An Entity-Aware Approach to Logical Fallacy Detection in Kremlin Social Media Content

Benjamin Shultz
Independent Resarcher
Washington, D.C., United States
0009-0008-0354-1917

***Abstract*—Logical fallacy detection has emerged as a novel and challenging task for language models, more complex than traditional fake news or hate speech detection. This research-in-progress examines an Entity-Aware Approach for logical fallacy detection adapted for a timely use case of Kremlin social media content. As part of this study, a curated dataset of tweets about the war in Ukraine published by Russian government accounts, *RuFal*, is introduced, on which the Entity-Aware Approach is tested. Preliminary results show the Entity-Aware Approach outperforms baseline pre-trained language models by at least 0.83% on the domain non-specific LOGIC dataset and when both directly transferred to and trained on the domain specific *RuFal* dataset, by at least 3.09% and 0.45%, respectively, showing the Entity-Aware Approach warrants further research.**

Keywords—Disinformation, Logical fallacy detection, Named-entity recognition, Russia, Social media

I. Introduction

Computational research around mis-, dis- and mal-information (MDM) has largely centered around the tasks of fake news detection and hate speech detection [1]. Through these tasks, numerous machine learning methods and models have been developed to classify text on a binary continuum, of "Real" or "Fake", or "Hate Speech" or "Not Hate Speech" [2]. MDM, however, is rarely this black and white and often presents in diverse ways. Logical fallacy detection (LFD) is a novel natural language inference task that has emerged, which goes beyond binary classification, attempting to discover the presence of *many* false or misleading communication strategies in a given set of content. LFD is a particularly challenging area of MDM research, as fallacious statements often contain true information, albeit presented in such a way as to deceptively make a point. Logical fallacies have long been researched in the philosophical realm, and largely encompass two types—formal and informal [10,11]. Formal fallacies typically contain a valid argument, but are presented in an invalid way or on the assumption of given conditions, whereas informal fallacies typically contain both an invalid argument and an invalid presentation. Language models have struggled to reliably detect informal logical fallacies, particularly in any given domain-specific manner.

Existing studies have primarily examined LFD in a domain-nonspecific manner, albeit with some research into issues such as climate change [3,4]. Additionally, some studies have considered the task of propaganda detection, sharing some overlap with LFD, examining the Communist Party of the People's Republic of China [6] and far-right news outlets [7]. These studies implement a variety of methods, including structure-aware classifiers [3], case-based reasoning [4] and fine-grained text analysis [5,7]. However, none specifically examine the role that named entities play in effecting LFD—a potentially significant gap considering the use case of Kremlin social media content. Moreover, prior works indicate there is a continuing need for reliable models that can detect fallacious arguments and help to mitigate the spread of MDM surrounding emerging geopolitical conflicts. Such conflicts, for instance, the full-scale Russian invasion of Ukraine, have taken on new forms as the boundaries between social media and information warfare have blurred. A hallmark of Russia's full-scale invasion has been its visual presence across social media. To this point, the Russian government has persistently used its official social media channels to spread false and misleading content about its war [8,9], including the use of fallacious arguments to distort researchable and documented facts for political gain.

This study seeks to examine this domain-specific challenge through the following research questions:

- How does the presence of named entities in content affect the performance of LFD models? Does the generalization of such entities contribute to stronger model performance or generalizability?

In addressing these research questions with a novel Entity-Aware Approach (EAA), this study makes two primary contributions to the broader canon of LFD work:

- This study presents *RuFal*, a novel dataset containing 700 fallacious English-language Kremlin tweets, manually annotated with 13 common logical fallacy types (examples entries are found in Appendix II).
- The EAA identifies and replaces named entities with pre-defined labels, and in preliminary results, outperforms baseline pre-trained models for both domain-specific and nonspecific LFD.

II. Data & Methods

In this section, evaluation data, baseline language models used for comparison and the specific use case of Kremlin social media content are detailed. As well, the importance of considering named entities in LFD is expanded upon.

