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

Rohan Leekha, Olga Simek and Charlie Dagli - Massachusetts Institute of Technology Lincoln Laboratory

Introduction

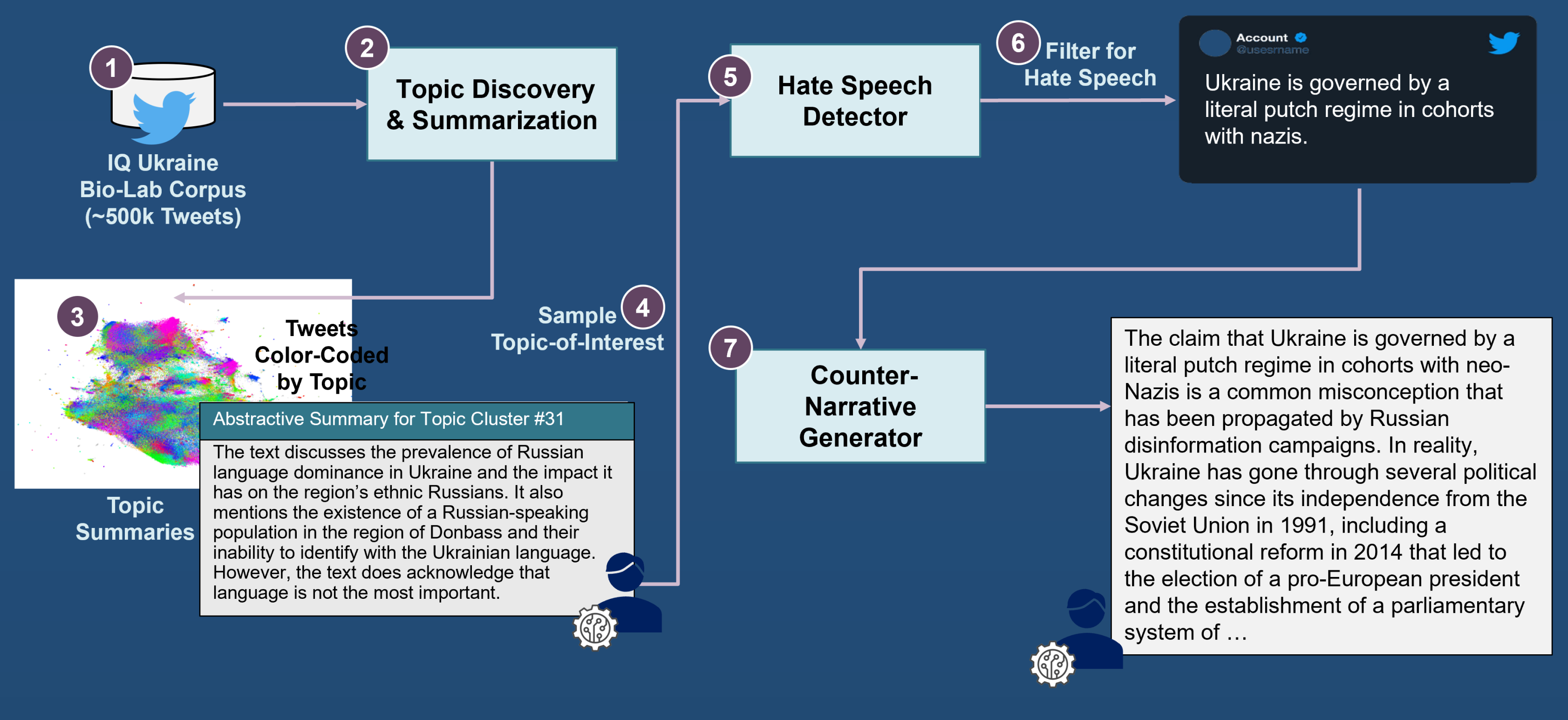
MOTIVATION

There is a need for a robust and effective defense strategy against malign influence spread by influence operation campaigns perpetrated by foreign state actors and other entities

PROBLEM

- Social media and, more recently, generative AI make it very easy to generate disinformation and hateful content at scale
- Monitoring via traditional topic/narrative modeling often focuses on low-level content that can be difficult to interpret
- Highly manual workflows typically used for content prioritization and response generation do not scale

Approach



Abstractive Summary for Topic Cluster #31

The text discusses the prevalence of Russian language dominance in Ukraine and the impact it has on the region's ethnic Russians. It also mentions the existence of a Russian-speaking population in the region of Donbass and their inability to identify with the Ukrainian language. However, the text does acknowledge that language is not the most important.

