NLP-MisInfo-2023 - Abstract - Countering Malicious Content Moderation Evasion in Online Social Networks: Simulation and Detection of Word Camouflage

Álvaro Huertas-García[1], Alejandro Martín[1], Javier Huertas-Tato[1] and David Camacho[1]

[1]*Department of Computer Systems Engineering, Universidad Politécnica de Madrid, St. Ramiro de Maeztu, 28040, Madrid, Spain*

Abstract

This research introduces novel methodologies and tools to combat content evasion in multilingual Natural Language Processing on social networks. A unique Python package, "pyleetspeak", is developed, offering a customizable system for simulating multilingual content evasion through word camouflage techniques. The study also presents a synthetic multilingual dataset of camouflaged words, facilitating the training of models for camouflage detection. In a comparative analysis of various models, the multilingual MPNET-ideal model, pre-trained on an extended mSTSb dataset, outperforms other models in detecting camouflaged content across languages. The research underscores the utility of the tool in improving content moderation, enhancing online security, and serving as a potential data augmentation tool for AI systems. This work constitutes a significant contribution towards combating information disorders on social networks and sets the stage for further research in this field.

Keywords

Information Disorders, Leetspeak, Word camouflage, Multilingualism, Content Evasion

Social networks are constantly contending with malicious information or 'information pollution' that can polarize media discourse [1, 2]. As a countermeasure, content moderation is employed, involving the review and removal of content that breaches platform rules. However, with users persistently devising methods to evade these efforts, the potential for harmful or illegal content to spread grows, posing a risk to the platform's reputation and its users' safety.

The emergence of the COVID-19 pandemic highlighted the urgency of effectively combating information disorders and enhancing content moderation [3, 4]. Malicious actors have been exploiting content moderation rules, resorting to evasion techniques such as leetspeak and word camouflaging [5, 6]. This study presents a unique and novel solution, introducing a customizable methodology for simulating multilingual content evasion[1], a curated synthetic

Woodstock'21: Symposium on the irreproducible science, June 07–11, 2021, Woodstock, NY

alvaro.huertas.garcia@upm.es (Á. Huertas-García); alejandro.martin@upm.es (A. Martín); javier.huertas.tato@upm.es (J. Huertas-Tato); david.camacho@upm.es (D. Camacho)

https://aida.etsisi.upm.es/ (D. Camacho)

0000-0003-2165-0144 (Á. Huertas-García); 0000-0002-0800-7632 (A. Martín); 0000-0003-4127-5505 (J. Huertas-Tato); 0000-0002-5051-3475 (D. Camacho)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://pypi.org/project/pyleetspeak/

dataset of camouflaged words[2], and a multilingual Transformer-based model[3] to identify various word camouflage techniques and prevent content evasion over 20 languages[4]. The efficacy of multilingual pre-training in semantic similarity for enhancing such models is also explored.

A novel system for simulating multilingual content evasion through word camouflage techniques is developed based on literature references [5, 6, 7, 8] and strategies observed on social media. This system includes three unique modules: LeetSpeaker, PunctuationCamouflage, and InversionCamouflage, all of which have been embedded into the Python package "pyleetspeak". The LeetSpeaker module uses 'leetspeak', a character replacement system (i.e., "vaccination" into "v@ccin@tion" or "v4ccin4tion"), while the PunctuationCamouflage module inserts punctuation marks within words to confound content moderation algorithms (i.e., "COVID-19" is transformed into "C.O.V.I.D.-1.9"). Lastly, the InversionCamouflage module scrambles words by reversing the order of syllables (i.e., "Methodology" can be changed to "Me-do-tho-lo-gy).

The "pyleetspeak"[1] package also showcases its utility as a data generator, which uses KeyBERT to extract semantically relevant words, apply camouflage methods, and generate data annotated in Spacy format. The data is tagged with four entities representing different camouflage methods, and a dictionary detailing parameters applied to each instance ensures process interpretability.

An experimental protocol was designed to address the problem of word camouflage in multilingual content. The protocol starts with the creation of a synthetic multilingual dataset from non-camouflaged text data. This dataset, curated from various sources (OPUS News-Commentary [9], OPUS ParaCrawl [9], TED2020 [10] and WikiMatrix [11]), is used to train models to recognize camouflaged entities in monolingual and multilingual contexts. After camouflaging, the data is divided into training, validation, and testing sets, ensuring the camouflage stems exclusively from our generator tool.

To handle the task of word camouflage detection, a variety of models is employed. These include *paraphrase-multilingual-mpnet-base-v2* (MPNET-base) [12], *mstsb-paraphrase-multilingual-mpnet-base-v2* (MPNET-ideal)[3] [13], *bloomz-560m* [14], *xlm-roberta-base* [15], and *bert-base-multilingual-cased* [16]. These models are fine-tuned using the Spacy interface, establishing a comprehensive training architecture for the task at hand.

The developed model and the curated dataset are made publicly available for broader research and application. The open accessibility of these resources promotes transparency, encourages reproducibility, and potentially enables further advancements in the field of content evasion detection.

In a research effort to develop the best multilingual NER model for word camouflage detection, the study conducted various experiments and presented impressive findings. The most striking result was the performance of the MPNET-ideal model, a version of the MPNET that was pre-trained using the semantic textual similarity task with a multilingual extended mSTSb dataset. The MPNET-ideal outperformed all other trained multilingual models across most datasets, demonstrating its superiority in word camouflage detection. Specifically, the model exhibited improved performance over the monolingual baseline models, with the most substantial enhancement in Italian language detection where the F1 score went from 0.7061 to

[2] https://github.com/Huertas97/XX_NER_WordCamouflage

[3] https://huggingface.co/Huertas97/xx_LeetSpeakNER_mstsb_mpnet

[4] ar, az, da, de, el, en, es, fi, fr, hu, id, it, kk, nb, ne, nl, pt, ro, ru, sl, sv, tg, tr

0.8913.

