Paying Attention to Deflections: Mining Pragmatic Nuances for Whataboutism Detection in Online Discourse

Khiem Phi **Noushin Salek Faramarzi** **Chenlu Wang** **Ritwik Banerjee**
Department of Computer Science
Stony Brook University, New York, USA
{kphi, nsalekfarama, chenlwang, rbanerjee}@cs.stonybrook.edu

Abstract

Whataboutism, a potent tool for disrupting narratives and sowing distrust, remains underexplored in quantitative NLP research. Moreover, past work has not distinguished its use as a strategy for misinformation and propaganda from its use as a tool for pragmatic and semantic framing. We introduce new datasets from Twitter[1] and YouTube, revealing overlaps as well as distinctions between whataboutism, propaganda, and the *tu quoque* fallacy. Furthermore, drawing on recent work in linguistic semantics, we differentiate the 'what about' lexical construct from whataboutism. Our experiments bring to light unique challenges in its accurate detection, prompting the introduction of a novel method using attention weights for negative sample mining. We report significant improvements of **4%** and **10%** over previous state-of-the-art methods in our Twitter and YouTube collections, respectively.[2]

1 Introduction

Whataboutism is the practice of deflecting criticism or avoiding an unfavorable issue by raising a different, more favorable matter, or by making a counter-accusation. Since its first use in 1974 (Zimmer, 2017), it has emerged as a common variation of the classical fallacy known as *tu quoque* (lit. "you also") – attacking the opponent's behavior or action for being inconsistent with their argument, thereby discrediting them. Despite significant work devoted to misinformation and propaganda, the detection of whataboutism has largely relied on the ease of tracking "*what about*" phrases. This, however, is informed by popular notions of the phenomenon, leading to a naïve linguistic treatment and subsequently, a neglect of the unique challenges to its detection. Many "what about" phrases do not, in fact, signal propagandist use (Fig. 1).

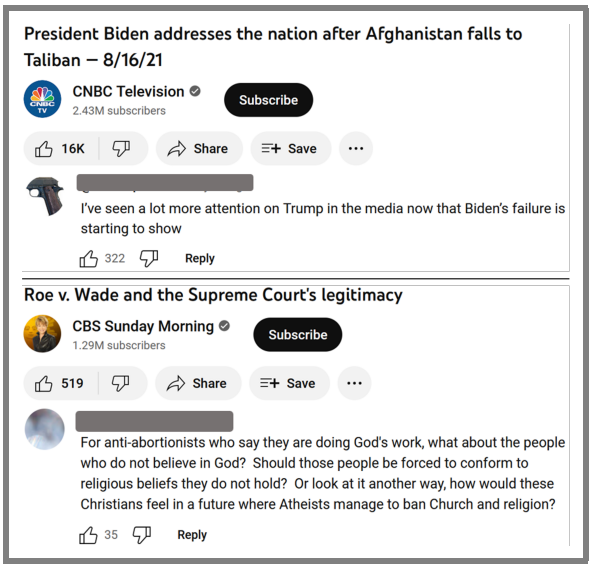


Figure 1: *What about* in YouTube comments: implicit use (top) to discredit the source and redirect the topic, and explicit use (bottom) as an attempt toward reasonable argumentation instead of propaganda.

Whataboutism drives information disorder[3] by derailing proper argumentation, instead of the relatively conspicuous act of misleading by directly lying. As epistemologist Fallis (2015, p. 420) argues, information disorder comprises two equally odious functions: (a) *creating false beliefs*, and (b) *preventing the creation of belief in truth*. Empirical efforts target the former through models for fact-checking and fake news detection (Guo et al., 2022). Whataboutism, on the other hand, is a prevalent and potent contributor to the latter as well. Given its deleterious effects on discourse and social cohesion, it is thus critical for scalable propaganda detection models to be informed by a deeper understanding of whataboutism so that valid argumentation is not conflated with propaganda.

[1]We use the name "Twitter", since our data collection and analysis was conducted while the platform still used that name.

[2]Code and data: github.com/KhiemPhi/wabt-det.

[3]Following the Council of Europe report by Wardle and Derakhshan (2017, pp. 10, 20), 'information disorder" includes dis/misinformation (falsehoods with/out the intention to mislead), as well as malinformation (*e.g.*, hate speech).

Findings of the Association for Computational Linguistics ACL 2024, pages 12628–12643
August 11-16, 2024

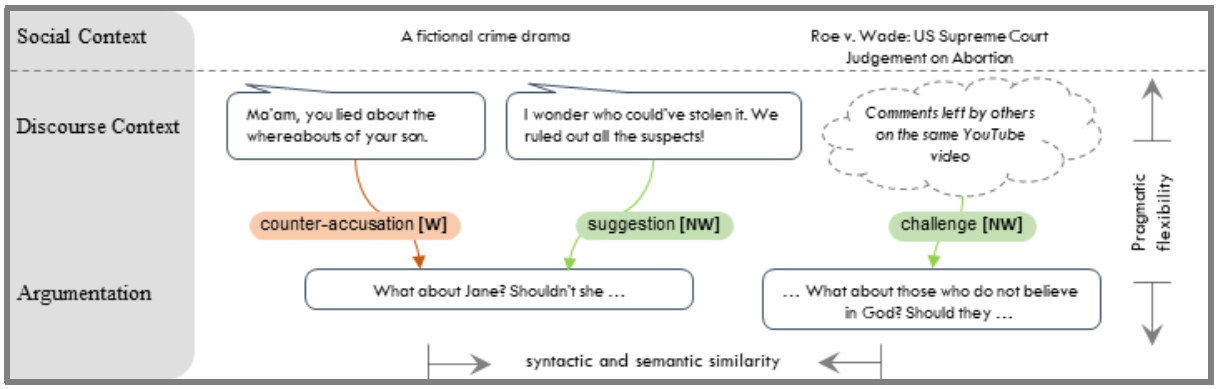


Figure 2: Syntactically and semantically similar (or even identical) responses may exhibit extreme *pragmatic flexibility*: being instances of whataboutism [W] or not [NW] depending on the discursive context. Furthermore, the latter category may include valid argumentation tools such as suggestion or challenge, instead of propaganda.

This work provides just such a scrutiny: with two new datasets (§3), we describe the first rigorous framework for the analysis and detection of whataboutism, distinguishing valid argumentation from propagandist attempts to derail a narrative. We find that instances of whataboutism, redirection, and *tu quoque* fallacies do not coincide in contemporary social media. Contrary to suppositions of prior work on propaganda detection, our analysis reveals that whataboutism serves purposes other than propaganda. Further, we demonstrate that a habitual application of transfer learning or contrastive learning are remarkably poor at detecting whataboutism.[4] To overcome this, we describe a novel method to mine negative samples using attention weights instead of contrast using cosine similarity (§4). With this new formulation of (dis) similarity achieving superior results (§5), our work holds implications for understanding complex discourse surrounding misinformation and propaganda, and modeling natural language pragmatics.

2 Pragmatic flexibility

"What about" $\not\Leftrightarrow$ whataboutism: We utilize the theory of *structured meaning approach* (Krifka, 2001) for a deeper examination beyond early datasets that appear to have conflated the two to a large extent (Da San Martino et al., 2020). Studies of the "what about" construct demonstrate extremely high semantic and pragmatic complexity, but has received little empirical attention (Bledin and Rawlins, 2021). It can be used with a broad range of syntactic constituents, yielding diverse semantics. Further, it has often been framed in terms of an antecedent *question under discussion* (QUD) (Roberts, 2012), wherein a diversion is achieved by establishing the prejacent as a referential topic while simultaneously raising a new question about its properties (Ebert et al., 2014). Thus, it is not the bigram "what about", but the nature of the newly raised question, which determines whether whataboutism is being introduced in a discourse. To wit, Fig. 1 depicts valid argumentation employing this phrase, while an instance of whataboutism does not.