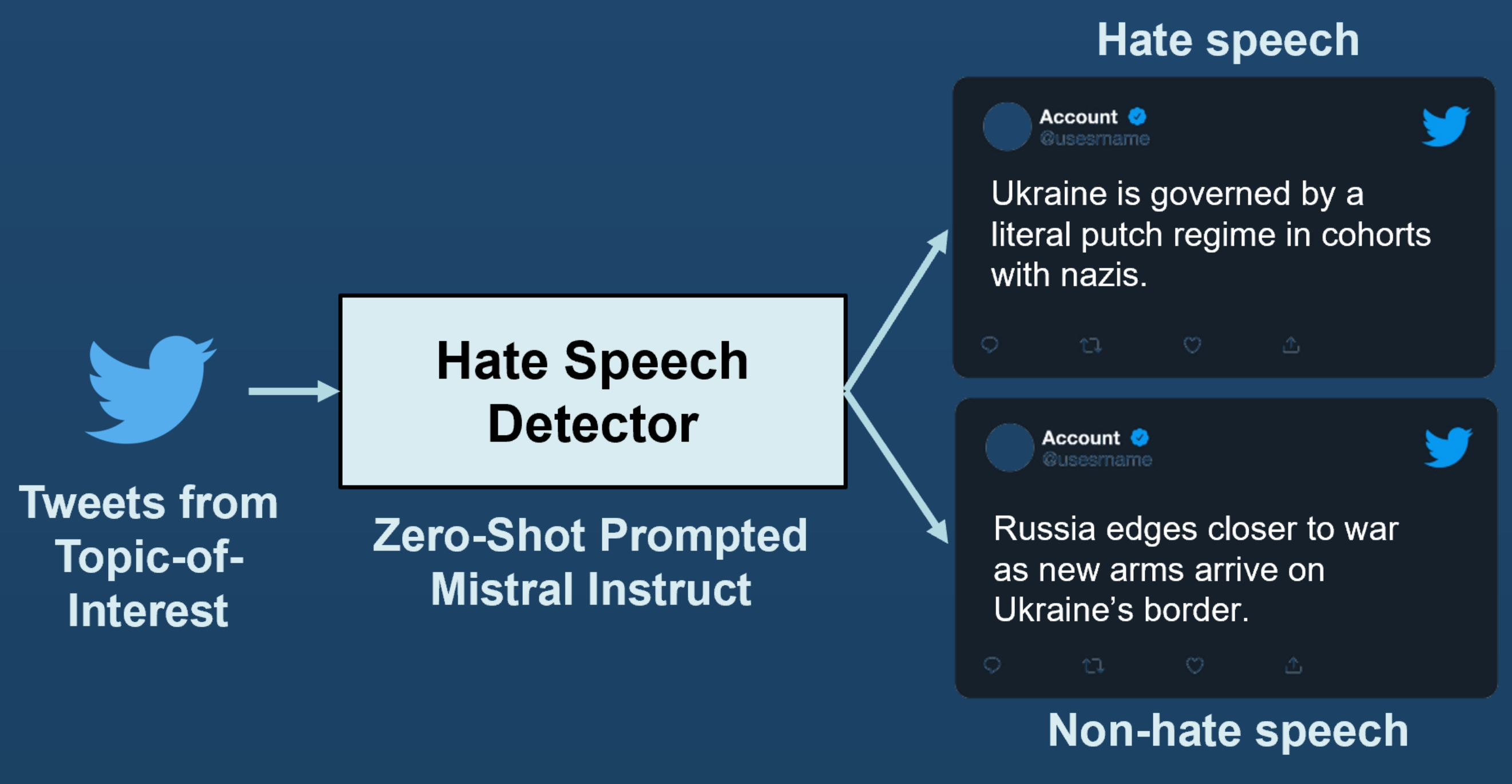
Ukraine is governed by a literal putch regime in cohorts with nazis.

