure (Wiegand et al. 2021). Additionally, many datasets overlap between class labels, as shown by Waseem (2016), who found an overlap of 2,876 tweets between the Waseem and Hovy dataset.

In their analysis, Alkomah and Ma (2022) showed that research requires more robust, reliable, and extensive datasets due to the broad applications of hate speech detection. Vashistha and Zubiaga (2020) created a robust and massive dataset by combining four well-known datasets. Their merged dataset included HASOC (Mandl et al. 2019) and SemEval, which are among the most popular datasets. HASOC is divided into three sub-tasks:

- the first focuses on identifying hate speech and offensive language,
- the second focuses on identifying the type of hate speech,
- the third focuses on identifying the target group (or individuals) of hate speech.

Basile et al. (2019) focused on multilingual hate speech detection against immigrants and women on the X platform using the SemEval Task 5 dataset. Zampieri et al. (2019a), in their study addresses the identification and categorization of offensive language on social media using the SemEval Task 6 dataset. The latest OLID dataset (Zampieri et al. 2019b) for offensive language identification contains

over 14,000 English tweets and is aimed at similar tasks as the HASOC dataset. The HASOC 2020 dataset (Mandl et al. 2020b) contains only 3,708 English tweet samples, but is considered substantial and competitive.

Mishra et al. (2020) achieved an F1 score of 51.52% in the first task for English when classifying tweets into two categories: whether a tweet is hateful and offensive or the opposite. In the second task, they achieved an F1 score of 23.41%, where tweets (labelled as hateful and offensive in the first task) were classified into three categories: hateful, offensive, and disrespectful.

ElSherief et al. (2018), in their study, compiled a dataset for hate speech containing 27,330 tweets. They also managed to extract 25,278 instigators of hate speech and 22,287 target accounts. Their research focused on comparing hate speech instigators, their targets, and general X users. They found that hate instigators tend to target more visible users and that participation in hateful discussions is associated with higher visibility. Additionally, it was shown that both instigators and targets of hate have unique personality traits that may contribute to hate speech, such as anger or depression.

Davidson et al. (2017), in their study, classified textual data into three categories (hateful, offensive, neutral). They found that racist and homophobic tweets are more likely to be classified as hate speech, whereas sexist tweets are generally classified as offensive. Other studies that also focus on dataset creation and classification are listed in Tab. 1, along with the corresponding categories and the number of tweets.

2.2 Comparison of Slovak datasets

The detection of toxicity, meaning the identification of hate speech and offensive language in the Slovak language, has so far been the subject of very few scientific studies. In Tab. 2 we have listed the available corpora of textual data in Slovak, where the focus of individual datasets and their size can also be seen. Most commonly, authors have classified hate speech into two categories (hateful, neutral). Alternatively, datasets have been divided into three categories (positive, negative, neutral) or even four categories (neutral, mildly toxic, moderately toxic, and highly toxic).

Author / Dataset Name / Reference	Dataset Size (No. Tweets)	Dataset Categories
Waseem and Hovy (2016a)	16,000	Racism, Sexism, Neither
Waseem et al. (2016b)	6,909	Racism, Sexism, Neither, Both
Davidson et al. (2017)	24,783	Hateful, Offensive, Neither
Harassment (Golbeck et al. 2017)	35,000	Harassing, Neutral
Twitter & Reddit SA (Shen and Rudzicz 2017)	162,980 & 37,249	Positive, Neutral, or Negative

ElSherief et al. (2018)	27,330	Archaic, Class-based, Disability, Ethnicity, Gender, Religion, Sexual Orientation
Founta et al. (2018)	80,000	Offensive, Abusive, Hateful, Aggressive, Cyberbullying, Spam, Normal
Amievalita (Fersini et al. 2018)	4,000	Misogynistic, Discrediting, Sexual Harassment, Stereotype, Dominance
Women (Fersini et al. 2018)	3,977	Misogyny, Stereotype, Dominance, Sexual Harassment, Discrediting, Misogyny Target
OLID (Zampieri et al. 2019a)	14,000	Offensive, Non-offensive, Targeted Insults. Individual, Group
L-HSAB (Mulki et al. 2019)	5,846	Hateful, Offensive, Normal, Targeted
HASOC (Mandl et al. 2019)	5,335	Hateful and Non-offensive
	7005	Hateful, Offensive, Vulgar
Ousidhoum et al. (2019)	5,647	Hateful, Offensive, Neither, Directness, Hostility, Target
MMHS150K (Winter et al. 2019)	150,000	Neutral, Religion, Sexism, Racism, Homophobia, Other Hate
AbusEval (Caselli et al. 2020)	18,740	Offensive, Non-offensive, Targeted, Non-targeted, Explicitly Insulting, Implicitly Insulting, Non-insulting
HatEval (Yang et al. 2020)	13,000	Hateful, Neutral, Individual Target, Group Target
HateXplain (Mathew et al. 2021)	20,148	Hateful, Offensive, Normal
Sentiment Analysis (Shrivastava 2023)	905,874	Positive, Negative
Flipkart (Vaghani et al. 2023)	205,053	Positive, Neutral, or Negative
Youtube Statistics (Patil 2023)	19,658	Positive, Negative, Neutral