Whether "what about" is explicit or not, it can suggest a solution, challenge a statement, add specification to prior discussion, or laterally redirect the QUD (Beaver et al., 2017; Bledin and Rawlins, 2020). This extreme pragmatic flexibility makes it difficult to *distinguish valid argumentation from propagandist use*. The problem is further compounded by the fact that the *pragmatic distinctions coexist with lexical and semantic similarity*. Fig. 2 illustrates how the same text may serve as (i) harmful deflection via accusation or diversion, which are characteristics of propaganda, and (ii) accommodation or "centering" of attention.[5] Thus, strategies relying solely on syntax or semantics may not differentiate these opposing discourse maneuvers.

3 The Datasets

Da San Martino et al. (2019b) curated a corpus for propaganda detection, containing 76 instances of whataboutism, which serves as a valuable founda-

[4] Identical methods (see §4) have shown state-of-the-art results in various downstream tasks, notably computational propaganda detection. This clear disparity underscores the importance of recognizing whataboutism as overlapping with propaganda, yet distinct from it (as shown in Fig. 2).

[5] The former characterize propaganda, while the latter is a valid act that, incidentally, has immense utility in NLP tasks like anaphora resolution (Grosz et al., 1983; Dekker, 1994).

tion for the study of propaganda. However, its size and assumption (that every occurrence is a propagandist use) limit its suitability for our study. We introduce the TQ$^+$ collections,[6] encompassing the social discourse around each instance. We hope this dataset, featuring ten times more labeled instances than the previous corpus, will spur deeper analyses of fallacies and inspire further research on the role of pragmatics in propaganda and misinformation.

YouTube comments have not received much attention in empirical studies. Our first collection comprises YouTube comments due to a balanced user distribution across demographics (Statista Research Department, 2022a,b,c,d) and a less regulated environment, fostering diverse opinions (Mejova and Srinivasan, 2012). Targeting socially divisive topics prone to whataboutism, we formulate search queries for six such topics on YouTube. From the top five most viewed videos per query, we collect English comments along with the title, transcript, the number of up-votes, and the publisher information. This corpus, TQ^+_{YT}, comprises 1,642 labeled comments from 17 videos across 6 topics, sorted by up-votes.

Similar to TQ^+_{YT}, our **Twitter dataset** TQ^+_{TW} focuses on socially relevant topics that tend to elicit strong emotional responses. We gather English tweets and their replies 8 such topics. The original tweet offers the pragmatic context for identifying whataboutism in responses. While six topics overlap with TQ^+_{YT}, the other two contain fewer tweets, so that our corpus can be used to analyze model performance on topics with limited data. For each topic, we gather tweets with significant engagement, filtering out threads with less than 200 messages. Within each thread, we exclude messages that do not directly respond to the original tweet, lack opinions or discursive content (*e.g.*, "wow"), are socially too inappropriate, consist solely of emojis/emotions, or contain images/videos. In TQ^+_{TW}, each datum comprises a tweet-reply pair, and the collection consists of 1,202 messages.

We use stratified partitioning to divide both datasets into training, validation, and testing sections, ensuring an even distribution of comments across topics in each. Within each topic, comments were split into 80% training, 5% validation, and 15% testing, maintaining the class distribution throughout. These stratified segments are then combined across topics to form the final training, validation, and test sets. The distribution of labeled data across these sets, as well as the class-wise distributions, are described in Appendix A.

We enlist three annotators with native fluency in English who independently assign binary labels to each comment.[7] To attain an understanding of the broader sociopolitical discourse pertaining to each topic, annotators dedicate considerable time to carefully review the entire YouTube video or read the complete Twitter conversation before labeling each instance. This meticulous process is expected to ensure diligent annotation and achieve a high level of data fidelity, despite limiting the size of the datasets. The final label is determined by majority vote, with inter-annotator agreements measured by Fleiss' kappa (Fleiss and Cohen, 1973) at $\kappa = 0.65$ (TQ^+_{YT}) and $\kappa = 0.75$ (TQ^+_{TW}).[8]

Following recent recommendations (Gebru et al., 2021; Bender and Friedman, 2018), we include two comprehensive transparency artifacts: (a) the data statement (Appendix A) and (b) the annotation guide/codebook, included in our data repository.

3.1 A thematic scrutiny

Inspired by the wide adoption of thematic analysis in qualitative research (Braun and Clarke, 2006; Guest et al., 2011), we offer a qualitative discussion to uncover recurring themes in our data. It stands not as an extension, but in complement, to the quantitative study based on corpus annotations. While rare in computational research, thematic analysis has been closely linked to foundational work on word senses and content analysis (Stone et al., 1966; Litkowski, 1997; Ide and Véronis, 1998).

Discourse coherence: Comment-threads on YouTube videos lack coherent discourse. A comment often serves as a standalone response to the video or earlier comments. This differs from discourse structures in articles, interviews, or debates, where whataboutism has been observed (Putz, 2016; Dykstra, 2020). While Twitter threads also show some deviation, our filtering ensures that TQ^+_{TW} offers a cleaner statement-response format.

Perceptions, framing, and *Tu Quoque*: We see that users seldom employ whataboutism as a de-

[6] To highlight that modern use of whataboutism surpasses the classical *tu quoque* fallacy, our datasets are dubbed TQ$^+$, with platform-specific subscripts serving as mnemonic aids.

[7] Shoukri (2003) suggests no additional benefits from more than three annotators for a dichotomous variable.

[8] There is no single threshold for a good value of κ. The widely used interpretations introduced by Landis and Koch (1977) regard these scores as "substantial agreement".

Original context	Comment	W	⇒	TQ
YT Video: One dead after 'Unite the Right' rally in Virginia (Fox News)	(1) if such things happened in china and russia, what would cnn and nytimes say?	✓	✓	
	(2) If the perp was an Islamist, imagine how the left would dismiss what he did as not being indicative of Islam (like they always do).	✓	✓	
YT Video: Outrage grows over Russian bounties (ABC News)	(3) They did the same thing to indigenous folks for bounty hunters and military, the US military deserves this.			✓
YT Video: Russians Flee Into Exile Because Of Putin's War With Ukraine: NYT (MSNBC)	(4) Fox News, Republicans and Trump all defended this monster. They are now hoping and praying that you forget that part.		✓	
YT Video: Russia is aware it is not winning the war in Ukraine: Brookings senior fellow (CNBC)	(5) Russia bombing buildings four miles from Kyiv's center doesn't sound like losing. How close have the Ukrainians gotten to Moscow?	✓		✓
	(6) USA baby not didn't war crime afganistan Iraq	✓	✓	✓
Tweet: We united the world to protect Ukraine, we will unite the world to restore justice. Russian invaders will be legally and fairly held to account for all war crimes. The terrorist state will be held to account for the crime of aggression. (@ZelenskyyUa; Mar 3, 2023)	(7) Coming from a govt that congratulated Bola Tinubu of Nigeria. Funny people	✓	✓	
	(8) And what of Ukraine's war crimes? Torture and murder of POWs, firing from civilian positions, using human shields or shooting civilians for receiving Russian ration packs. ALL war crimes need to be prosecuted, not just those by Russians.	✓	✓	✓

Table 1: Comments and their corresponding contexts (YouTube video title, or tweet to which they are responding) along with three facets: whataboutism (**W**), topic redirection (⇒), and conformity with classical *tu quoque* (**TQ**).

fensive tactic to deflect criticism. Instead, it is frequently wielded to preclude the relevant issue by *introducing* an accusation, diverging from the traditional view of whataboutism as a response with a *counter*-accusation. Users frequently employ whataboutism to shape perceptions, understanding, or evaluation of an issue, particularly through the **lens of domestic political divides**. Table 1 illustrates this trend: comments on a Fox News video critique liberal media outlets like CNN and the New York Times (1, 2), while criticisms of conservative entities appear under content from liberal organizations (4), and even centrist channels (3).[9]

Users commonly employ whataboutism to highlight perceived **selection bias** or **hypocrisy** in news coverage, reframing overlooked issues in a new context. Assessing whether such reframing constitutes propaganda is complex, given its prevalence even in journalism from reputable organizations.[10] Social media users similarly engage in this practice: in (5), a central claim is challenged with what the author views as overlooked evidence;[11] and (7) demonstrates the use of whataboutism in an appeal to hypocrisy but the deflection is not a counter-accusation, diverging from the *tu quoque* fallacy.