The models were also evaluated across different camouflage techniques, revealing that detection of inversion camouflage was more challenging compared to punctuation or leetspeak camouflage. The results suggested that the MPNET-ideal multilingual model could accurately detect camouflaged entities across multiple languages and different types of text with high precision and recall. It was further demonstrated that the model could effectively differentiate between different camouflage techniques and handle a variety of languages. For instance, the confusion matrices revealed the difficulty of differentiating "MIX" entities from "LEETSPEAK" or "PUNCT_CAMO" entities due to the mixed elements, but the MPNET-ideal model still performed admirably.

Finally, the research validated the model's performance using an external tool, AugLy [17]. Though designed for monolingual data augmentation, AugLy could apply transformations that resembled camouflage techniques, making it an apt tool for external validation. The study discovered that the model could accurately detect new camouflage strategies, such as upside-down letters or emoticons in place of letters. However, it struggled to detect modifications in less semantically meaningful words like articles and pronouns. This shortcoming highlighted the importance of focusing on semantically meaningful words when dealing with camouflage detection. Overall, the MPNET-ideal model's validation results underlined its impressive capabilities in detecting various camouflage techniques, cementing its position as an effective tool for multilingual word camouflage detection.

To conclude, this research offers significant insights and practical solutions for addressing content evasion in multilingual Natural Language Processing. The novel tool "pyleetspeak" and the robust multilingual NER camouflage detection model effectively enhance content moderation and improve online security. The tool's utility extends beyond its immediate application, indicating its potential in data augmentation for AI systems and future expansion to other languages and evasion strategies.

This summary encapsulates the key findings from [18] research paper, highlighting the development and utilization of a synthetic multilingual dataset and the Python package "pyleetspeak" for addressing the issue of content evasion in social networks. The original article presents more in-depth insights and discusses the broader impacts of word camouflage on content moderation. This research represents a significant stride towards combating information disorders on social networks and provides a solid foundation for future research in this crucial area.

Acknowledgments

This research has been supported by the Spanish Ministry of Science and Education under FightDIS (PID2020-117263GB-I00) and XAI-Disinfodemics (PLEC2021-007681) grants, by Comunidad Autónoma de Madrid under S2018/ TCS-4566 (CYNAMON), by BBVA Foundation grants for scientific research teams SARS-CoV-2 and COVID-19 under the grant: "*CIVIC: Intelligent characterisation of the veracity of the information related to COVID-19*", and by IBERIFIER (Iberian Digital Media Research and Fact-Checking Hub), funded by the European Commission under the call CEF-TC-2020-2, grant number 2020-EU-IA-0252. Finally, David Camacho has been supported by the Comunidad Autónoma de Madrid under "Convenio Plurianual with

the Universidad Politécnica de Madrid in the actuation line of *Programa de Excelencia para el Profesorado Universitario*"

References

[1] F. Fagan, Optimal social media content moderation and platform immunities, European Journal of Law and Economics 50 (2020) 437–449. doi:10.1007/s10657-020-09653-7.

[2] F. Sharevski, R. Alsaadi, P. Jachim, E. Pieroni, Misinformation warnings: Twitter's soft moderation effects on covid-19 vaccine belief echoes, Computers & Security 114 (2022) 102577. doi:https://doi.org/10.1016/j.cose.2021.102577.

[3] Y. Gerrard, Beyond the hashtag: Circumventing content moderation on social media, New Media & Society 20 (2018) 4492–4511. doi:10.1177/1461444818776611.

[4] A. Martín, J. Huertas-Tato, Á. Huertas-García, G. Villar-Rodríguez, D. Camacho, FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference, Knowledge-Based Systems 251 (2022) 109265. doi:10.1016/j.knosys.2022.109265.

[5] M. Kavanagh, Bridge the generation gap by decoding leetspeak, Inside the Internet 12 (2005) 11.

[6] A. Romero-Vicente, Word camouflage to evade content moderation, 2021. URL: https://www.disinfo.eu/publications/word-camouflage-to-evade-content-moderation/.

[7] K. Blashki, S. Nichol, Game geek's goss: linguistic creativity in young males within an online university forum, 2005.

[8] J. Fuchs, Gamespeak for n00bs - a linguistic and pragmatic analysis of gamers' language, Ph.D. thesis, University of Graz, 2013. URL: https://unipub.uni-graz.at/obvugrhs/content/titleinfo/231890?lang=en.

[9] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218.

[10] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, arXiv preprint (2020). doi:arXiv:2004.09813.

[11] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia, 2019. arXiv:1907.05791.

[12] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, 2020. arXiv:2004.09297.

[13] Á. Huertas-García, J. Huertas-Tato, A. Martín García, D. Camacho, Countering Misinformation Through Semantic-Aware Multilingual Models, in: Intelligent Data Engineering and Automated Learning – IDEAL 2021, Springer International Publishing, 2021, pp. 312–323. doi:10.1007/978-3-030-91608-4_31.

[14] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, 2022. doi:10.48550/ARXIV.2211.01786.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave,

M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, 2019. doi:10.48550/ARXIV.1911.02116.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[17] Z. Papakipos, J. Bitton, Augly: Data augmentations for robustness, 2022. arXiv:2201.06494.

[18] Álvaro Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage, Applied Soft Computing 145 (2023) 110552. doi:https://doi.org/10.1016/j.asoc.2023.110552.