The claim that Ukraine is governed by a literal putch regime in cohorts with neo-Nazis 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 …

Counter-Narrative Generation Prototype

CLASSIFICATION OF SOCIAL MEDIA POSTS CONTAINING INFLAMATORY SPEECH

- Identify extreme content via hate-speech classifier and rank-order tweets in topic clusters
- Zero-shot prompt-based prediction using Mistral Instruct & Twitter's hate speech guidelines
- ~1k hand-labeled evaluation set
- Compare to pre-trained HateBERT1, Fine-Tuned RoBERTa2, prompted LLaMA-7b/LLaMA2-7b



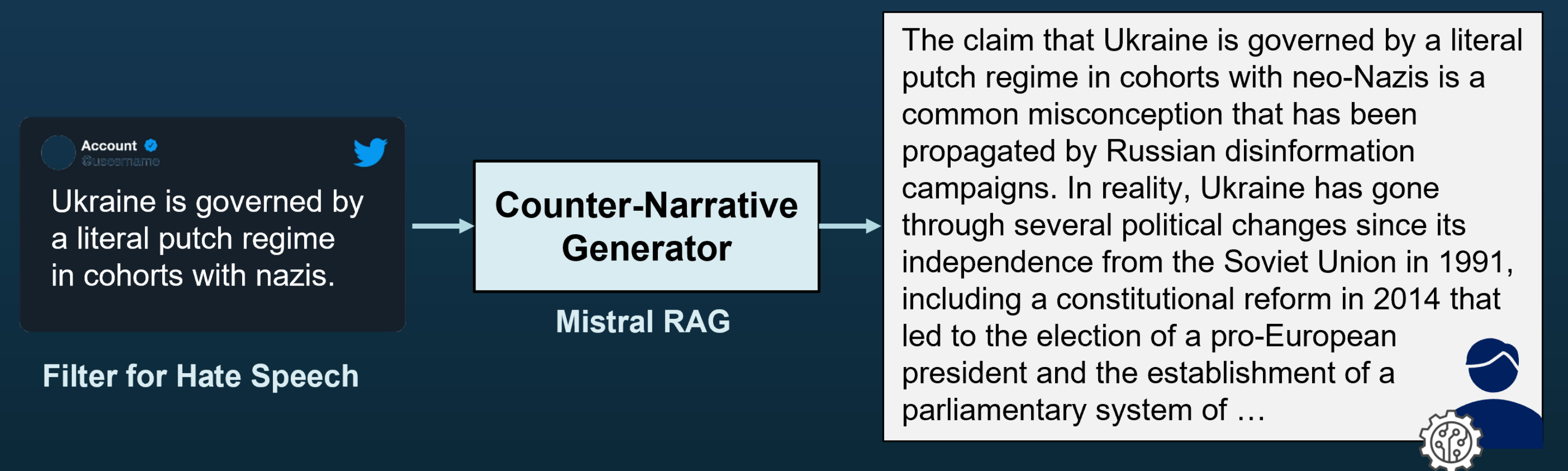
Ukraine is governed by a literal putch regime in cohorts with nazis.

Russia edges closer to war as new arms arrive on Ukraine's border.

Hate Speech Classification Pipeline

AUTOMATED RESPONSE TO INFLAMATORY SPEECH

- Retrieval Augmented Generation (RAG) using relevant factual context (Wikipedia, News Articles) to ground auto-generated response (via FAISS over MPNet paragraph embeddings)
- Mistral-7B prompted to make use of factual arguments utilizing the RAG database to generate contextually relevant and correct counter-speech
- Human evaluation on Likert scale of 5 generations for each of 20 random hate speech tweets; 2 raters, 5 quality dimensions



Ukraine is governed by a literal putch regime in cohorts with nazis.