There are similarities between whataboutism in propaganda, whataboutism in general discourse, and the "what about" bigram. Our thematic analysis, however, reveals distinct characteristics, preventing one from implying the other, despite topical redirection. Some comments, like Table 1:(4), overtly redirect without using whataboutism, while others, like (5), attempt to reframe the same topic. Furthermore, contemporary whataboutism differs from the classical fallacy as they deflect to entities not directly involved in the conversation. Rather than a two-person interaction, commenters may perceive the video or previous posts as representative of "tu". For instance, in Table 1:(3, 6), they may view the videos as representing the U.S. government and comment on past military actions. However, instances like (2) introduce accusations unrelated to the discourse ("the left"). Twitter examples, like (8), align more closely with classical *tu quoque*, likely due to the direct tweet-response structure.

Sarcasm and irony: Since subjectivity and opinion influence people significantly (Picard, 2000), we analyze our datasets for subjective expressions and emotive content. We find that explicit sentiment polarity or emotions are rare, while sarcasm or irony abound, as seen in Table 1:(1, 6, 7).

[9]The liberal, conservative, or centrist leanings mentioned in this discussion are obtained from Ad Fontes Media (2022).

[10]For example: (1) Dhanesha, N. *Climate fixes are all aimed at property owners. What about renters?* July 27, 2022. Vox. Accessed: 06.23.2023. (2) Hawkins, A. and Davidson, H. *As the west tries to limit TikTok's reach, what about China's other apps?* April 12, 2023. The Guardian. Accessed: 01.03.2024.

[11]Table 1:(5) implies there are the only possible outcomes, showcasing the false dichotomy fallacy.

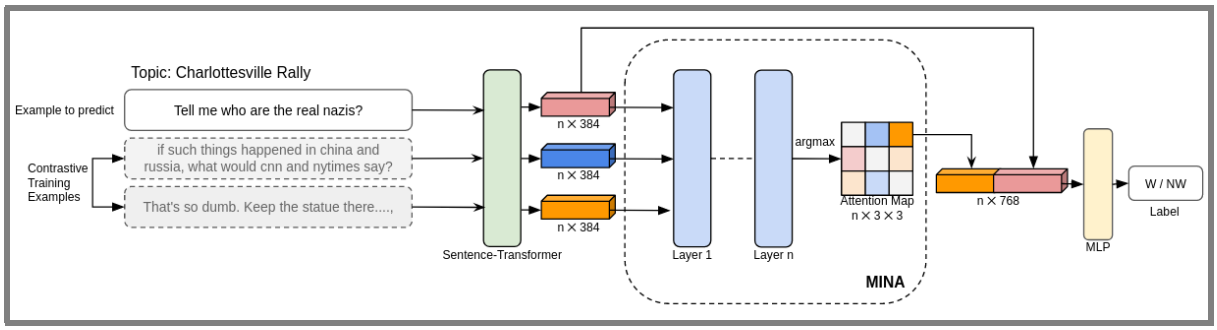


Figure 3: MINA (**Mi**ning **N**egatives with **A**ttention) employs attention weights in the final layer of the Transformer encoder as a measure of pragmatic contrast. The complete architecture is shown here *in situ*.

4 Experiments

We start with two sets of experiments before describing our novel approach: (i) transfer-learning techniques akin to recent empirical research on propaganda detection, and (ii) possible enhancements to these baselines by incorporating conventional measures of semantic (dis-) similarity.

4.1 Baselines: (Habitual) Transfer learning

To establish competitive baselines, we look to the best performing models from the SemEval-2020 propaganda detection task (Da San Martino et al., 2020).[12] We incorporate a fully connected layer and fine-tune them using the following configurations on a single NVIDIA Titan XP GPU: (i) batch size of 80 for BERT-base and 40 for RoBERTa, (ii) 10 epochs, (iii) maximum sequence length of 256, and (iv) learning rate (η) set to 1×10^{-4}.

Further, along the lines adopted by Vlad et al. (2019) and Yu et al. (2021), we include another baseline that integrates affective language with task-specific fine-tuning. Due to our corpus being rich in irony and sarcasm but sparse in explicit expressions of sentiment, we fine-tune a RoBERTa model pretrained on irony detection (Barbieri et al., 2020) with the same training configuration.

4.2 (Conventional) Semantic similarity

Given the diversionary nature of whataboutism, identifying it hinges on discerning a sentence's relation to the antecedent question under discussion (QUD). For improvements over baseline results, we thus explore models that promise a deeper semantic understanding within a topic's discourse. Given the inadequate performance – particularly on YouTube comments – of language models pretrained on token prediction, we adopt SBERT (Reimers and Gurevych, 2019), which is trained on natural language inference (NLI) datasets (Bowman et al., 2015; Williams et al., 2018), and excels in NLI and semantic textual similarity (STS) benchmarks. We fine-tune its embeddings with a configuration identical to that described earlier for BERT.

We also explore two additional contrastive learning methods: (i) "Correct and Smooth" (C&S) (Huang et al., 2021), which constructs a graph where the embeddings serve as nodes and cosine similarity provides edge weights, and tunes the embeddings via label propagation (Zhou et al., 2003); and (ii) SimCSE (Gao et al., 2021), which has achieved notable results in STS tasks with supervision from annotated pairs in NLI datasets (with *entailment* and *contradiction* pairs serving as positive and negative samples, respectively).

Notwithstanding their proficiency in STS benchmarks, the results of our experiments (§5) reveal these models to be inadequate in understanding the pragmatic variations prevalent within each topic's social discourse. A common thread in their methodology is the use of cosine similarity to differentiate between instances from positive and negative samples. Cosine similarity may present challenges in high-dimensional vector spaces, as it becomes highly probable that any two vectors are nearly orthogonal (Appendix B offers a formal proof). Further, as experiments reveal significant room for improvement over these models, we conjecture that approaches relying on cosine measure may struggle to detect whataboutism. This challenge drives us to explore beyond semantic similarity to capture pragmatic differences in modern social media discourse, with empirical results described in §5.

[12] While Piskorski et al. (2023) present a related task, we opt not to use the top models of that task due to: (a) the multilingual corpus; (b) their use of fine-tuned Transformer models (or their ensembles), much like the models in the earlier task; and (c) the poorer English-language whataboutism detection results, compared to the SemEval-2020 task.

