Retrieving false claims on Twitter during the Russia-Ukraine conflict

Valerio La Gatta*
Information Sciences Institute,
University of Southern California
Los Angeles, California, USA
University of Naples Federico II
Naples, Italy
valerio.lagatta@unina.it

Chiyu Wei*
Information Sciences Institute
University of Southern California
Los Angeles, California, USA
chiyuwei@isi.edu

Luca Luceri
Information Sciences Institute
University of Southern California
Los Angeles, California, USA
lluceri@isi.edu

Francesco Pierri
Polititecnico di Milano
Milan, Italy
Information Sciences Institute,
University of Southern California
Los Angeles, California, USA
francesco.pierri@polimi.it

Emilio Ferrara
Information Sciences Institute
University of Southern California
Los Angeles, California, USA
ferrarae@isi.edu

ABSTRACT

Nowadays, false and unverified information on social media sway individuals' perceptions during major geo-political events and threaten the quality of the whole digital information ecosystem. Since the Russian invasion of Ukraine, several fact-checking organizations have been actively involved in verifying stories related to the conflict that circulated online. In this paper, we leverage a public repository of fact-checked claims to build a methodological framework for automatically identifying false and unsubstantiated claims spreading on Twitter in February 2022. Our framework consists of two sequential models: First, the *claim detection* model identifies whether tweets incorporate a (false) claim among those considered in our collection. Then, the *claim retrieval* model matches the tweets with fact-checked information by ranking verified claims according to their relevance with the input tweet. Both models are based on pre-trained language models and fine-tuned to perform a text classification task and an information retrieval task, respectively. In particular, to validate the effectiveness of our methodology, we consider 83 verified false claims that spread on Twitter during the first week of the invasion, and manually annotate 5,872 tweets according to the claim(s) they report. Our experiments show that our proposed methodology outperforms standard baselines for both *claim detection* and *claim retrieval*. Overall, our results highlight how social media providers could effectively leverage semi-automated

approaches to identify, track, and eventually moderate false information that spreads on their platforms.

KEYWORDS

claim detection, claim retrieval, fact-checking, Twitter

1 INTRODUCTION

On 24th February 2022, Russia started its still ongoing invasion of Ukraine, causing unprecedented backlashes for the rest of the world.[1] Soon afterward, concerns were raised about the presence of Russian disinformation campaigns on online social media, aimed at repurposing the invasion as a "special operation" against alleged Nazis in Ukraine, or attempting to blame NATO's expansion for causing the invasion [9, 10, 22, 26, 27]. Russian interference with other countries' democratic processes is anything but new, as extensively reported in the past, especially in the context of the 2016 U.S. Presidential election [3, 16]. During major geo-political events, manual fact-checking represents the leading strategy to debunk false information through domain experts' analyses, crowd-sourcing approaches [30], and semi-automatic systems assessing news truthfulness [7]. However, given the impossibility of manually checking every piece of information circulating online [24], social media providers still struggle with keeping track of and moderating false information that spreads online [21, 25].

Given these premises, in this work, we aim to build a methodological framework to automatically identify false and unsubstantiated claims – verified by news agencies and fact-checking organizations (e.g., Politifact, Snopes) – that were shared on Twitter at the dawn

*V. La Gatta and C. Wei contributed equally to this work.

[1]https://en.wikipedia.org/wiki/2022_Russian_invasion_of_Ukraine

of the Russian invasion of Ukraine. To this end, we collect 83 false claims that were verified in the first weeks of the invasion and annotate 5,872 original tweets based on the claim(s) they discuss. Then, we deploy an automatic pipeline that comprises two models: (i) the *claim detection* model identifies whether an input tweet contains a (false) claim among the 83 present in our collection; and, assuming the input tweet reports a claim, (ii) the *claim retrieval* model ranks those 83 claims according to their relevance with the input tweet. For these models, we leverage modern transformer-based architectures and adopt transfer learning on the annotated dataset to optimize the performance on both tasks.

Previous work to design automatic tools for fact-checking follows two main directions. On the one hand, research extensively focuses on the problem of assessing the truthfulness of pieces of information [19], also supporting the veracity prediction with factual evidence [4, 31]. On the other hand, relatively less research is devoted to the *claim detection* problem, which can be formulated as a ranking task where several input sentences are ranked by check-worthiness [2, 20], or as a binary classification task to predict whether an input sentence constitutes a claim [8, 11, 12]. In this paper, we consider the latter setting as we aim to identify false claims and we are less interested in prioritizing tweets that require fact-checking.

