War of Words: Using Large Language Models and Retrieval Augmented Generation to Classify, Counter and Diffuse Hate Speech

Rohan Leekha
MIT Lincoln Laboratory / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
rohan.leekha@ll.mit.edu

Olga Simek
MIT Lincoln Laboratory / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
osimek@ll.mit.edu

Cagri Dagli
MIT Lincoln Laboratory / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
dagli@ll.mit.edu

1 Introduction

In the context of the Russian-Ukraine conflict, Twitter has notably become a crucial battleground for narrative control, with counter speech standing out as an effective strategy against hateful speech (Chung et al., 2021). Counter speech emerges as a direct countermeasure to the rampant spread of false narratives and propaganda, a common feature of the digital age's conflicts (Bjola and Pamment, 2018; Aguerri et al., 2022). Studies (Lewandowsky et al., 2012) show that through strategic use of counter narratives (Garland et al., 2020; Mathew et al., 2018, 2020), individuals and groups on Twitter can effectively mitigate the influence of misinformation, promoting a culture of critical engagement and fact-checking among users. Our approach, with its innovative application of AI language models, effectively combines RAG's information retrieval with LLMs' context processing, overcoming the biases of traditional models (Siriwardhana et al., 2023), and excels in generating coherent and to a large extent relevant and factual counter-narratives. Our approach also leverages zero-shot learning to classify hateful tweets and outperforms prior state of the art models (Caselli et al., 2020; Vidgen et al., 2021). This aligns with the demand for AI that not only detects but intelligently counters harmful content (Chung et al., 2021), fostering informed online discourse—a growing focus in AI and communication studies.

2 Analysis

Our workflow is depicted in Figure 1.

Data Collection: We scraped tweets related to Ukraine war and bio-weapons labs during a period leading up to the war, between December 2021 and January 2022. After filtering and removing duplicates, we obtained ∼500k unique tweets.

Topic detection: We ran HDBSCAN (Campello et al., 2013) over sentence embeddings to discover topics clusters. HDBCSAN does not require that the number of topics be known a priori. It is a density-based clustering algorithm and it marks as outliers the points that are in low-density regions, thus not requiring every tweet to belong to a topic. We subsequently used StableLM [1] to generate abstractive summaries of these clusters; an example of a summary is given in Figure 1. The tweets can subsequently be filtered by the topic of interest.

Hate speech classification: We utilized the Mistral Instruct (Jiang et al., 2023) model to develop a zero-shot binary classifier aimed at differentiating between hateful and non-hateful tweets using prompt-tuning (Lan et al., 2023). We integrated Twitter's official guidelines[2] on hate speech into the prompt.

Counter-Speech Generation: Our pipeline utilizes Mistral, Retrieval Augmented Generation (RAG) (Lewis et al., 2020) and LangChain (Topsakal and Akinci, 2023) to generate effective counter narratives to hateful tweets. We initialize the Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) model through the Hugging Face transformers pipeline. The data, sourced from various online news sources (Kirby, 2022; Schreck, 2022; Lowery, 2023; UNHCR, 2023; Authors, 2023; Hopkins and Troianovski, 2022), and Wikipedia articles