(a) Whataboutism detection on YouTube comments

Model	W			NW		
	P	R	F_1	P	R	F_1
Transfer learning (baseline models):						
BERT	0.72	0.36	0.48	0.92	0.98	0.95
RoBERTa	0.23	0.71	0.34	0.94	0.64	0.76
RoBERTa $_{\text{Irony}}$	0.30	0.30	0.30	0.93	0.99	0.96
Based on conventional measures of semantic similarity:						
SBERT	0.13	0.05	0.07	0.96	0.99	0.97
C&S	0.45	0.45	0.45	0.91	0.91	0.91
SimCSE	0.27	0.68	0.38	0.94	0.72	0.82
Mining negatives with attention (MINA):						
SBERT	0.63	0.53	0.58*	0.94	0.96	0.95
RoBERTa $_{\text{Irony}}$	0.57	0.47	0.52	0.94	0.95	0.94

(b) Whataboutism detection on Twitter replies

Model	W			NW			Params
	P	R	F_1	P	R	F_1	(M)
Transfer learning (baseline models):							
BERT	0.64	0.76	0.69	0.85	0.76	0.80	110
RoBERTa	0.64	0.80	0.71	0.87	0.75	0.81	124
RoBERTa $_{\text{Irony}}$	0.54	0.89	0.67	0.91	0.57	0.70	124
Based on conventional measures of semantic similarity:							
SBERT	0.64	0.76	0.69	0.85	0.76	0.80	22.7
C&S	0.57	0.60	0.59	0.95	0.91	0.93	22.7
SimCSE	0.64	0.80	0.71	0.87	0.74	0.80	22.7
Mining negatives with attention (MINA):							
SBERT	0.73	0.78	0.75*	0.87	0.84	0.85	31
RoBERTa $_{\text{Irony}}$	0.91	0.57	0.70	0.54	0.89	0.67	133

Table 2: Whataboutism detection results on (a) YouTube and (b) Twitter: baseline results (top) using transfer-learning with fine-tuning; (middle) conventional semantic similarity measures; and (bottom) dis/similarity based on **M**ining **N**egative samples with **A**ttention (MINA). Macro-average (**P**)recision, (**R**)ecall, and $\boldsymbol{F}_1$ for the target class are shown on column-wise color gradients, with blue (■) indicating the best performance and yellow (■) the worst.

4.3 Mining negatives with attention

Given the centrality of topical redirection in whataboutism, any representation distinguishing it requires an understanding of the pragmatic context, namely the antecedent QUD. We thus model each comment by training on tuples comprising the comment t (which we seek to classify), along with c comments each from the W and NW classes. For our collection of YouTube comments ($\text{TQ}^{+}_{\text{YT}}$), the $2c$ comments are selected from the same video as t; and for $\text{TQ}^{+}_{\text{TW}}$, from the same thread as the original tweet. This hyperparameter c controls context incorporation during training. We reason that such tuples can better capture semantic redirection *within the ambient context of the social topic*, than embeddings based solely on global distributional semantics (Lenci et al., 2022). With this setup[13], the most dissimilar examples can be identified using a context tuple. Then, combining them with a representation of t will encode pragmatic shifts through an implicit *grounding of the text in the surrounding social discourse*.[14]

Specifically, we apply a Transformer encoder with d layers and h attention heads (Vaswani et al., 2017) to incorporate cross-attention between the elements of the context tuple. We then extract the cross-attention map from the encoder's final layer for a similarity matrix. After identifying the example with the highest numeric value, we concatenate its text with t. This combined embedding is fed into a multi-layer perceptron for final classification. Termed ***mi***ning ***n***egatives with ***a***ttention (MINA), this method employs cross-attention scores to mine the most pragmatically dissimilar examples in the discourse surrounding t. Fig. 3 is an *in situ* illustration of its complete architecture. Ablation studies (§5) dictate our chosen configuration of 2 encoder layers, 32 attention heads, SBERT embeddings as input, and a context size of 1.

5 Results and analysis

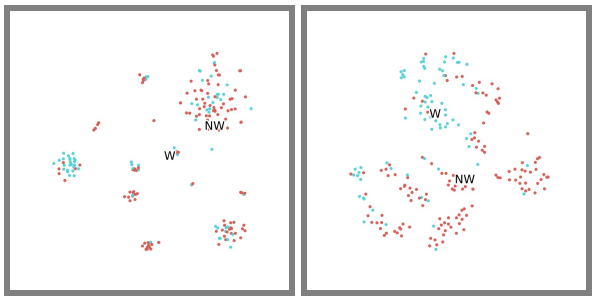
Table 2 displays the results of baseline transfer-learning on pretrained Transformer-based models, models based on semantic similarity, and MINA.

Baseline models: The baseline models perform reasonably well on Twitter, but not on $\text{TQ}^{+}_{\text{YT}}$: BERT achieves higher precision but lower recall, while RoBERTa shows the opposite behavior. It is thus evident that models adept at propaganda detection have room for improvement in identifying whataboutism, especially given their tendency to assign the majority label NW to test instances.

Semantic similarity models: When applying the standard transfer-learning technique (*i.e.*, fine-tuning on task-specific labeled data) to language

[13]The language model m, which provides the representation of t and contextual comments, can be viewed as a discrete parameter of the complete approach, where one model may be swapped out for another.

[14]In effect, we surmise that the pragmatic act is grounded not just in the context of immediately surrounding texts, but in a sampling of the entire social discourse about that topic. This is similar in spirit to visually grounding objects through cross-modal attention (Ilinykh and Dobnik, 2022).



(a) Transfer-learning from pre-trained language models. (b) Contrasting with MINA to capture pragmatic differences.

Figure 4: t-SNE visualization of the target class (W; blue), amid everything else (NW; red) with (a) transfer-learning with pretrained language models, and (b) mining negative samples using MINA. The latter demonstrates better separability of the target class.

(a) *Varying h, with* $d = 4, c = 1$ ($\sigma^2 = 0.0086$):		
h (attention heads)	$\overline{F_1}$ (YouTube)	$\overline{F_1}$ (Twitter)
32	0.52	0.71
64	0.52	0.70
128	0.51	0.70
384	0.48	0.69
(b) *Varying c, with* $d = 4, h = 32$ ($\sigma^2 = 0.0072$):		
c (context size)	$\overline{F_1}$ (YouTube)	$\overline{F_1}$ (Twitter)
1	0.51	0.71
2	0.50	0.70
3	0.51	0.67
4	0.49	0.69
(c) *Varying m, with* $d = 4, h = 32, c = 1$ ($\sigma^2 = 0.0089$):		
m (language model)	$\overline{F_1}$ (YouTube)	$\overline{F_1}$ (Twitter)
SBERT	0.52	0.71
RoBERTa $_{\text{Irony}}$	0.51	0.70

Table 3: Ablation study on MINA hyperparameters: d (encoder layers), h (attention heads), c (context size), and m (input embeddings). Mean F_1 and its variance σ^2 are reported over 50 runs.

models known for their excellent performance on various semantic similarity tasks, we find no consistent improvement over baseline models. Once again, all models display a strong inclination to classify $\text{TQ}^{+}_{\text{YT}}$ test instances into the majority class, with SBERT exhibiting the poorest performance in this regard. SimCSE, trained on hard negative samples from NLI datasets, shows relatively better performance, while the simplest model in this category, C&S, achieves the highest F_1 score of 0.45. On $\text{TQ}^{+}_{\text{TW}}$, the results are consistently better, with SBERT in particular showing remarkable improvement. But these models fail to notably surpass the baseline transfer-learning methods.