A. Data

This study leverages two primary datasets. First, five pre-trained language models, as well as one with the EAA (a total of six models) are further trained on the LOGIC dataset [3] for baseline, domain non-specific LFD. Consisting of 2,449 manually annotated samples of 13 logical fallacy types (see

ASONAM '23, November 6 – 9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11...$15.00
http://doi.org/10.1145/3625007.3627988

Appendix I for a full list of definitions), LOGIC is a gold-standard dataset for LFD. Second, to test the EAA for the use case of Kremlin social media content, a novel LFD dataset, *RuFal*, is presented. It contains 700 English-language tweets posted by Russian government accounts between September 2022—March 2023, manually annotated with the same 13 logical fallacy types as LOGIC. All six models are directly transferred and further trained onto *RuFal* for analysis. Tweets were extracted from the Ukraine Conflict Twitter Dataset [12], an open-source dataset hosted on Kaggle that contains more than 10M tweets about Russia's war in Ukraine. *RuFal* is split into 595 train samples and 105 test samples. Charts detailing both the descriptive statistics and count of the fallacy types in *RuFal* can be found below in Tables 1 and 2. In emphasizing the novelty of this analysis, to the best of the author's knowledge, *RuFal* is the first publicly released LFD dataset specific to the domain of Kremlin social media content.

TABLE I. RUFAL DESCRIPTIVE STATISTICS

	Samples	***Tokens***	***Vocab***
Total	700	44,677	5,543
Train	595	24,128	4,190
Test	105	20,549	1,353

TABLE II. RUFAL LOGICAL FALLACY TYPE COUNT

Fallacy Type	***Count***
Intentional Fallacy	211
Appeal to Emotion	113
Fallacy of Extension	67
Ad hominem	58
Fallacy of Relevance	52
Faulty Generalization	44
Ad populum	40
False Causality	33
Fallacy of Credibility	27
False Dilemma	26
Equivocation	13
Circular Claim	10
Deductive Fallacy	6

B. Leveraging Pre-trained Language Models

Five different language models are used to test the EAA on LOGIC and *RuFal*: BERT cased and uncased [13], Electra [14], ALBERT [15] and DeBERTa [16]. First, an initial round of training and testing on LOGIC is conducted using the Transformers package, finding DeBERTa to produce the best F1 score. Weighted Precision and Recall, in addition to F1, are reported. Then, separately, the EAA is implemented with DeBERTa and trained and tested with LOGIC. These steps are repeated for both further training and a direct transfer to *RuFal*. Standard parameters, including the Adam Optimizer and a learning rate of 2^{e-5} are used for all models, and all models were run for three epochs on each dataset.

C. The Impact of Entities in Kremlin Social Media Content

This study posits that the presence of named entities in a corpus can directly impact the ability of language models to reliably engage in LFD. In a normal embedding process, named entity descriptors will be assigned their own unique embedding. For instance, in the case of Kremlin social media content, "Lavrov", "S.Lavrov" and "Sergey Lavrov" will be embedded differently from one another. This is despite all three descriptors representing the same Russian Foreign Ministry official, and being included in content published by many different Kremlin accounts. In the case of LFD, when a model is meant to decipher patterns in a corpus to determine which logical fallacy is present in a given piece of content, an oversized number of unique embeddings—one representing each different descriptor—may create unnecessary confusion for the model, making it less likely to return the correct label [18]. Prior works [3,4] expand on the importance of argument structures and case-based reasoning in LFD, however, the EAA simplifies what we find to be complex approaches to content generalization while still achieving solid performance.

Through the EAA results, shown below in Fig. 1, this challenge is accounted for by replacing four basic categories of named entities (people, organizations, locations and miscellaneous entities) with an entity type label as identified by FlairNLP's zero-shot, pre-trained named entity recognition (NER) package.