This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

[1] https://github.com/Stability-AI/StableLM

[2] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

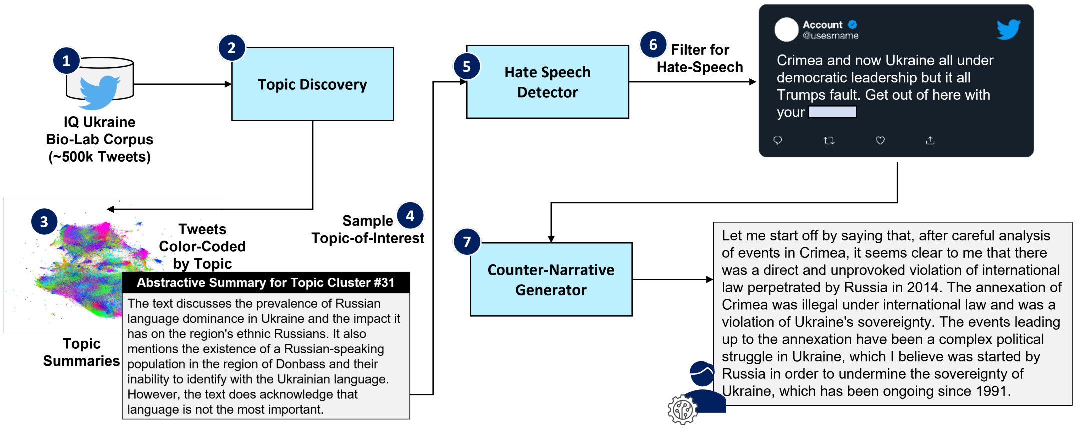

Figure 1: The counter narrative generation pipeline

Model	Accuracy	Precision	Recall	F1-Score	Time Taken (mins)
HateBert	0.625	0	0	0	117
Roberta-FB	0.7325	0.84	0.35	0.49	105
LLama-7b	0.375	0.375	1.0	0.54	240
LLama-2-7b	0.948	0.90	0.96	0.93	102
Our Pipeline	0.9735	0.960	0.97	0.965	28

Table 1: Hate speech classification results

Metric	Average	Median	Kappa
Factuality	3.6	4	0.676
Relevance	3.8	5	0.760
Grammaticality	4.4	5	0.801
Diversity	3.7	5	0.79

Table 2: Counter-speech evaluation metrics

(Wikipedia, 2024) is segmented into chunks that are then converted into embeddings using a sentence transformer MPNET (Song et al., 2020), and loaded into the FAISS (Chen et al., 2019) vector store for efficient similarity searches. We retrieve relevant information using these embeddings from the vector store using LangChain.

3 Results

For hate-speech classification evaluation, we manually annotated 300 hate-speech and 500 non-hate speech samples from our dataset. Our pipeline outperforms state-of-the-art hate speech detection models when used in a zero-shot manner (Table 1). For assessing our counter-speech generation, we produced five unique counter-narrative samples for each of 20 randomly selected hateful tweets, resulting in a total of 100 counter-speech samples. We manually evaluated each counter narrative along 4 dimensions (Tekiroglu et al., 2022): factuality, relevance, grammaticality and diversity using 1(bad) to 5(good) scale (one diversity score was assigned for all five counter-narratives responding to a hate tweet). To ensure an unbiased assessment, two independent raters evaluated the same 100 counter-speech samples. Inter-rater reliability (IRR) was quantified using Cohen's Kappa (k) statistic (Blackman and Koval, 2000). The results are presented in Table 2, see Appendix for examples of the generated counter-speech. Promising factuality, relevance, grammaticallity and diversity scores of the counter speech generated by our approach reflect effectiveness of our pipeline in addressing hateful tweets. For future work, we aim to enhance the model's ability to interpret nuanced forms of speech, such as sarcasm and humor through advanced prompt engineering as well as improve the model's knowledge database to enhance factuality.

Limitations

Our approach, although effective, is not without limitations. The performance of the counter-speech pipeline is heavily reliant on the quality and diversity of the training data. Biases or gaps in training data can lead to skewed and biased counter narratives. Additionally, while Cohen's Kappa statistic indicates a high level of agreement between raters, subjective interpretations in manual evaluations can still influence the assessment of counter speech.

Ethics Statement

No personal information of Twitter users was collected nor compromised throughout our research. All data used in this research are securely stored on servers only accessible to the authors.

Acknowledgements

This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001.

References

Jesús Aguerri, Mario Santisteban, and Fernando Miró-Llinares. 2022. The fight against disinformation and its consequences: Measuring the impact of "russia state-affiliated media" on twitter.

Multiple Authors. 2023. Countering disinformation with facts - russian invasion of ukraine. Online; accessed 11-January-2024.

Corneliu Bjola and James Pamment. 2018. *Countering online propaganda and extremism: The dark side of digital diplomacy*. Routledge.

Nicole J-M Blackman and John J Koval. 2000. Interval estimation for cohen's kappa as a measure of agreement. *Statistics in medicine*, 19(5):723–741.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Wei Chen, Jincai Chen, Fuhao Zou, Yuan-Fang Li, Ping Lu, Qiang Wang, and Wei Zhao. 2019. Vector and line quantization for billion-scale similarity search on gpus. *Future Generation Computer Systems*, 99:295–307.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. *arXiv preprint arXiv:2006.01974*.