Once verified that the input tweet reports a (false) claim, we aim to retrieve the most relevant verified claim(s) in our collection. Shaar et al. [33] have recently proposed and defined *claim retrieval* as the task of ranking a corpus of verified documents according to their relevance to an input text. Recently, competitors at CheckThat!2021 ([20]) have shown that fine-tuning transformer-based models lead to promising performance improvements with respect to standard information retrieval approaches (e.g., BM25 [29]). Accordingly, Mansour et al. [17] obtain similar results under multilingual COVID-19 claims. Conversely, considering a political debate scenario, Shaar et al. [32] evaluate the impact of modeling the claim's global and local contexts on the ranking performance. Overall, the above contributions compete on benchmark datasets and assume that input sentences contain a claim. Instead, we define a more challenging scenario where *claim retrieval* is combined with *claim detection* in a pipeline, whose goal is to understand whether a generic tweet reports a false claim related to the specific topic of the Russian invasion of Ukraine.

Contributions of this work. To summarize, our contributions are as follows:

- We collect 83 false claims that spread on Twitter during the first weeks of the invasion and we manually annotate 5,872 original tweets to determine whether they discuss a (false) claim and, if so, which claim(s) they discuss.
- We develop an automatic pipeline to detect and retrieve tweets discussing any of the 83 false claims in our collection. In particular, our models are based on modern transformer-based architectures and perform, in order, *claim detection* and *claim retrieval*.
- We show the effectiveness of the proposed approach in retrieving the claim(s) that are referenced in the input tweet. In addition, we also observe how our models generalize to new claims that were unseen during the training process.

2 METHODOLOGY

In this section, we describe the data collected for the analysis and the methodology employed to annotate tweets and their corresponding claims. Then, we present our methodological framework by formally defining the *claim detection* and the *claim retrieval* tasks and corresponding models.

2.1 Data Collection

To identify false and unsubstantiated claims that circulated online during the same period, we rely on the Russia-Ukraine Conflict-Misinfo Dashboard[2], which provides a collection of true and false claims and rumors verified by fact-checking outlets such as *USA Today* and *Snopes*. Specifically, we collect 83 English false claims that were verified in the period 22nd February - March 1st.

In addition, we leverage an existing dataset[3] of tweets matching over 30 conflict-related keywords in English, Russian and Ukrainian language, which were identified through a snowball sampling approach, collected via Twitter's Filter v1.1 Streaming API[4]. In particular, we focus on English-language tweets shared during the initial weeks of the invasion, i.e., from February 22nd, 2022 to March 8th, 2022. Note that we consider the tweets posted until one week after the above-mentioned false claims were verified (March 1st) so as to capture their propagation on Twitter. Overall, in this observation period, the collected dataset contains more than 2M English tweets with original content, i.e., original tweets, replies, and quotes. We purposely exclude retweets as their textual content is exactly the same.

2.2 Data Annotation

Given the collected fact-checked information, our goal is to find tweets reporting such verified claims. However, as manually annotating the whole corpus of tweets in our dataset is not a suitable solution, we deploy a machine learning-based annotation strategy to maximize the likelihood of finding tweets related to one of the claims under analysis.

Specifically, we first use the RoBERTa transformer [15] to extract the vector embeddings of both claims and tweets, and then we compute the cosine similarity between each claim and tweet in our data, retaining the top-100 most similar tweets for each claim. Consequently, we end up with 8,300 unique tweet-claim pairs that were inspected via a *manual* labeling process. It is worth noting that the choice of the RoBERTa transformer depends on the higher similarity scores provided by this model with respect to other transformers (*ms-marco-MiniLM-L-4-v2* and *quora-roberta-base*), which allows us to maximize the chance of finding a matching pair.