Mining negatives with attention (MINA): Even while employing contextual embeddings, the inability of the above models to effectively detect whataboutism suggests a limitation in capturing the kind of pragmatic nuances illustrated earlier (Fig. 1, 2). MINA, however, is designed to leverage cross-attention scores across comments within a topic to model pragmatic context. To demonstrate its efficacy, we assess the *worst*-performing models from previous categories, focusing on F_1 for the target class. On $\text{TQ}^{+}_{\text{YT}}$, F_1 of RoBERTa $_{\text{Irony}}$ increases by 22%, rising from 0.3 to 0.52, while SBERT shows an exceptional improvement of 51%, from 0.07 to 0.58: a 10% boost over the next best result.[15] When evaluated on $\text{TQ}^{+}_{\text{TW}}$, SBERT with MINA outperforms other models, showing a 6% improvement over vanilla SBERT and a 4% improvement over RoBERTa and SimCSE. RoBERTa $_{\text{Irony}}$ initially exhibits high recall but low precision due to being predisposed to irony. However, MINA often flips the output when irony is present in comments but not in other context tuple elements. The significant improvements we observe with MINA are reflected in its effects on target class separability. Comparing the t-SNE (Hinton and Roweis, 2002) visualizations in Fig. 4a and Fig. 4b, it becomes clear that while fine-tuning on a moderate amount of task-specific training data has limited utility, contrasting with MINA improves the ability to distinguish whataboutism from everything else.

Ablation experiments: To determine the optimal configuration for MINA, we conduct a series of experiments varying hyperparameters d (number of encoder layers), h (number of attention heads), d (number of comments for context size), and m (language model for input embeddings) (see Table 3). The models are trained end-to-end with the configuration described earlier (§4). Each experiment varies one hyperparameter while keeping others constant. Due to random selection of comments in context tuples, each ablation undergoes 50 runs, with mean F_1 and variance reported. The best hyperparameter configuration and its best run informs the results in Table 2.[16]

The results indicate no benefit in increasing the

[15] To compare, the SemEval-2020 task on propaganda detection (Da San Martino et al., 2020) saw the best performance in detecting whataboutism, straw man arguments, and red herrings, achieve $F_1 = 0.269$. (these three closely-related fallacies were merged into a single group due to insufficient data.

[16] For reproducibility, context tuples and model weights from the best run are made available with our code and data.

Method	$\overline{F_1}$ (YouTube)	$\overline{F_1}$ (Twitter)
MINA	**0.52**	**0.71**
Random	0.50	0.67
Cosine-Sim	0.50	0.67

Table 4: MINA's novel approach using cross-attention weights to construct pragmatically dissimilar contexts is a superior sample mining strategy.

encoder layers, attention heads, and context size beyond 4, 32, and 2, respectively. For the model m, we use SBERT and RoBERTa $_{\text{Irony}}$. Across 50 trials, SBERT yields marginally higher average F_1 scores for YouTube (0.52, against 0.51 for RoBERTa $_{\text{Irony}}$) and Twitter (0.71, against 0.70 for RoBERTa $_{\text{Irony}}$). The variance across all 50 trials for selecting the embeddings was very low, at $\sigma^2 = 0.0089$.

A comparison of mining strategies: We compare MINA's negative sample mining with two common strategies in recent contrastive mining literature: cosine similarity-based mining (*e.g.*, Wang et al., 2021) and random sampling (Jiang et al., 2021; Xu et al., 2022). Table 4 shows MINA's superior result, while the use of cosine measure is akin to random sampling for tuple construction. This further supports our conjecture that conventional measures of semantic similarity or contrast are inadequate for capturing pragmatic differences.

Challenges, negative results, and insights: In exploring several models and techniques to detect whataboutism, the path to MINA, which uses negative sampling inspired by cross-attention, was marked by several unexpected outcomes. First, we experimented with the advanced pre-trained encoder DeBERTa (He et al., 2021) on the collection of YouTube comments, but its performance was unexpectedly inferior to BERT, achieving an F_1 score of 0.455 compared to BERT's 0.48.

Although *modern large language models* (LLMs) like Llama-2-7b, Llama-2-13b (Touvron et al., 2023), and GPT-3.5-turbo (Brown et al., 2020) have shown impressive capabilities in various tasks, they did not yield satisfactory results in detecting whataboutism. The limitations we observe align with those noted by Ruiz-Dolz and Lawrence (2023) in detecting fallacious argumentation. On Twitter comments, for instance, the highest F_1 score was 0.65, 4% lower than the BERT baseline (Table 2). Even with sophisticated contextual prompting, no significant improvements were observed. GPT-3.5-turbo excelled with canonical whataboutism that closely mirrored syntactic patterns found in popular media, but struggled with nuanced, rephrased instances. Consequently, these LLMs were excluded from subsequent stages of experimentation in this work.

Our experiments included *incorporating video transcripts as contextual information*, for which we leveraged Longformer (Beltagy et al., 2020). Contrary to initial expectations, these yielded lower precision and recall across all models. Analyzing the errors revealed that comments from other users are a better predictor than the contextual information derived from the video. This observation, plus our thematic analysis (§3.1) and the success of MINA strongly indicate that while the video remains important in setting the topic, detecting whataboutism in a comment often depends more on the broader social discourse surrounding the topic.

Lastly, we decided against *batch-contrastive learning* due to two potential concerns raised by Khosla et al. (2020) and Qu et al. (2021): (i) large batch sizes are vital for it to be effective, and (ii) it typically involves augmentation, which introduces noise that hinders pragmatic understanding. Further, it would still use cosine similarity on its own, which we have questioned in §4.2.

6 Related Work

Propaganda detection: Early propaganda detection (Barrón-Cedeño et al., 2019) categorized sources based on external inputs, but concerns were raised due to noisy data and questions about epistemic bias (Uscinski and Butler, 2013; Marietta et al., 2015). Traditionally, whataboutism was viewed as the tu quoque fallacy (Fischer, 2021). Another body of work, however, views it in the broader pragmatic and dialectic framework (Aikin, 2008). While this has gained prominence in theoretical studies (Bowell, 2023), computational approaches remain confined to viewing whataboutism purely as propaganda (Da San Martino et al., 2019b; Sahai et al., 2021; Baleato Rodríguez et al., 2023). Our work connects to the pragma-dialectic view of whataboutism: overlapping with propaganda but not subsumed by it.

NLP tasks have traditionally favored utilizing pretrained models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), fine-tuning them for specific tasks. Da San Martino et al. (2019b) accordingly approached propaganda detection, creating a corpus of news articles later used in shared tasks (Da San Martino et al., 2019a,

2020), where these pretrained models achieved top results (Yoosuf and Yang, 2019; Morio et al., 2020). While related, our research differs substantially. They target propaganda detection, treating whataboutism only as propagandist language while the limited dataset size precludes deeper analysis. Da San Martino et al. (2020) nevertheless note the difficulty in identifying whataboutism.

Since identifying the precise textual span of propaganda is laborious and error-prone, we look to Sahai et al. (2021), who label informal fallacies in social media comments. However, they continue to treat whataboutism purely as propaganda, while also turning to transfer learning through fine-tuning the models (*i.e.*, the approach we use as baselines).

Framing and persuasion: Unlike earlier studies, Piskorski et al. (2023) view whataboutism as strategic framing and persuasion rather than outright propaganda, thus offering a taxonomy that aligns more closely with our study. However, instances of whataboutism are sparse in their multilingual corpus (only 25 in English), limiting its utility in our work. Additionally, fine-tuned Transformer models or ensembles, as utilized by top-performing participants, yielded remarkably low F_1 scores for whataboutism and other related categories (see Wu et al., 2023). The results of this shared task are noteworthy, yet they are confounded by ambiguity (Purificato and Navigli, 2023; Liu et al., 2023) and low inter-annotator agreement. Addressing these challenges is crucial for an accurate interpretation or analysis of the findings.

Modeling (dis) similarity. Whataboutism, a discursive maneuver for diversion and centering, relates to semantic textual similarity (STS) (Cer et al., 2017). Modeled as a graph, STS becomes a node classification task using graph neural networks (GNN). Huang et al. (2021), however, show that minor modifications to classical graph-based learning algorithms (*e.g.*, Zhou et al., 2003) outperform large GNNs despite far fewer parameters and less training time. Their approach (C&S) does not require careful early stopping criteria or a large validation set, making it another competitive baseline.