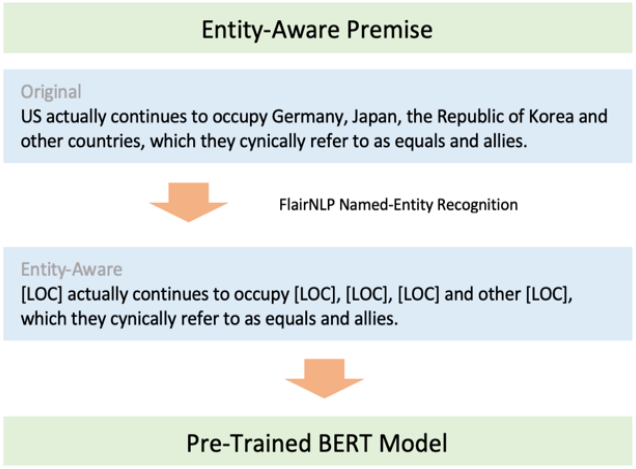

Fig. 1. This Entity-Aware Approach uses FlairNLP to replace entities in the corpus with labels to make more identifiable the key parts of domain-specific fallacious arguments. An example of a Kremlin tweet is shown. This approach may better enable a pre-trained language model to analyze text data for LFD (see Table 3 for model performance on the *RuFal* dataset).

III. PREMLINARY RESULTS & DISCUSSION

Preliminary results show the EAA in combination with the DeBERTa base model to outperform the baseline DeBERTa model, and others tested, by at least 0.83% when trained and tested the LOGIC dataset, and by 0.45% when further trained and tested on the *RuFal* dataset. A direct transfer of the model trained on LOGIC, without seeing *RuFal*, was conducted, and DeBERTa with EAA outperformed the baseline DeBERTa by 3.09%. These preliminary results suggest that leveraging zero-shot, pre-trained NER models to pre-process data can improve LFD results, and in further stages of this research, we aim to expand further test cases to include different pre-trained NER models and additional pre-trained language models. Below in Table 3 the results for all six models across the three tests conducted are displayed. The best F1 score in each instance is bolded and Precision and Recall are additionally reported.

From a policymaking perspective, the collective research effort into LFD is in its infancy—plainly seen by the metrics of this study, as well as prior works [3,4,5,6,7]. However, LFD has the potential to better inform policymakers about the MDM narratives they encounter, as well as counter-narratives and responses [17]. Broadly speaking, the ability to connect certain fallacy types with particular arguments or talking points would be beneficial in counter-MDM spaces.

TABLE III. MODEL PERFORMANCE

Pre-trained models trained on LOGIC			
	F1	P	R
BERT-base-cased	46.47	49.11	51.00
BERT-base-uncased	45.67	46.69	50.67
Electra	15.45	11.21	29.00
ALBERT	50.66	50.00	55.00
DeBERTa	60.02	60.19	61.33
DeBERTa *EAA*	**60.85**	61.19	62.67
Direct transfer to RuFal			
	F1	P	R
BERT-base-cased	9.58	9.36	13.33
BERT-base-uncased	6.72	5.60	8.57
Electra	1.58	1.60	5.71
ALBERT	8.02	9.98	9.52
DeBERTa	9.69	10.86	11.43
DeBERTa *EAA*	**12.78**	19.25	15.24
Further training on RuFal			
	F1	P	R
BERT-base-cased	26.58	23.71	32.38
BERT-base-uncased	29.85	25.90	38.10
Electra	12.29	8.00	26.67
ALBERT	31.52	31.19	35.24
DeBERTa	33.41	32.74	38.10
DeBERTa *EAA*	**33.86**	32.43	40.00

For instance, if the content around a given Kremlin narrative is incorporating the "Fallacy of Extension" or an "Appeal to Emotion," policymakers might be ill-advised to respond to such a narrative for fear of amplifying false or misleading content that resonates with its consumers, which emotional content is shown to do more than factual content [19]. Whereas, if a Kremlin narrative leverages the "Fallacy of Credibility", in other words, either falsely stating a person of note has made a particular claim, or falsely ascribing views to a person of note, policymakers may find themselves in a better position to debunk such claims with a response countering such false credibility.

This raises an additional area of overlap between LFD, propaganda detection, sentiment and emotional analysis that would be well-served with future computational studies. Where we believe the EAA possesses the strongest advantage over existing approaches to LFD lies in the fact that social media and microblogging content, in most political contexts, can lack a defined-enough structure to successfully implement a structure-aware analysis or reasoning-based approach. This is seen in Appendix II via the example entries of Kremlin tweets for each logical fallacy type, which vary greatly in terms of length, tone and subject matter. This has posed a challenge in other areas of NLI and NLP, such as semantic similarity analysis, and remains so for LFD. Named entities, however, unlike a defined set of logical fallacy structures, are universal across content regardless of structure; the EAA, therefore, yields promising results for domain-specific, political-oriented LFD.