Next, we annotate the above-mentioned tweet-claim pairs considering a strict definition of relevance, i.e., the tweet discusses the claim if it explicitly mentions the same entities and events reported in the fact-checked information. Table 1 shows some examples of false claims and matched tweets, which highlight how the matched tweet can discuss a claim without expressing any stance (examples no. 1 and 2) or can support or refute the claim (examples no. 3 and 4, respectively). Finally, it is worth noting that the number of unique
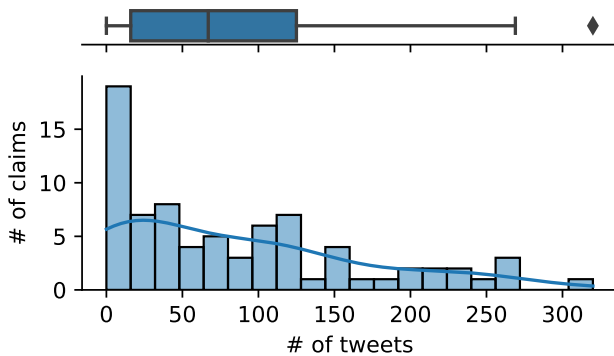
---

[2]www.shorturl.at/puN37
[3]Reference omitted for blind review.
[4]https://developer.twitter.com/en/docs/twitter-api/v1

Table 1: Examples of some tweet-claim pairs annotated in the dataset

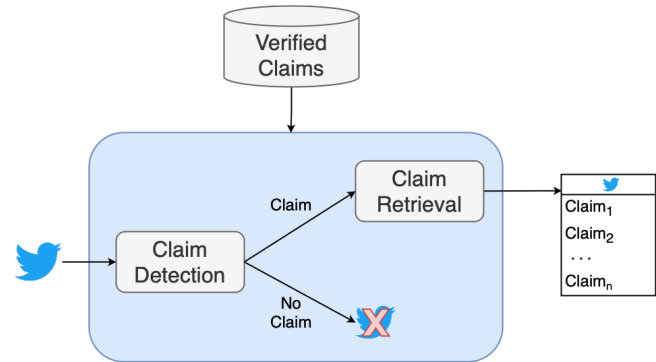| No.	Claim	Tweet
1	Russian President Vladimir Putin threatened India against getting involved in the Ukraine crisis.	Putin has warned India that don't try to interfere in their matter, otherwise be ready to face the consequences
2	The President Of Ukraine, Volodymyr Zelenskyy, Is On The Ground With His Fellow Troops	Volodymyr Zelenskyy the president of Ukraine has decided to stay behind and fight among his people against the Russian army send to kyiv […]
3	The Russian armed forces are not striking at the cities of Ukraine; they are not threatening the civilian population.	It is clear that the Russian army does not want to harm civilians, its strikes were directed only at military targets, […] life seems almost normal in Kiev.
4	The Russian armed forces are not striking at the cities of Ukraine; they are not threatening the civilian population.	Russian forces continue strikes in multiple cities […]. This is premeditated mass murder and must be responded to as such.

Figure 1: Distribution of the number of tweets with respect to the number of claims

Figure 2: Our proposed methodological framework: the *claim detection* model detects whether the input tweet reports a fact-checked (false) claim. If a claim is detected, the *claim retrieval* model retrieves the most relevant claims (within the corpus of "Verified Claims") related to the tweet.

tweets is different from the number of pairs because a tweet can be related to multiple claims. In particular, we find 5,872 unique tweets in the ranking of the 83 claims.

Our manual annotation results in 2,359 (out of 5,872 – 40.2%) tweets associated with at least one claim. Figure 1 shows the distribution of the number of tweets with respect to the number of claims: most of the tweets are related to less than five claims and only 13 tweets discuss more than 10 claims. Overall, each claim has at least one matching tweet and the most matched claim has 100 related tweets. Conversely, we find 3,513 (out of 5,872 – 59.8%) tweets that do not report any claim.

2.3 Our Framework

Figure 2 describes the two-step pipeline proposed in our methodological framework. The first step is characterized by a *claim detection* model, which verifies whether an input tweet discusses a (false) claim included in a predetermined set of already verified claims. Next, if the input tweet actually discusses a false claim in the *Verified Claims* corpus, a *claim retrieval* model ranks the claims in the corpus according to their relevance with the input tweet. It is worth noting that, even if we target false claims, our methodology

does not depend on the claim's truthfulness and can be adopted to retrieve tweets matching any claim associated with a specific topic.

2.3.1 Claim Detection. Overall, we frame the claim detection problem as a binary classification task: given an input tweet t, the objective is to establish whether t contains a false claim among the ones collected in the corpus of *Verified Claims*. Therefore, we leverage BERT-based language models [5], which have proven their effectiveness in several text classification tasks [14] due to the extensive knowledge they acquire during the pre-training process. In particular, we fine-tune the *BERT-base-cased* with the Huggingface library[5] for five epochs using Adam optimizer, categorical cross-entropy as a loss function, and 20 tweets as batch size.