STS models do not capture whataboutism's pragmatic complexity. We address this by sharing a core intuition with contrastive learning: bring similar instances closer while pushing apart dissimilar ones (Weinberger and Saul, 2009). Augmentation methods used in contrastive learning, however, are ill-suited for pragmatics. Instead, we opt for SimCSE (Gao et al., 2021). While hard negative mining has proven effective (Xiong et al., 2021), the *use of attention weights to model pragmatic differences* is a novel contribution of this work. Here, notable work includes Le et al. (2018), who integrate attention weights with syntax, and Yamagiwa et al. (2022), who fuse self-attention matrices with word-mover's distance to obtain promising STS results.

7 Conclusion

Our study underscores areas where state-of-the-art propaganda detection models often fail to distinguish propagandist use of whataboutism from valid argumentation tools or figurative language use like irony, revealing how their suitability for this task may be improved. Addressing this within the pragma-dialectic framework, we illustrate the need to develop models capable of understanding pragmatic variations. We find traditional similarity measures ineffective for this purpose. Cross-attention proves valuable instead. We thus propose a novel methodology, MINA, for mining negative samples based on cross-attention, which achieves superior results in whataboutism detection and suggests its utility in other tasks that require grounding in ambient social discourse. This study is based on our contribution of two annotated datasets that can facilitate further research in related areas.

8 Limitations

Emotive topics may influence the labels assigned to specific instances. We report Fleiss' kappa, adjusting for the possibility of chance agreement, but our approach does not take into account individual viewpoints. We thus advocate imbibing this into *perspectivist* research (Cabitza et al., 2023).

We have shared model weights, code, and datasets for replicability. MINA's reliance on randomized tuple batches, however, may cause slight changes in retrained models in spite of extremely low variance (reported in Table 3).

The insights regarding the underwhelming performance of modern LLMs are crucial for understanding their limitations in handling tasks like whataboutism detection, where pragmatic nuances are important. In this regard, a deeper exploration of in-context learning is warranted.

Finally, we note that the TQ^+ collections are imbalanced, with a minority target class. This, however, mirrors real-world scenarios and may not be a disadvantage after all.

References

Ad Fontes Media. 2022. Interactive Media Bias Chart. adfontesmedia.com/interactive-media-bias-chart. Accessed: 04.24.2024.

Scott Aikin. 2008. Tu Quoque Arguments and the Significance of Hypocrisy. *Informal Logic*, 28(2):155–169.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper Bullets: Modeling Propaganda with the Help of Metaphor. In *Findings of the Assoc. Comput. Linguist.: EACL 2023*, pages 472–489.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Assoc. Comput. Linguist.: EMNLP 2020*, pages 1644–1650.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Inf. Process. Manag.*, 56:1849–1864.

David I. Beaver, Craige Roberts, Mandy Simons, and Judith Tonhauser. 2017. Questions Under Discussion: Where Information Structure Meets Projective Content. *Annu. Rev. Linguist.*, 3(1):265–284.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *ArXiv*, abs/2004.05150v2.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Justin Bledin and Kyle Rawlins. 2020. Resistance and Resolution: Attentional Dynamics in Discourse. *J. Semant.*, 37(1):43–82.

Justin Bledin and Kyle Rawlins. 2021. About "What About": Topicality at the Semantics-Pragmatics Interface. Handout presented at the 31st Semantics and Linguistic Theory Conference.

Tracy Bowell. 2023. Whataboutisms: The Good, the Bad and the Ugly. *Informal Logic*, 43(1):91–112.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proc. 11th Int. Workshop Semant. Eval. (SemEval-2017)*, pages 1–14.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. In *Proc. 2nd Workshop Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proc. 14th Workshop Semant. Eval.*, pages 1377–1414.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-Grained Analysis of Propaganda in News Article. In *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, pages 5636–5646.

Paul Dekker. 1994. Predicate logic with anaphora. In *Proceedings of the 4th Semantics and Linguistic Theory Conference*, pages 79–95. Linguistic Society of America.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conf. N. Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. Vol. 1 (Long and Short Papers)*, pages 4171–4186.

Alan Dykstra. 2020. The rhetoric of "whataboutism" in american journalism and political identity. *Res Rhetorica*, 7(2):2–16.

Christian Ebert, Cornelia Ebert, and Stefan Hinterwimmer. 2014. A unified analysis of conditionals as topics. *Linguist. Philos.*, 37(5):353–408.

Don Fallis. 2015. What is disinformation? *Library Trends*, 63(3):401–426.

John Martin Fischer. 2021. How We Argue Now. *The Philosopher's Magazine*, pages 30–35.

Joseph L. Fleiss and Jacob Cohen. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ. Psychol. Meas.*, 33(3):613–619.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proc. 2021 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pages 6894–6910.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM*, 64(12):86—-92.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a Unified Account of Definite Noun Phrases in Discourse. In *21st Annu. Meet. Assoc. Comput. Linguist.*, pages 44–50. Association for Computational Linguistics.

Greg Guest, Kathleen M. Macqueen, and Emily E. Namey. 2011. *Applied Thematic Analysis*. Sage Publications, Thousand Oaks, California, USA.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguist.*, 10:178–206.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Geoffrey Hinton and Sam Roweis. 2002. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 857–864. MIT Press.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. 2021. Combining Label Propagation and Simple Models out-performs Graph Neural Networks. In *9th International Conference on Learning Representations (ICLR)*.

Nancy Ide and Jean Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):1–40.

Nikolai Ilinykh and Simon Dobnik. 2022. Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. 2021. Improving Contrastive Learning on Imbalanced Data via Open-World Sampling. In *Advances in Neural Information Processing Systems*, volume 34, pages 5997–6009. Curran Associates.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates.

Manfred Krifka. 2001. For a Structured Meaning Account of Questions and Answers. In Wolfgang Sternefeld and Caroline Féry, editors, *Audiatur Vox Sapientiae: A Festschrift for Arnim von Stechow*, volume 52 of *Studia Grammatica*. Walter de Gruyter GmbH.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Yuquan Le, Zhi-Jie Wang, Zhe Quan, Jiawei He, and Bin Yao. 2018. ACV-tree: A New Method for Sentence Similarity Modeling. In *Proc. 27th Int. Jt. Conf. Artif. Intell., IJCAI-18*, pages 4137–4143.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Lang. Resour. Eval.*, 56(4):1269–1313.

Kenneth C. Litkowski. 1997. Desiderata for Tagging with WordNet Synsets or MCCA Categories. In *Tagging Text with Lexical Semantics: Why, What, and How?*

Genglin Liu, Yi Fung, and Heng Ji. 2023. NLUBot101 at SemEval-2023 Task 3: An Augmented Multilingual NLI Approach Towards Online News Persuasion Techniques Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1636–1643, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Morgan Marietta, David C. Barker, and Todd Bowser. 2015. Fact-Checking Polarized Politics: Does The Fact-Check Industry Provide Consistent Guidance on Disputed Realities? *The Forum*, 13(4):577–596.

Yelena Mejova and Padmini Srinivasan. 2012. Political Speech in Social Media Streams: YouTube Comments and Twitter Posts. In *Proc. 4th Annu. ACM Web Sci. Conf.*, WebSci '12, pages 205—208.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 Task 11: An Empirical Study of Pre-Trained Transformer Family for Propaganda Detection. In *Proc. 14th Workshop Semant. Eval.*, pages 1739–1748.

Rosalind W. Picard. 2000. *Affective Computing*. MIT Press.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proc. Jt. 15th Linguist. Annot. Workshop (LAW) 3rd Des. Mean. Represent. (DMR) Workshop*, pages 133–138.