Tab. 1. Comparison of Global Corpora

Author / Dataset Name / Reference	**Dataset Size (No. Tweets)**	**Dataset Categories**
Sentigrade (Krchnavy and Simko 2017)	1,584	Positive, Negative, Neutral
Švec et al. (2018)	80,000	Hateful, Neutral
Machová et al. (2022a)	24,000	Positive, Negative, Neutral
Machová et al. (2022b)	3,092	Neutral, Mildly Toxic, Moderately Toxic, Very Toxic
Mojžiš and Kvassay (2022)	2,283	Hateful, Neutral
	10,000	Hateful, Neutral
Papcunová et al. (2023)	283	Hateful, Neutral

Tab. 2. Comparison of Slovak Corpora

3 DATASETS

In machine learning tasks, a dataset is required to train a model for performing various machine learning or deep learning tasks. The reason why a dataset is necessary is that machine learning heavily depends on data. Without data, artificial intelligence cannot learn, making it the most important aspect that enables the training of machine learning algorithms. Regardless of the skills or knowledge of the team and the size of the dataset, if the dataset is not of sufficient quality, the entire artificial intelligence project will not achieve satisfactory results.

Criteria	Value
Number of Annotators	7
Age	25–40
Gender	Women and Men
Education	PhD Students and Research Assistants From DEMC

Tab. 3. Basic characteristics of the annotators of ToxicSK and SentiSK datasets

Criteria	Value
Number of Annotators	60
Age	18–22
Gender	Women and Men
Education	1st and 2nd Year Bachelor's Students

Tab. 4. Basic characteristics of the annotators of hate_speech_slovak dataset

When working with artificial intelligence, we largely rely on the dataset. From training, tuning, model selection, to testing, we use a dataset divided into three sets: training, validation, and test sets. The training set is used to train the model, the validation set is used to adjust weights and fine-tune the model, and the test set is used to evaluate the trained model. Often, simply gathering data is not enough; on the contrary, in most artificial intelligence tasks, classifying and annotating the dataset takes the majority of the time, especially for corpora that are sufficiently accurate to reflect a realistic vision of the world.

In this section, we present the created datasets SentiSK, ToxicSK, and hate_speech_slovak. In Tab. 3, we provided the basic characteristics of the annotators who participated in annotating the created ToxicSK and SentiSK datasets. In Tab. 4, we outlined the key characteristics of the annotators involved in labeling the hate_speech_slovak dataset. All comments contained in these datasets were obtained through our custom-developed web scraping tool and were publicly accessible at the time of collection. The preprocessing pipeline involved the removal of duplicate entries and URLs.

3.1 Dataset: ToxicSK

The ToxicSK dataset (TUKE-KEMT/toxic-sk 2024) was created as part of a research task focused on detecting toxicity on social media. We focused on the Slovak language. The comments is a collection of public posts on the Facebook social network.

The collected comments were annotated using the Prodigy tool into two categories: toxic (1) and non-toxic (0). The ToxicSK dataset is class-balanced and contains 4,420 toxic and 4,420 non-toxic comments.