In addition, assuming a realistic scenario where only a minority of tweets report specific factual events related to the war [1] (i.e., 2.3k out of 5.9k tweets identified by our annotation procedure), we evaluate our model by performing random oversampling of the minority class (i.e., tweets with a specific claim). In other words, we

---

[5]https://huggingface.co

randomly replicate tweets containing a claim so as to have the same number of tweets belonging to the positive and negative classes.

2.3.2 Claim Retrieval. Following the cascade architecture of our framework (see Fig. 2), we are now interested in finding the most relevant claims associated with the input tweets that report a claim, as determined with the *claim detection* model.[6] Formally, we leverage the information retrieval formulation related to the task of identifying fact-checked claims [33]: given an input tweet t containing a claim, rank a set of verified (false) claims $\{c_1, c_2, \cdots, c_n\}$ based on their relevance to t.

In particular, we consider the tweet-claim pairs as input of the system and use the similarity score to rank the verified claims. However, performing the supervised training process, we need to build the negative class by finding instances of unrelated tweet-claim pairs. This problem is common when addressing information retrieval tasks [6], and several (negative) sampling strategies have been evaluated across different ranking tasks [13, 35]. Among the others, we find that random negative sampling yields good results in our use scenario, i.e., for each tweet we randomly select 10 unrelated claims.

Next, as for the claim detection model, we leverage a transformer-based architecture. Indeed, when optimized for ranking tasks, pre-trained language models have proven their effectiveness in a variety of information retrieval tasks, including semantic search [23] and semantic textual similarity [28]. Specifically, we finetune the *stsb-roberta-base* cross-encoder within the SBERT library[7].

We train the model for three epochs using categorical cross-entropy loss and 16 tweet-claim pairs as batch size.

3 EXPERIMENTS

This section presents the experiments we have performed to evaluate our framework. All experiments are conducted on a machine equipped with CPU Intel Xeon-4116, RAM 32 GB, and one NVIDIA A100.

3.1 Evaluation Data and Metrics

In this section, we describe both the data sets and metrics used to evaluate our methodological framework. We distinguish these for the claim detection and claim retrieval models, given the different prediction goals of the two tasks. For the claim detection task, we consider the annotated tweets described beforehand. Our dataset includes 2,359 (out of 5,872 – 40,2%) tweets reporting at least one (false) claim and 3,513 (out of 5,872 – 59,8%) tweets that do not incorporate any claim. Given the classification task, we measure the performance of our models considering binary classification metrics such as precision, recall, f-score, and accuracy.

For the claim retrieval task, we consider the 2,359 tweets related to the 83 claims gathered from the Russia-Ukraine ConflictMisinfo Dashboard. We build up a dataset of 40,007 tweet-claim pairs as previously discussed, consisting of 3,637 *positive* tweet-claim pairs and 36,370 *negative* tweet-claim pairs. Consistently with previous work [33, 34], we evaluate the performance of the claim retrieval

---

[6]Note that if the claim detection model assesses the input tweet to be unrelated to any claim, the input tweet is automatically discarded and not considered in the claim retrieval model.
[7]https://www.sbert.net

Table 2: Claim detection: performance with and without random oversampling

Metric	No Oversampling	With Oversampling
Precision	79.90%	**81.39%**
Recall	80.19%	**80.24%**
F1-score	79.35%	**80.57%**
Accuracy	80.11%	**81.59%**

model with the HitRatio@K, i.e., whether the correct claim (i.e., the claim matching the tweet) is among the top-k results of the ranking. It is worth noting that, since most of the tweets match only one claim, HitRatio@K is almost equal to Recall@K [34]. On the one hand, from the end-user's perspective, the HitRatio on lower values of k (e.g. $k \in \{1, 3, 5\}$) might be indicative of the system utility in easing manual fact-checkers works, i.e., experts would spot in real-time if the top-ranked results are relevant to the input claim. On the other hand, the HitRatio on higher values of k (e.g. $k \in \{10, 20, 50\}$) should be considered in offline settings and/or in an automated fact-checking pipeline where results should be further processed as evidence for the veracity prediction.