Antonio Purificato and Roberto Navigli. 2023. APatt at SemEval-2023 Task 3: The Sapienza NLP System for Ensemble-based Multilingual Propaganda Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 382–388, Toronto, Canada. Association for Computational Linguistics.

Catherine Putz. 2016. Donald Trump's Whataboutism. thediplomat.com/2016/07/donald-trumps-whataboutism. Accessed: 08.09.2022.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding. In *International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, pages 3982–3992.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models. In *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore. Association for Computational Linguistics.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions. In *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. (Vol. 1: Long Papers)*, pages 644–657.

Mohamed M. Shoukri. 2003. *Measures of Interobserver Agreement and Reliability*, 1st edition. Chapman & Hall/CRC Biostatistics Series. CRC Press.

Statista Research Department. 2022a. Distribution of Twitter users worldwide as of January 2022, by gender. www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender. Accessed: 04.24.2022.

Statista Research Department. 2022b. Leading countries based on number of Twitter users as of January 2022. www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries. Accessed: 04.24.2022.

Statista Research Department. 2022c. Leading countries based on YouTube audience size as of January 2022. www.statista.com/statistics/280685/number-of-monthly-unique-youtube-users. Accessed: 04.24.2022.

Statista Research Department. 2022d. Most popular social networks worldwide as of January 2022, ranked by number of monthly active users. www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users. Accessed: 2022-04-24.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, Massachusetts, USA.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.

Joseph E. Uscinski and Ryden W. Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, NIPS '17, pages 6000—-6010. Curran Associates.

George-Alexandru Vlad, Mircea-Adrian Tanase, Cristian Onose, and Dumitru-Clementin Cercel. 2019. Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-capsule model. In *Proc. 2nd Workshop Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 148–154. Association for Computational Linguistics.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 2021. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033.

Claire Wardle and Hossein Derakhshan. 2017. Information Disorder: Towards an Interdisciplinary Framework for Research and Policy-Making. Technical report, Council of Europe.

Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.*, 10(2):207–244.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1995–2008, Toronto, Canada. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations (ICLR)*.

Lanling Xu, Jianxun Lian, Wayne Xin Zhao, Ming Gong, Linjun Shou, Daxin Jiang, Xing Xie, and Ji-Rong Wen. 2022. Negative Sampling for Contrastive Representation Learning: A Review.

Hiroaki Yamagiwa, Sho Yokoi, and Hidetoshi Shimodaira. 2022. Improving word mover's distance by leveraging self-attention matrix.

Shehel Yoosuf and Yin Yang. 2019. Fine-Grained Propaganda Detection with Fine-Tuned BERT. In *Proc. 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91. Association for Computational Linguistics.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. 2021. Interpretable Propaganda Detection in News Articles. In *Proc. Int. Conf. Recent Adv. Nat. Lang. Process. (RANLP 2021)*, pages 1597–1605.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pages 321–328. MIT Press.

Ben Zimmer. 2017. The Roots of the 'What About?' Ploy (9 June, 2017). *The Wall Street Journal*. www.wsj.com/articles/the-roots-of-the-what-about-ploy-1497019827. Accessed: 05.05.2022.

A Datasheet

This datasheet is included based on the recommendations of Gebru et al. (2021). It serves to document the creation, composition, intended uses, and maintenance of the TQ$^+$ dataset ("tu quoque and beyond") released with this work. We hope this will facilitate its better usage, and further encourage transparency, accountability, and reproducibility.[1]

Why was the dataset created? TQ$^+$ was created to enable research on identifying expressions of whataboutism in social media posts made by active participating users: given an English (en-US) comment made in the specific context, identify whether the comment expresses whataboutism. Intentionally created for this task, the dataset comprises of several hot-button sociopolitical topics, where whataboutism is more likely to be employed. While there exist a few datasets for propaganda detection (Da San Martino et al., 2019a), there is no corpus exclusively devoted to the study of whataboutism, which has unique sociopolitical and linguistic properties (ranging from semantics to pragmatics) that are distinct from other propagandist maneuvers. Furthermore, many who employ whataboutism are neither the creators nor the intentional disseminators of propaganda. Instead, many instances are attempts made by users to challenge or re-center a narrative (better captured in the taxonomic changes seen in Piskorski et al. (2023)). These factors distinguish TQ$^+$ from earlier datasets.

What other tasks could the dataset be used for? TQ$^+$ can be used for various types of modeling or comprehension of whataboutism in social media commentary. For instance, one may study syntactic patterns and their correlation with expressions of whataboutism, or how other users react to comments that express whataboutism. It can also be used to learn from adversarial settings, wherein this dataset can be used to automatically generate whataboutism on politically divisive topics. Furthermore, we expect TQ$^+$ to be useful for discourse analyses and other studies conducted by social scientists and media communication researchers.

Has the dataset been used for any tasks already? This is a novel dataset. As of June 2023, it has not been used for any other task or publication.

[1]For the sake of brevity, we have not included information that is included in the main body of this paper.

A.1 Dataset composition

TQ$^+$ comprises two sub-datasets: TQ$^+_{\text{YT}}$, which consists of comments made by active users on YouTube videos, and TQ$^+_{\text{TW}}$, which consists of Twitter posts.

Fig. 5a shows the distribution of these comments and their labels across the topics in the YouTube corpus.The Twitter corpus has 6 topics that are in TQ$^+_{\text{YT}}$, and additionally, *Tiktok ban* and *Trump indictment*. The data distribution across all these topics, along with the distribution of the binary ground-truth labels, is shown in Fig. 5b.

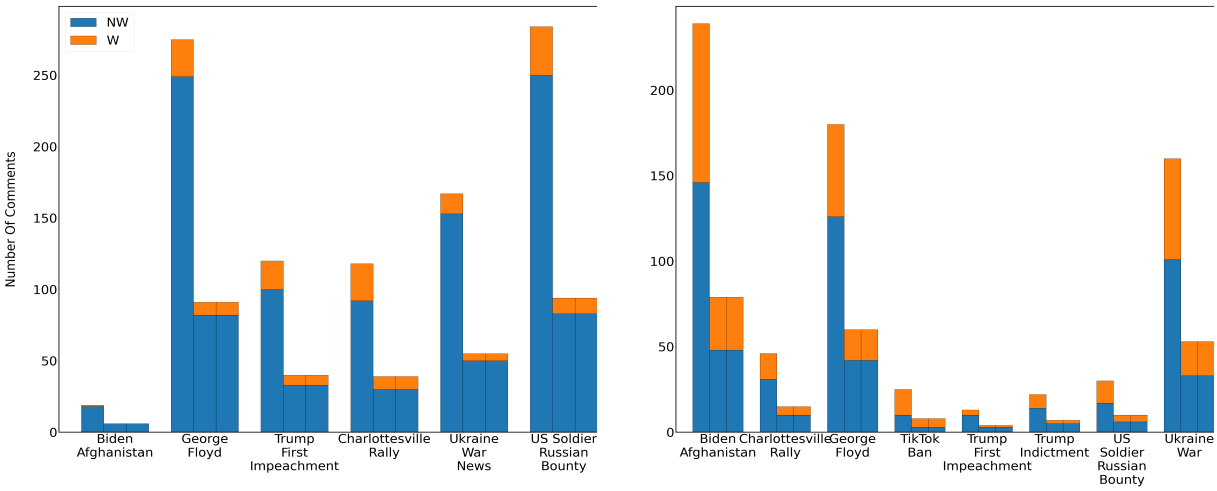
What are the instances? (1) TQ$^+_{\text{YT}}$: Each instance is a comment written by a YouTube user. Each comment is provided along with a link to the corresponding YouTube video and its title, and labeled as *whataboutism* or not (1/0). (2) TQ$^+_{\text{TW}}$: Each instance consists of a tweet written by a Twitter user and a comment written by a different Twitter user in response to that tweet. A link to the corresponding tweets is provided. Each comment is labeled as *whataboutism* or not (1/0).