IV. CONCLUSION & RESEARCH TRAJECTORY

This research, while still in progress, makes two key contributions to the literature through 1) presenting a novel dataset for LFD tailored to the timely use case of Russian government disinformation on social media; and 2) finding an Entity-Aware Approach to LFD to outperform baseline language models in domain-nonspecific, domain-specific and direct transfer instances. While there are limitations on this work, further expansion of the selection of NER models tested—or perhaps ensembling several NER models—will increase both the explainability and robustness of these early results. As will cross-domain analysis on further LFD datasets, such as LOGIC Climate, a companion dataset to LOGIC. We also hope to leverage future resources to continue manually annotating Kremlin tweets to increase the size of the *RuFal* dataset, and eventually expand its inputs to include more than solely tweets. In a time with seemingly continuous advances in machine learning, in combination with a volatile geopolitical situation in continental Europe for the first time in nearly 30 years, this study conducts an interdisciplinary analysis which provides tangible contributions to the literature and policy sphere alike.

REFERENCES

[1] Giachanou, A., & Rosso, P. (2020). The Battle Against Online Harmful Information. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. CIKM '20: The 29th ACM International Conference on Information and Knowledge Management. ACM. https://doi.org/10.1145/3340531.3412169

[2] Sourati, Z., et al. (2023). Robust and explainable identification of logical fallacies in natural language arguments. In Knowledge-Based Systems (Vol. 266, p. 110418). Elsevier BV. https://doi.org/10.1016/j.knosys.2023.110418

[3] Jin, Z., et al. (2022). Logical Fallacy Detection. In Findings of the Association for Computational Linguistics: EMNLP 2022. https://doi.org/10.18653/v1/2022.findings-emnlp.532

[4] Sourati, Z., Ilievski, F., Sandlin, H.-Â., & Mermoud, A. (2023). Case-Based Reasoning with Language Models for Classification of Logical Fallacies (V2). arXiv. https://doi.org/10.48550/ARXIV.2301.11879

[5] Alhindi, T., Chakrabarty, T., Musi, E., & Muresan, S. (2022). Multitask Instruction-based Prompting for Fallacy Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.560

[6] Chang, R.-C., Lai, C.-M., Chang, K.-L., & Lin, C.-H. (2021). Dataset of Propaganda Techniques of the State-Sponsored Information Operation of the People's Republic of China (V1). arXiv. https://doi.org/10.48550/ARXIV.2106.07544

[7] Martino, G. D. S., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-Grained Analysis of Propaganda in News Articles. arXiv. https://doi.org/10.48550/ARXIV.1910.02517

[8] Shultz, B. (2023). In the Spotlight: The Russian Government's Use of Official Twitter Accounts to Influence Discussions About its War in Ukraine. In Proceedings of the 2nd ACM Workshop on Multimedia AI against Disinformation (MAD'23). ICMR '23. ACM. https://doi.org/10.1145/3592572.3592843

[9] Klepper, D. For Russian diplomats, disinformation is part of the job. (2022). Associated Press. apnews.com/article/russia-ukraine-covid-technology-health-business-628cf047adf9fde93c0d7f820e46f8e4.

[10] Korb, K. (2004). Bayesian Informal Logic and Fallacy. In Informal Logic (Vol. 24, Issue 1). University of Windsor Leddy Library. https://doi.org/10.22329/il.v24i1.2132

[11] Sahai, S., Balalau, O., & Horincar, R. (2021). Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.53

[12] Kaggle AI/ML Community. (2022). Ukraine Conflict Twitter Dataset. https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows

[13] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/ARXIV.1810.04805

[14] Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv. https://doi.org/10.48550/ARXIV.2003.10555.

[15] Lan, Z., et al. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (Version 6). arXiv. https://doi.org/10.48550/ARXIV.1909.11942

[16] He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2006.03654

[17] Tumber, H., & Waisbord, S. (Eds.). (2021). The Routledge companion to media disinformation and populism. Routledge.