3.2 Experimental Protocol

For each task, we perform a 5-fold cross-validation to evaluate the performance of our models. In particular, we consider two evaluation settings:

- Leave Tweet Out (LTO) Assessment: for both the claim detection and retrieval tasks, we ensure that the tweets in the training, validation, and testing sets do not overlap.
- Leave Claim Out (LCO) Assessment: for the claim detection task, we ensure that the tweets in the training, validation, and testing sets match different claims. For the claim retrieval task, we ensure that the claims within the tweet-claim pairs in the training, validation, and testing sets do not overlap.

The LTO assessment refers to the classical evaluation of supervised machine learning approaches and is consistent with experiments from previous works [20, 33, 34]. The LCO assessment focuses on the generalization of the models to unseen claims and is close to a real-world scenario. Indeed, when deployed "in the wild", both claim detection and retrieval should be performed on tweets and claims that the models have not seen during the training process.

When evaluating the performance of the claim detection task, we compare our BERT-based model to a TF-IDF baseline [12], also assessing the effectiveness of the oversampling strategy detailed in Section 2.3.1. For the claim retrieval task, we compare the ranking performance of our BERT-based cross-encoder with respect to Sentence-BERT [28].

3.3 Results

3.3.1 Claim Detection. Table 2 shows classification performance with and without the random oversampling strategy. In particular, we can notice that oversampling the minority class improves the predictive power of the claim detection model for every classification metric.

Table 3: Claim detection: performance comparison per class, and their 95% confidence interval, between TF-IDF baseline and our approach (bold indicates best on average, * indicates statistical significance ($p < 0.01$)

Metric	Claim (Positive class)		No Claim (Negative class)	
	TF-IDF	Ours	TF-IDF	Ours
Precision	67.88% ± 1.57%	**79.75% ± 3.05%***	80.97% ± 2.92%	**83.02% ± 1.87%**
Recall	73.01% ± 4.99%	**73.31% ± 2.09%**	76.84% ± 0.81%	**87.17% ± 3.85%***
F1-score	70.33% ± 3.05%	**76.17% ± 1.71%***	78.84% ± 1.40%	**84.97% ± 1.39%***

Table 4: Claim detection: LTO and LCO assessment

Metric	Settings	
	LTO	LCO
Precision	81.39% ± 1.14%	78.07% ± 1.48%
Recall	80.24% ± 1.36%	77.58% ± 1.67%
F1-score	80.57% ± 1.38%	77.63% ± 1.60%
Accuracy	81.59% ± 1.42%	78.03% ± 1.56%

Table 3 shows the performance marginalized for each class of the *claim detection* task. In particular, our BERT-based model outperforms the TF-IDF baseline for both classes and all metrics. Overall, we can observe that performance is better on the negative class, i.e., the system has higher precision and recall in the detection of tweets that do not report any claim. We do not report the accuracy metric as it is equivalent to the recall computed on the single class.

Finally, Table 4 shows the aggregated performance under the LTO and LCO evaluation settings. We can observe that LCO evaluation is more challenging since the model has to deal with claims that it has never seen during the training process.

3.3.2 Claim Retrieval. Table 5 shows the ranking performance of our claim retrieval model in comparison with sentence-BERT [28]. In both evaluation settings (i.e., LTO and LCO), our model outperforms the baseline, and the performance gap is even larger when considering the top positions in the ranking ($k \in \{1, 3\}$). This result highlights that our system can actually ease fact-checkers' work because it can spot tweets that discuss false claims that were already verified. In addition, we can notice again that the performance under LTO settings is slightly better than the ones under LCO settings, even if the gap is smaller with respect to what we found in the claim detection task.

3.3.3 Evaluation in the wild. Finally, we apply our claim detection and retrieval models on the whole Twitter dataset to assess how our methodology performs in a real scenario. In this case, we do not have any annotation for the LTO and LCO settings, and thus we cannot perform a formal quantitative analysis. However, we randomly sample 100 (resp. 100) tweets assessed as reporting a claim (resp. not reporting any claim) by the claim detection model. Then, we manually verify whether the models make the correct prediction, i.e., whether the tweet discusses a false claim and, if so, to which false claim it was related.

Out of the 100 tweets that are predicted by the *claim detection* model as not related to any (false) claim, we find 12 misclassified instances, i.e., tweets that are actually related to a false claim. This

result confirms that the *claim detection* model has a high precision (> 80%) on the negative class (see Table 3 for further details).