The claim that Ukraine is governed by a literal putch regime in cohorts with neo-Nazis 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 …

Counter-Narrative Generation Pipeline

Results

DATASET AND METRICS FOR EVALUATING HATE SPEECH DETECTION

- Scraped tweets related to Ukraine war and bio-weapons labs during a period leading up to the war
- Time period: between December 2021 and January 2022
- After filtering and removing duplicates, dataset contained ~500k unique tweets

Model	Precision (⇧)	Recall (⇧)	F1 score (⇧)	Accuracy (⇧)	Time to run in mins (⇩)
HateBERT Pre-Trained	0.00	0.00	0.00	0.63	117
RoBERTa Fine-Tuned	0.84	0.35	0.49	0.73	105
LLaMA-7B Zero-Shot	0.38	**1.00**	0.54	0.38	240
LLaMA2-7B Zero-Shot	0.90	0.96	0.93	0.95	102
Mistral-Instruct Prompted	**0.96**	0.97	**0.97**	**0.97**	**28**

Hate Speech Detection Evaluation (BioLab Eval Set)

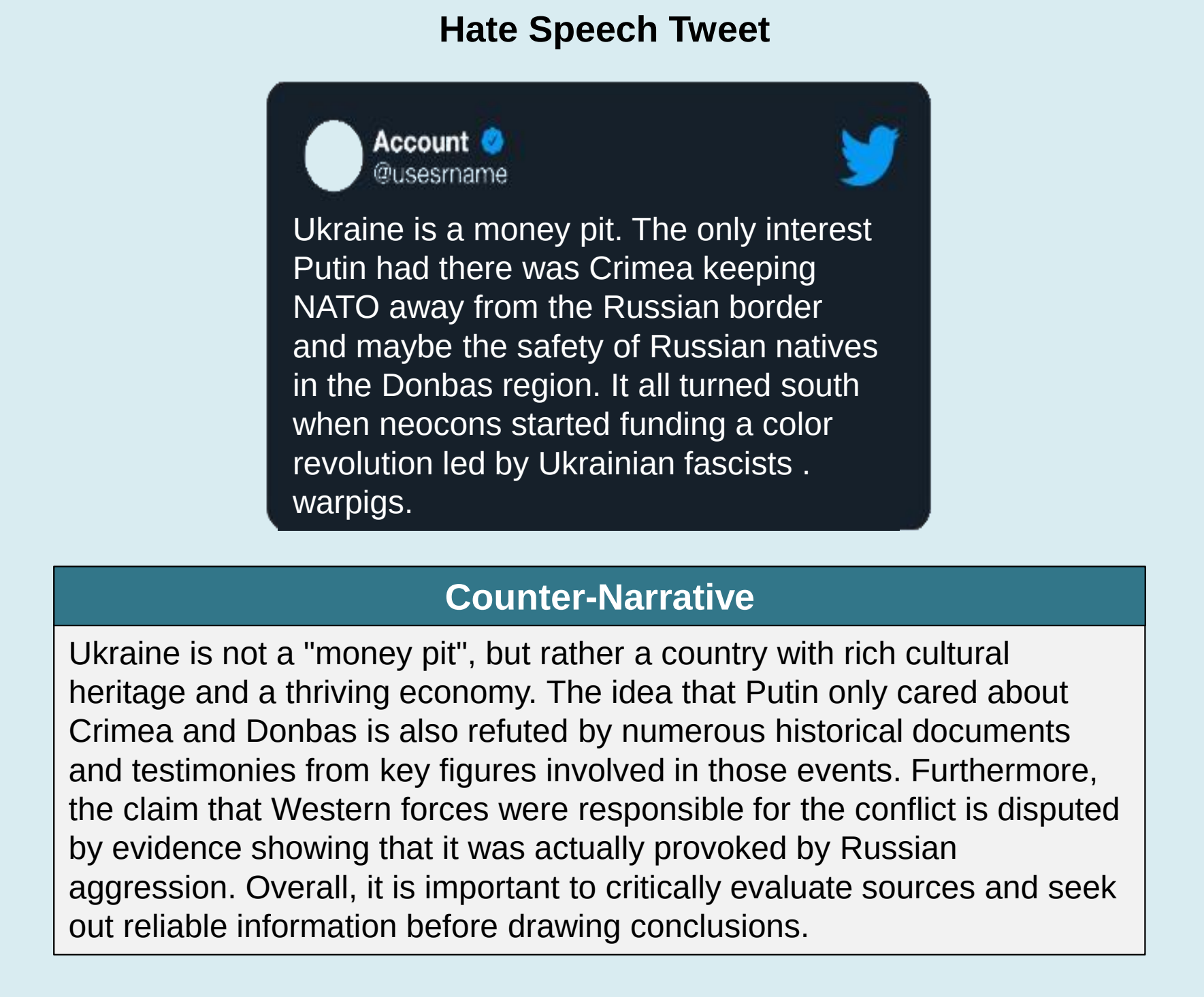
METRICS FOR EVALUATING COUNTER-SPEECH GENERATION