[18] Alhindi, T., Chakrabarty, T., Musi, E., & Muresan, S. (2022). Multitask Instruction-based Prompting for Fallacy Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.5

[19] Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. In Cognitive Research: Principles and Implications (Vol. 5, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s41235-020-00252-3

APPENDIX I. LOGICAL FALLACY TYPOLOGY & DEFINITIONS[1]

Fallacy Type	*Definition*
Intentional Fallacy	Some intentional (sometimes subconscious) action or choice to incorrectly support an argument.
Appeal to Emotion	Manipulation of the recipient's emotions using loaded or strong language in order to win an argument.
Fallacy of Extension	Attacking an exaggerated or caricatured version of your opponent's position.
Ad hominem	Irrelevantly attacking an entity or some aspect of the entity who is making the argument.
Fallacy of Relevance	An appeal to evidence or examples that are not relevant to the argument at hand; often known as whataboutism.
Faulty Generalization	A conclusion is drawn about all or many instances of a phenomenon on the basis of one or a few instances of it.
Ad populum	An argument which is based on affirming that something is real or better because the majority thinks so.
False Causality	A statement that jumps to a conclusion implying a causal relationship without evidence.
Fallacy of Credibility	An appeal is made to some form of ethics, authority, expertise, or credibility.
False Dilemma	A claim presenting only two options or sides when there are many options or sides.
Equivocation	Likening two ambiguous keywords or phases to each other that lie on either end of an argument.
Circular Claim	When the end of an argument comes back to the beginning without proving itself.
Deductive Fallacy	An error in the logical structure of an argument.

[1] Fallacy typology and definitions largely borrowed from Jin et al. [3] and Sourati et al. [4].

APPENDIX II. RUFAL EXAMPLE ENTRIES[1]

Fallacy Type	*Example Entry*
Intentional Fallacy	A patriotic event "Crimea🇷🇺 is in my heart" was held in Yalta!
Appeal to Emotion	For us, this is primarily about fighting for our people who live in these territories. You see, we are a multi-ethnic country; this is the Russian world after all.
Fallacy of Extension	An unprecedented sanctions aggression has been launched against Russia. It was aimed at crushing our economy, wrecking our national currency by stealing our currency reserves & provoking a devastating inflation in a short span of time. This plan has fallen through.
Ad hominem	Claims by @JamesCleverly and @trussliz that the West has supposedly never threatened Russia or impinged upon Russian territorial integrity should be taken with a significant pinch of salt. The truth is just the opposite.
Fallacy of Relevance	On March 24, 1999, NATO forces led by the USA in gross violation of the UN Charter began barbaric carpet bombing of Yugoslavia.
Faulty Generalization	'We're gonna f**king kill you all' – says the "brave" AFU "liberator" in Kherson after trashing a shop belonging to someone who he thinks to be a 'collaborator'. Regrettably, this is just the tip of the iceberg which the West stubbornly refuses to notice.
Ad populum	The special military operation in Ukraine is aimed at ensuring security not only of Russia, but the whole world.
False Causality	The purpose of this operation is to protect people who, for eight years now, have been facing humiliation and genocide perpetrated by the Kiev regime.
Fallacy of Credibility	The inconvenient truth about the pervasive nature of Nazism in Ukraine, ignored by the MSM, was voiced by ex-US military John McIntyre, who as a UAF volunteer witnessed the neo-Nazis, crimes & hatred first-hand.
False Dilemma	The arms black market operating in Ukraine creates serious challenges. Cross-border criminal groups smuggle these arms to other regions. There is a persistent risk of criminals getting hold of powerful weapons, incl portable air defence systems & precision weapons.
Equivocation	Sneaky double standards by UK: Call for sanctions against Russia and embargo on Russian oil, leading to oil price rise, along with claims to stop buying Russian oil. Yet at the same time secretly buy Russian oil from third countries. All of this at the expense of UK citizens.
Circular Claim	The Kiev regime, guided by US, blocked the peace talks it itself had initiated. Russia has always stood for dialogue but Ukraine had legislatively blocked the negotiation process back in September 2022.
Deductive Fallacy	Responsibility for inciting and escalating Ukraine conflict and sheer number of casualties lies entirely with the Western elites and today's Kiev regime, which is serving not national interests but interests of third countries.

[1] The full *RuFal* dataset can be accessed on the Hugging Face Hub under RuFal_fallacy_detection.