On the contrary, out of the 100 tweets that are predicted by the *claim detection* model as discussing a (false) claim, we find 64 false positives, i.e., tweets do not discuss any false claim but generally report information about the Ukraine-Russia war. This result confirms the performance differences between the positive and negative classes (see Table 3) of the *claim detection* model, and probably depends on data imbalance, i.e., the number of tweets with no claim is higher than the number of tweets reporting a claim.

In addition, we apply the *claim retrieval* model for the 36 tweets that were actually related to some false claims. In particular, we retrieve the top-3 relevant false claims and find that the correct matching claim was retrieved 34 times (out of 36).

Overall, these results confirm the promising performance of the *claim retrieval* model but highlight the limitations of the *claim detection* model, which overestimates the number of tweets discussing false claims.

4 CONCLUSIONS AND FUTURE WORK

We presented a methodological framework to detect over 80 false and unsubstantiated claims that were shared on Twitter during the first week of the conflict. Our framework first performs *claim detection* to identify whether an input tweet contains a (false) claim or not. Then, assuming the input tweet reports a claim, the system performs *claim retrieval* to rank a set of already-verified false claims according to their relevance with the input tweet. When fine-tuning modern BERT-based models, our methodology achieves promising performance to automate both tasks. Indeed, our models show good generalization capabilities, i.e., they reach good performance even when the claims that need to be detected were never seen during the training process.

Despite the promising performance, we highlight some limitations of our approach. First, we considered a limited set of false claims because of the short observation window, i.e., the first week of the invasion. However, when focusing on a longer time period, the number of verified claims increases as well as what was considered false at the beginning of the conflict could become true afterward (e.g., NATO's members providing Ukraine with military weapons). Second, when collecting tweets discussing false claims, we did not consider their stance and implicitly assumed they support or divulge the claims [18]. However, we actually find some tweets that reported (false) claims debating their veracity.

There is a number of avenues to explore in future research. First, we aim to extend our analysis to multimodal data, i.e., images and videos that were shared within the tweets could help our system

Table 5: Claim retrieval: performance comparison, and their 95% confidence interval, between the sentence-BERT baseline and our approach (bold indicates best on average, * indicates statistical significance ($p < 0.01$)

Setting	Model	HitRatio@k				
		$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$
LTO	Sentence-BERT	$85.25\% \pm 2.07\%$	$94.87\% \pm 1.69\%$	$97.24\% \pm 0.96\%$	$98.77\% \pm 0.43\%1$	$99.27\% \pm 0.39\%$
	Ours	$\mathbf{86.05\% \pm 0.95\%}$	$\mathbf{96.35\% \pm 0.71\%}^*$	$\mathbf{98.04\% \pm 0.57\%}^*$	$\mathbf{99.27\% \pm 0.36\%}$	$\mathbf{99.78\% \pm 0.11\%}^*$
LCO	Sentence-BERT	$77.60\% \pm 0.1196$	$95.68\% \pm 6.74\%$	$98.01\% \pm 3.66\%$	$99.63\% \pm 0.66\%$	$99.78\% \pm 0.00\%$
	Ours	$\mathbf{82.25\% \pm 10.81\%}^*$	$\mathbf{96.42\% \pm 2.59\%}$	$\mathbf{98.26\% \pm 1.45\%}$	$\mathbf{99.88\% \pm 0.02\%}$	$\mathbf{99.96\% \pm 0.00\%}$

to improve its classification and ranking performance. Second, we plan to apply our methodology "into the wild" to investigate the diffusion of the most shared (false) claims on Twitter and unveil the communities that support or are most susceptible to false narratives. Third, we will deploy our framework within a fact-checking pipeline and assess the extent to which such a system could improve manual fact-checking by allowing debunkers to focus on brand-new claims and ignoring similar claims that have been already verified.

ACKNOWLEDGMENTS

REFERENCES

[1] Liesbeth Allein and Marie-Francine Moens. 2020. Checkworthiness in Automatic Claim Detection Models: Definitions and Analysis of Datasets. *CoRR* abs/2008.08854 (2020). arXiv:2008.08854 https://arxiv.org/abs/2008.08854

[2] Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic Fact-Checking Using Context and Discourse Information. *J. Data and Information Quality* 11, 3, Article 12 (may 2019), 27 pages. https://doi.org/10.1145/3297722

[3] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. IEEE, 258–265.