Dataset	ToxicSK
Number of Comments	8,840
Number of Categories	2
Type of Categories	Toxic (1), Non-toxic (0)
Number of Negative Comments	4,420
Number of Positive Comments	4,420
Number of Words	89,756
Number of Characters	476,170
Average Number of Words per Sentence	10.15
Number of Unique Words	18,883
Number of Unique Words	11,602
Number of Stopwords	20,958
Data Source	Facebook

Tab. 5. Specification of the ToxicSK dataset

3.2 Dataset: hate_speech_slovak

The hate_speech_slovak dataset (TUKE-KEMT/hate_speech_slovak 2024) contains posts from a social network that have been annotated by humans. Each post is labelled by 1, if contains hateful or offensive language, and by 0 if not. The data was collected from a variety of public pages on topics such as sports, politics, and general discussions. To ensure the quality of the data, the collected posts underwent a cleaning process using text clustering. The annotations were provided by a group of students from the Technical University of Košice in Slovakia.

To maintain reliability, the dataset underwent a filtering process to remove annotations from users who showed a low level of agreement with others. Annotations were evaluated based on a scoring system: annotators received positive points when their annotations aligned with others and negative points when they differed. Any annotator with a low agreement ratio (below 70%) was excluded from the dataset. Additionally, for each post, votes for the positive, neutral, and negative categories were calculated from the remaining reliable annotators, with posts where the neutral class was the majority being discarded. Despite these efforts, some bias remains in the dataset due to the personal opinions of the annotators. For most items, the class was determined by the votes of trustworthy annotators, but in some cases, items had only a single vote.

Dataset	hate_speech_slovak
Number of Comments	13,189
Number of Categories	2
Type of Categories	Hate Speech (1), Neutral (0)
Number of Hate Speech Comments	3,605
Number of Neutral Comments	9,584
Number of Sentences	11,870
Number of Words	218,984
Number of Characters	1,130,860
Average Number of Words per Sentence	18.45
Number of Unique Words	42,031
Number of Unique Words	28,649
Number of Stopwords	50,151
Data Source	Facebook

Tab. 6. Specification of the hate_speech_slovak dataset

3.3 Dataset: SentiSK

The SentiSK dataset (TUKE-KEMT/senti-sk 2024) was created as part of research focused on sentiment analysis in the Slovak language. The SentiSK dataset contains 34,006 comments from the social media platform Facebook. The comments were collected using a Python tool for extracting data from websites, specifically comments under posts by three Slovak politicians. Data preprocessing involved cleaning the text of unwanted characters, as well as removing empty lines, extra spaces, periods, etc. The NLTK library was used for preprocessing. The dataset was annotated using the Prodigy annotation tool provided by the Department of Electronics and Multimedia Communications (DEMC). The SentiSK dataset was annotated into three sentiment categories: 20,668 negative comments, 9,581 neutral comments, and 3,779 positive comments. The distribution of comments in these categories indicates that the SentiSK dataset is class-imbalanced. Since the data were taken from the posts by Slovak politicians, there was a high number of negative comments.

Dataset	SentiSK
Number of Comments	34,006
Number of Categories	3
Type of Categories	Negative, Neutral, Positive
Number of Negative Comments	20,668
Number of Neutral Comments	9,581
Number of Positive Comments	3,779
Number of Words	401,937

Number of Characters	2,213,773
Average Number of Words per Sentence	11.82
Number of Unique Words	65,049
Number of Unique Words	43,365
Number of Stopwords	90,376
Data Source	Facebook

Tab. 7. Specification of the SentiSK dataset

4 CONCLUSION

This research highlights the growing need for Slovak-specific datasets in the field of toxic language, hate speech, and sentiment analysis. While significant progress has been made in detecting harmful language in widely spoken languages, Slovak remains underexplored, limiting the effectiveness of moderation systems (Cao et al. 2023; Jaggi et al. 2024; Lee et al. 2024; Hee et al. 2024). By analyzing 26 existing datasets and introducing three new annotated datasets—ToxicSK, SentiSK, and hate_speech_slovak—we contribute to closing this gap and provide a solid foundation for future advancements in Slovak NLP.

Our findings emphasize that native datasets significantly improve detection accuracy compared to machine-translated alternatives. Furthermore, we underscore the importance of automated detection systems in combating online toxicity and its real-world consequences. Moving forward, future research should focus on expanding dataset size, improving annotation consistency, and integrating advanced machine learning techniques to enhance detection models.

Given recent advances in large language models, future research should consider leveraging pre-trained and instruction-finetuned LLMs for toxicity detection in Slovak, as these approaches may offer improved performance even in under-resourced settings.

By fostering a more robust NLP ecosystem for Slovak, this work aims to support safer and healthier online interactions while contributing to multilingual NLP advancements.