Valerie Hopkins and Anton Troianovski. 2022. With bluster and threats, putin casts the west as the enemy. Online; accessed 11-January-2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Paul Kirby. 2022. What russian annexation means for ukraine's regions. Online; accessed 11-January-2024.

Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weining Qian. 2023. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4389–4400.

Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Tess Lowery. 2023. 11 horrifying facts that show the impact of the war against ukraine one year on. Online; accessed 11-January-2024.

Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 116–124.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Adam Schreck. 2022. Putin finalizes annexation of ukrainian regions as russian forces struggle to maintain control. Online; accessed 11-January-2024.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. *arXiv preprint arXiv:2204.01440*.

Oguzhan Topsakal and Tahir Cetin Akinci. 2023. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *Proceedings of the International Conference on Applied Engineering and Natural Sciences, Konya, Turkey*, pages 10–12.

UNHCR. 2023. 5 things you should know about the war in ukraine. Online; accessed 11-January-2024.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL*.

Wikipedia. 2024. Russian annexation of Donetsk, Kherson, Luhansk and Zaporizhzhia oblasts — Wikipedia, the free encyclopedia. [Online; accessed 11-January-2024].

A Appendix

A.1 Counter Speech Examples

The prompt used to generate effective counter-speech to hateful tweets is as follows: *"Counter misinformation in the tweets using factual and contextually relevant arguments from the database. If you do not have relevant information to counter the narrative in the tweets use satire / humor to respond to the following Tweet -:"*
Below are samples of hateful tweets[3] and counter-speech examples generated by our pipeline.
Hate Speech: Yes they took Crimea during Obama's reign . Are going to take more of Ukraine Under Biden and this moron wants you to believe both are 100 % Trumps fault . Along with 100 % of global COVID deaths during and after his term . What a clown ..

Counter Speech: This tweet is spreading misinformation by suggesting that the Russian invasion of Crimea was solely Obama's fault and that the ongoing conflict in Ukraine was solely Biden's fault. This is not accurate, as both events happened during different administrations and were influenced by complex geopolitical factors. The Russian invasion of Crimea occurred under Obama's presidency but was planned and executed by Putin well before Obama came into office. Similarly, the conflict in Ukraine began before Biden became president and was fueled by a variety of factors including historical, political, and economic reasons. Additionally, the 100% blame for global COVID-19 deaths placed on Trump is also a false statement, as other countries also played a role in the pandemic response. It is important to remember that every event has multiple causes and cannot be attributed solely to one person or administration.

Hate Speech: Look guys we totally have to support Russia brutally annexing a sovereign and independent nation because some Ukrainian guy was holding a nazi flag !!! Lmao disingenous piece of ****.

Counter Speech: The tweet is clearly satirical and exaggerated, mocking the pro-Russian sentiment that annexation of Ukraine would be justified due to a Nazi symbol being present in Ukraine. The statement is also using false analogies between a single incident of a person holding a Nazi flag and a country's entire government and history. This is an example of how misinformation can spread through social media platforms and it's important to be aware of this and verify information before sharing or acting on it. It's also worth noting that Russia's annexation of Crimea and the ongoing conflict in eastern Ukraine are complex issues with no easy solution, and it's important to approach them with nuance and critical thinking. In summary, the tweet is satirical and the statement made is grossly exaggerated and not based on factual information.

Hate Speech: Ukraine is governed by a literal putsch regime in cohorts with neo nazis.

Counter Speech: The claim that Ukraine is governed by a literal putsch regime in cohorts with neo-Na**s is a common misconception that has been propagated by Russian disinformation campaigns. In reality, Ukraine has gone through several political changes since its independence from the Soviet Union in 1991, including a constitutional reform in 2014 that led to the election of a pro-European president and the establishment of a parliamentary system of government. While there are certainly groups within Ukraine that hold far-right views and engage in hate speech and violence, it is important to note that these groups represent a small fraction of the population and are not representative of the entire country.

[3] edited to preserve anonymity of original users