[4] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: Generative Evidence Retrieval for Fact Verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. https://doi.org/10.1145/3477495.3531827

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[6] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic Models for the First-Stage Retrieval: A Comprehensive Review. *ACM Trans. Inf. Syst.* 40, 4, Article 66 (mar 2022), 42 pages. https://doi.org/10.1145/3486250

[7] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. https://doi.org/10.1162/tacl_a_00454

[8] Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic Encapsulation and Semantic Amalgamation Based Generalised Claim Detection from Online Content. *CoRR* abs/2101.11891 (2021). arXiv:2101.11891 https://arxiv.org/abs/2101.11891

[9] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2022. " A Special Operation": A Quantitative Approach to Dissecting and Comparing Different Media Ecosystems' Coverage of the Russo-Ukrainian War. *arXiv preprint arXiv:2210.03016* (2022).

[10] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2022. Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War On Reddit. *arXiv preprint arXiv:2205.14484* (2022).

[11] Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting Check-Worthy Factual Claims in Presidential Debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1835–1838. https://doi.org/10.1145/2806416.2806652

[12] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats* 2, 2, Article 14 (apr 2021), 16 pages. https://doi.org/10.1145/3412869

[13] Jia Li, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Sampling Matters! An Empirical Study of Negative Sampling Strategies for Learning of Matching Models in Retrieval-based Dialogue Systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1291–1296. https://doi.org/10.18653/v1/D19-1128

[14] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A Survey of Transformers. *CoRR* abs/2106.04554 (2021). arXiv:2106.04554 https://arxiv.org/abs/2106.04554

[15] 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. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[16] Luca Luceri, Silvia Giordano, and Emilio Ferrara. 2020. Detecting troll behavior via inverse reinforcement learning: A case study of Russian trolls in the 2016 US election. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 417–427.

[17] Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2022. Did I See It Before? Detecting Previously-Checked Claims over Twitter. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 367–381. https://doi.org/10.1007/978-3-030-99736-6_25

[18] Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 651–662. https://doi.org/10.1609/icwsm.v16i1.19323

[19] Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. *CoRR* abs/2103.07769 (2021). arXiv:2103.07769 https://arxiv.org/abs/2103.07769

[20] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mücahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. *CoRR* abs/2109.12987 (2021). arXiv:2109.12987 https://arxiv.org/abs/2109.12987

[21] Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. 2022. The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter. In *14th ACM Web Science Conference 2022*. 348–358.

[22] Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2022).

[23] Manish Patel. 2019. TinySearch - Semantics based Search Engine using Bert Embeddings. *CoRR* abs/1908.02451 (2019). arXiv:1908.02451 http://arxiv.org/abs/1908.02451

[24] Francesco Pierri and Stefano Ceri. 2019. False news on social media: a data-driven survey. *ACM Sigmod Record* 48, 2 (2019), 18–27.

[25] Francesco Pierri, Matthew R DeVerna, Kai-Cheng Yang, David Axelrod, John Bryden, and Filippo Menczer. 2023. One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study. *Journal of Medical Internet Research* 25 (2023), e42227.

[26] Francesco Pierri, Luca Luceri, and Emilio Ferrara. 2022. How does Twitter account moderation work? Dynamics of account creation and suspension during major geopolitical events. *arXiv preprint arXiv:2209.07614* (2022).

[27] Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2023. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. In *WebSci'23 – 15th ACM Web Science Conference*.

[28] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[29] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389. https://doi.org/10.1561/1500000019

[30] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts?. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 1736–1746. https://doi.org/10.1145/3511808.3557279

[31] Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving Evidence Retrieval for Automated Explainable Fact-Checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*. Association for Computational Linguistics, Online, 84–91. https://doi.org/10.18653/v1/2021.naacl-demos.10

[32] Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The Role of Context in Detecting Previously Fact-Checked Claims. In *Findings of the Association for Computational Linguistics: NAACL-HLT 2022 (NAACL-HLT '22)*. Seattle, Washington, USA.

[33] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3607–3618. https://doi.org/10.18653/v1/2020.acl-main.332

[34] Nguyen Vo and Kyumin Lee. 2020. Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7717–7731. https://doi.org/10.18653/v1/2020.emnlp-main.621

[35] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *CoRR* abs/2007.00808 (2020). arXiv:2007.00808 https://arxiv.org/abs/2007.00808