- Produced five unique counter-narrative samples for each of 20 randomly selected hateful tweets, resulting in a total of 100 counter-speech samples
- Manually evaluated each counter narrative along 4 dimensions: factuality, relevance, grammaticality and diversity using 1(bad) to 5(good) scale
- Note: 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 was quantified using Cohen's Kappa statistic

Metric	Mean (⇧)	Median (⇧)	Cohen Kappa (⇧)
Factuality	3.6	4	0.68
Relevance	3.8	5	0.76
Grammaticality	4.4	5	0.80
Diversity	3.7	5	0.79

Counter-Speech Evaluation Metrics

EXAMPLES OF AUTOMATICALLY GENERATED COUNTER-NARRATIVES

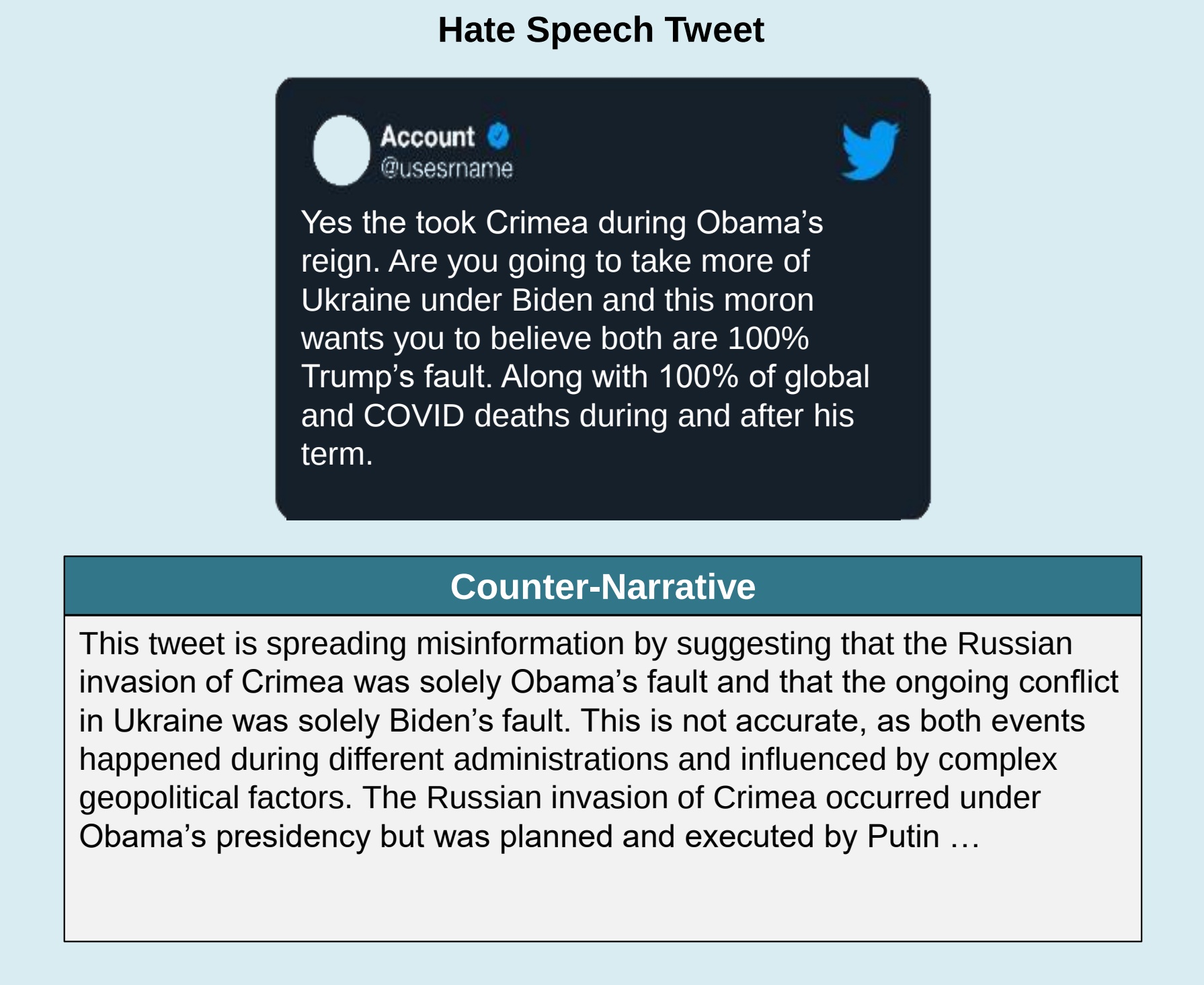

Hate Speech Tweet

Account
@usesrname

Ukraine is a money pit. The only interest Putin had there was Crimea keeping NATO away from the Russian border and maybe the safety of Russian natives in the Donbas region. It all turned south when neocons started funding a color revolution led by Ukrainian fascists . warpigs.

Counter-Narrative

Ukraine is not a "money pit", but rather a country with rich cultural heritage and a thriving economy. The idea that Putin only cared about Crimea and Donbas is also refuted by numerous historical documents and testimonies from key figures involved in those events. Furthermore, the claim that Western forces were responsible for the conflict is disputed by evidence showing that it was actually provoked by Russian aggression. Overall, it is important to critically evaluate sources and seek out reliable information before drawing conclusions.


Hate Speech Tweet

Account
@usesrname

Yes the took Crimea during Obama's reign. Are you going to take more of Ukraine under Biden and this moron wants you to believe both are 100% Trump's fault. Along with 100% of global and COVID deaths during and after his term.

Counter-Narrative

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 influenced by complex geopolitical factors. The Russian invasion of Crimea occurred under Obama's presidency but was planned and executed by Putin …

Conclusion

- Our research demonstrates the effectiveness of zero-shot learning and LLM prompt engineering for nuanced hate speech classification surpassing prior state-of-the-art models
- Our approach effectively combines RAG's information retrieval with LLMs' context processing, overcoming the biases of traditional models and excels in generating coherent and to a large extent relevant and factual counter-narratives

FUTURE WORK

- Reduce LLM-generated hallucinations (i.e., untruthful information and claims) by expanding RAG database and other approaches
- Develop effective techniques for prompt engineering in order to improve relevance of the generated counter-narratives
- Investigate Langchain's capability to incorporate chat history iterations in order to generate conversational response

DISTRIBUTION STATEMENT A. Approved for public re-lease. Distribution is unlimited. This material is based upon work supported by the Department of Defense 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 Defense.