For both collections, three annotators worked independently to generate the labels. Since multiple annotator judgments are obtained before the majority vote, we take cognizance of the recommendations made by Prabhakaran et al. (2021), and provide the individual annotator labels in the released datasets for flexible future use.

Are relationships between instances made explicit in the data?
TQ$^+$ YouTube: There are no explicit relationships between any two comments, other than their correspondence with the same YouTube video.
TQ$^+$ Twitter: There is no direct connection between any two comments, except for the fact that they both correspond to the same tweet.

How many instances of each type are there?
TQ$^+$ YouTube: There are $1,642$ instances in total, with 202 comments labeled as whataboutism (after aggregation via majority voting).
TQ$^+$ Twitter: There are $1,202$ instances in total, with 508 comments labeled as whataboutism.

Is everything included, or does the data rely on external resources? Everything is included. Due to the dynamic nature of social media and potential policy changes, however, future use of the data collection script may not gather the same data even if the comments remain on the platform.



(a) $\text{TQ}^{+}_{\text{YT}}$ (b) $\text{TQ}^{+}_{\text{TW}}$

Figure 5: Labeled data distribution, spanning six divisive sociopolitical topics for the collection of YouTube comments, and eight such topics for the collection of Twitter responses. Each topic is partitioned into three sets (training, validation, and test), each with comments labeled as whataboutism (W) or not (NW).

Are there recommended data splits or evaluation measures? The dataset comes with specified training, validation, and test splits (80%, 5%, and 15%, respectively). Due to class imbalance in the naturally occurring label distribution, the recommended measures are macro average precision, recall, and F_1 scores.

A.2 Data preprocessing

Comments in any language other than English (en-US) were discarded. Non-linguistic characters such as emojis were removed from each instance.

A.3 Dataset distribution and maintenance

The dataset is distributed together with the code, under the MIT license. There are no fees or access/export restrictions on this dataset.

A.4 Legal & ethical considerations

If the dataset relates to people, or was generated by people, were they informed about the data collection? The data was collected from public web sources, through publicly available APIs. The authors of the comments collected in the dataset are presumably aware that their posts are public, but there was no mechanism of informing them explicitly about the development of the dataset.

Ethical review applications or approvals: N/A.

Were there any provisions of privacy guarantees? No. Neither the dataset nor the data collection process includes any names or user handles.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? TQ^{+} does not contain any personal data of the authors of the collected comments. Further, it does not contain sensitive or confidential information.

Does the dataset contain information that might be considered inappropriate or offensive? Some YouTube comments and Twitter posts may contain inappropriate or offensive language, but their presence is negligible. For example, in a random sampling of 273 YouTube comments (7.8% of the complete dataset), the authors of this publication found no such instance.

B Proof of near-orthogonality

Formal proof of the statement (§4.2): Cosine similarity may present challenges in high-dimensional vector spaces, as it becomes highly probable that any two vectors are nearly orthogonal.

Theorem 1 (Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $|X_i| \leq 1$ for all i. Let $X = \sum_{i \in [n]} X_i$, and let σ^2 denote the variance of X_i. Then,*

$$\mathbb{P}[|X| \geq k\sigma] \leq 2e^{-k^2/4n}$$

Definition 1. *A d-dimensional **ball** is defined as $B_d = \{(x_1, \ldots, x_d) : \|\mathbf{x}\|_2 \leq 1\}$.*

Definition 2. *A vector $\mathbf{g} \in \mathbb{R}^d$ is a **Gaussian vector** if each $g_i\,(1 \leq i \leq d)$ is a uniformly and independently chosen $\mathcal{N}(0,1)$ random variable.*

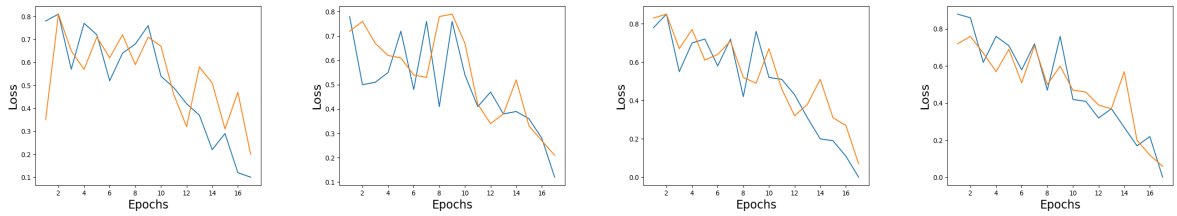


Figure 6: Training (■) and validation (■) losses over all epochs for the MINA-enhanced models reported in Table 2. From left to right: SBERT and RoBERTa$_{\text{Irony}}$ on YouTube comments; and then, SBERT and RoBERTa$_{\text{Irony}}$ on Twitter comments. Consistent trends between training and validation losses, where both decrease without significant divergence, indicate that a model is effectively generalizing to unseen data and not overfitting to the training set.

Theorem 2. *Let* $\mathbf{v}$ *be a unit vector in* $\mathbb{R}^d$*, and let* $\mathbf{x} = (x_1, \ldots, x_d)$ *be a Gaussian vector on the surface of* B_d *obtained by choosing each* $x_i \in \{\pm 1\}$ *and then normalizing via multiplication by* $1/\sqrt{n}$*. Further, let* X *denote the random variable* $\mathbf{v} \cdot \mathbf{x} = \sum_{1 \leq i \leq d} v_i x_i$*. Then,*

$$\mathbb{P}\left(|X| \geq \epsilon\right) \leq 2e^{-d\epsilon^2/4}$$

Proof.

$$\begin{aligned}
\mu &= \mathbb{E}\left[(X)\right] = \mathbb{E}\left[\left(\sum_{i=1}^{d} v_i x_i\right)\right] \\
&= \sum_{1 \leq i \leq d} \mathbb{E}\left[v_i x_i\right] = 0. \\
\sigma^2 &= \mathbb{E}\left[\left(\sum_{i=1}^{d} v_i x_i\right)^2\right] \\
&= \mathbb{E}\left[\sum_{1 \leq i \leq j \leq d} v_i v_j x_i x_j\right] \\
&= \sum_{1 \leq i \leq d} v_i^2 \mathbb{E}[x_i^2] + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} v_i v_j \mathbb{E}[x_i x_j].
\end{aligned}$$

Since $\mathbb{E}(x_i^2) = 1/n$, and if $i \neq j$, $\mathbb{E}(x_i x_j) = 0$,

$$\sigma^2 = \sum_{i=1}^{d} \frac{v_i^2}{d} = \frac{1}{d}.$$

By Bernstein's inequality (Theorem 1),

$$\mathbb{P}\left(|X| \geq \epsilon\right) \leq 2e^{-\epsilon^2 \cdot d/4}$$

□

Thus, if two unit vectors $\mathbf{x}$ and $\mathbf{y}$ are chosen at random from $\mathbb{R}^d$, and the random variable X denotes the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$, the upper bound of Theorem 2 shows that for high values of d, the inner product will have a very small value with high probability. In other words, in high-dimensional vector spaces, two randomly picked vectors are *nearly orthogonal* with high probability.

C Training and Validation Loss Analysis

This appendix presents the learning curves for the models enhanced with MINA, illustrating the training and validation loss over all epochs. These plots, shown in Fig. 6, depict their performances and generalization capabilities, addressing potential concerns about overfitting due to the moderate size of the TQ$^+$ collections.