ACKNOWLEDGEMENTS

The research in this paper was supported by the Ministry of Education, Research, Development, and Youth of the Slovak Republic under the research projects KEGA 049TUKE-4/2024 and KEGA 041TUKE-4/2025, and by the Slovak Research and Development Agency under the research projects APVV-22-0261 and APVV-22-0414. There was no additional external funding received for this study.

References

Alkomah, F., and Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. Information, 13(6), 273 p.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic evaluation, pp. 54–63.

Cao, Y. T., Domingo, L. F., Gilbert, S. A., Mazurek, M., Shilton, K., and Daumé III, H. (2023). Toxicity detection is not all you need: Measuring the gaps to supporting volunteer content moderators. Accessible at: arXiv preprint arXiv:2311.07879.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., and Granitzer, M. (2020, May). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In Proceedings of the twelfth language resources and evaluation conference, pp. 6193–6202.

Chen, M. B., Lau, J. H., and Frermann, L. (2023). The uncivil empathy: Investigating the relation between empathy and toxicity in online mental health support forums. In Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association, pp. 136–147.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media, 11(1), pp. 512–515.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In Proceedings of the International AAAI Conference on Web and Social Media, 12(1).

Ferko, V., (2024). Anotácia a vyhodnotenie slovenskej databázy nenávistnej reči. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, 55 p. Vedúci práce: doc. Ing. Daniel Hládek, PhD.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). In CEUR workshop proceedings, Vol. 2263, pp. 1–9. CEUR-WS.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the international AAAI conference on web and social media, 12(1).

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., ... and Wu, D. M. (2017, June). A large labeled corpus for online harassment research. In Proceedings of the 2017 ACM on web science conference, pp. 229–233.

Hee, M. S., Sharma, S., Cao, R., Nandi, P., Nakov, P., Chakraborty, T., and Lee, R. (2024). Recent advances in online hate speech moderation: Multimodality and the role of large models. Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 4407–4419.

Jaggi, H., Murali, K., Fleisig, E., and Bıyık, E. (2024). Accurate and Data-Efficient Toxicity Prediction when Annotators Disagree. Accessible at: arXiv preprint arXiv:2410.12217.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., and Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. Information Processing & Management, 58(5), 102643.

Krchnavy, R., and Simko, M. (2017). Sentiment analysis of social network posts in Slovak language. In 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 20–25.

Kvassay, M. (2022). New Public Dataset for Classification of Inappropriate Comments in Slovak language. In 2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA), pp. 437–441.

Lee, N., Jung, C., Myung, J., Jin, J., Camacho-Collados, J., Kim, J., and Oh, A. (2023). Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. Accessible at: arXiv preprint arXiv:2308.16705.

Machová, K., Mach, M., and Vasilko, M. (2022). Recognition of toxicity of reviews in online discussions. Acta Polytechnica Hungarica, 19(4).

Machová, K., Mach, M., and Adamišín, K. (2022). Machine learning and lexicon approach to texts processing in the detection of degrees of toxicity in online discussions. Sensors, 22(17), 6468.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation, pp. 14–17.

Mandl, T., Modha, S., Kumar M, A., and Chakravarthi, B. R. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In Proceedings of the 12th annual meeting of the forum for information retrieval evaluation, pp. 29–32.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI conference on artificial intelligence, 35(17), pp. 14867–14875.

Mishra, A. K., Saumya, S., and Kumar, A. (2020). IIIT_DWD@ HASOC 2020: Identifying offensive content in Indo-European languages. In FIRE (working notes), pp. 139–144).

Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (2019). L-hsab: A levantine twitter dataset for hate speech and abusive language. In Proceedings of the third workshop on abusive language online, pp. 111–118.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D. Y. (2019). Multilingual and multi-aspect hate speech analysis. Accessible at: arXiv preprint arXiv:1908.11049.

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., ... and Adamkovič, M. (2023). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. Complex & intelligent systems, 9(3), pp. 2827–2842.

Park, K., Baik, M. J., Hwang, Y., Shin, Y., Lee, H., Lee, R., ... and Park, S. (2024). Harmful Suicide Content Detection. Accessible at: arXiv preprint arXiv:2407.13942.

Patil, A., (2023). Youtube Statistics, Accessible at: https://www.kaggle.com/datasets/advaypatil/youtube-statistics.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation, 55, pp. 477–523.