Nexus at ArAIEval Shared Task: Fine-Tuning Arabic Language Models for Propaganda and Disinformation Detection

Yunze Xiao[1], **Firoj Alam**[2]
[1]Carnegie Mellon University in Qatar, Doha, Qatar
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
yunzex@andrew.cmu.edu,falam@hbku.edu.qa,

Abstract

The spread of disinformation and propagandistic content poses a threat to societal harmony, undermining informed decision-making and trust in reliable sources. Online platforms often serve as breeding grounds for such content, and malicious actors exploit the vulnerabilities of audiences to shape public opinion. Although there have been research efforts aimed at the automatic identification of disinformation and propaganda in social media content, there remain challenges in terms of performance. The ArAIEval shared task aims to further research on these particular issues within the context of the Arabic language. In this paper, we discuss our participation in these shared tasks. We competed in subtasks 1A and 2A, where our submitted system secured positions 9th and 10th, respectively. Our experiments consist of fine-tuning transformer models and using zero- and few-shot learning with GPT-4.

Propagandistic text:
مقطع فيديو جديد يظهر محاولة الزميلة #شيرين_أبو_عاقلة الاحتماء من رصاص الاحتلال الإسرائيلي بالجدار والشجرة في موقعها قبيل اغتيالها\\n#الأخبار LINK

Translation: A new video clip shows the attempt of colleague #Sherine_Abu_Aqla to take refuge from the bullets of the Israeli occupation using a wall and a tree at her location before her assassination\n#News LINK

Disinformative (hate speech) text:
@NourOusama@ therachellekayr البابا تبعكم فيه كورونا وعم ينشر الكورنا في لبنان عن طريق المسيحيين اللبنانيين الراجعين من روما والي ماتوا كلهم يعنى كلهم منكم مش من مناطق الشيعه

Translation: @therachellekayr @NourOusama Your Pope has Corona and is spreading Corona in Lebanon through the Lebanese Christians returning from Rome and all the people who died, I mean all of them, are from your group, they are not from the Shia area

Figure 1: Examples of propagandistic and disinformative text.

1 Introduction

In various communication channels, propaganda, also known as persuasive techniques, is disseminated through a wide set of methods. These techniques can range from appealing to the audience's emotions—known as the *"emotional technique"* — to employing logical fallacies. Examples of such fallacies include *"straw man"* arguments, which misrepresent someone's opinion; covert *"ad hominem"* attacks; and *"red herrings"*, which introduce irrelevant data to divert attention from the issue at hand (Miller, 1939).

Previous research in this area has taken various approaches to identify propagandistic content. These include assessing content based on writing style and readability levels in articles (Rashkin et al., 2017; Barrón-Cedeno et al., 2019), examining sentences and specific fragments within news articles using fine-grained techniques (Da San Martino et al., 2019), as well as evaluating memes for propagandistic elements (Dimitrov et al., 2021a).

Moreover, malicious actors manipulate media platforms to shape public opinion, disseminate hate speech, target individuals' subconscious minds, spread offensive content, and fabricate falsehoods, among other. These efforts are part of broader strategies to influence people's thoughts and actions (Zhou et al., 2016; Alam et al., 2022a; Sharma et al., 2022).

In a broader context, the proliferation of such disinformation can pose significant threats to societal harmony and undermine the trust individuals have in reliable sources (Mubarak et al., 2023). Currently, these manipulative strategies are widespread across various online platforms, where they are employed to influence public opinion and distort perceptions, taking advantage of the vulnerabilities of unsuspecting audiences (Oshikawa et al., 2018, 2020).

The far-reaching consequences of misinformation and propaganda include the incitement of prejudices and discriminatory behaviors, as well as the exacerbation of social divisions and polarization (Fortuna and Nunes, 2018; Zampieri et al., 2019, 2020; Da San Martino et al., 2019). In extreme cases, such false narratives can even fuel radicalization, threatening societal stability. Ultimately, the spread of misinformation undermines democracy

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 576–582
December 7, 2023

by depriving citizens of the accurate information needed for informed decision-making (Li et al., 2016). The digital age has expanded the reach of propaganda, subtly influencing individuals' perspectives even in their most private spheres.

Since propaganda can manifest in a variety of forms, detecting it and other types of misinformation has always been a challenging task. This task necessitates a deeper analysis of the context in which the content is presented. Therefore, the goal of the shared task is to advance research by developing methods and algorithms for identifying disinformation and propagandistic content. In Figure 1, we provide examples that depict such content.

In the ArAIEval shared task at ArabicNLP 2023 (Hasanain et al., 2023a), there are two tasks with two subtasks each: *(i)* **Task 1 Persuasion Technique Detection** and *(ii)* **Task 2: Disinformation Detection**. Each has two subtasks. We used pre-trained transformer-based models to fine-tune them on the task specific datasets.

We participated in subtasks 1A and 2A, where we fine-tuned pretrained models to predict whether the texts contain persuasion techniques (1A) or are disinformative (2A). We also explored zero-shot and few-shot learning using GPT-4 to understand its performance for these tasks. Both subtasks in which we participated fall under binary classification settings.

2 Related Work

In this section, we discuss the research related to the automatic detection of persuasion techniques and disinformation.

Over the past few decades, the use of persuasion techniques, often in the form of propaganda, has proliferated on social media platforms, aiming to influence or mislead audiences. This has become a major concern for a wide range of stakeholders, including social media companies and government agencies. In response to this growing issue, the emerging field of "computational propaganda" aims to automatically identify such manipulative techniques across various forms of content—textual, visual, and multimodal (e.g., memes).

Recently, the study by (Da San Martino et al., 2019) curated a variety of persuasive techniques. These range from emotional manipulations, such as using *Loaded Language* and *Appeal to Fear*, to logical fallacies like *Straw Man* (misrepresenting someone's opinion) and *Red Herring* (introducing irrelevant data). The study primarily focused on textual content, such as newspaper articles. In a similar vein, (Da San Martino et al., 2020) organized a shared task on the "Detection of Propaganda Techniques in News Articles." Building on these previous efforts, (Dimitrov et al., 2021b)[1] orchestrated the *SemEval-2021 Shared Task 6 on Detection of Propaganda Techniques in Memes* in 2021. This task had a multimodal setup, integrating both text and images, and challenged participants to construct systems capable of identifying the propaganda techniques employed in specific memes. Efforts have also been made towards multilingual propaganda detection. (Hasanain et al., 2023b) demonstrates that multilingual models significantly outperform monolingual ones, even in languages that are unseen.

While most of these efforts have focused primarily on English, Alam et al. (2022b) organized a shared task on fine-grained propaganda techniques in Arabic to enrich the field of Arabic AI research. This event attracted numerous participants.

In addition to the use of propaganda, malicious social media users frequently disseminate disinformative content—including hate speech, offensive material, rumors, and spam—to advance social and political agendas or to harm individuals, entities, and organizations. To address this issue, the current literature has explored automated techniques for detecting disinformation on social media platforms. For example, the study by Demilie and Salau (2022) investigated the detection of fake news and hate speech in Ethiopian social media. The researchers found that a hybrid approach, combining both deep learning and traditional machine learning techniques, proved to be the most effective in identifying disinformation in that context.

In the field of Arabic social media, numerous researchers have used various approaches for disinformation detection. For example, the study by Boulouard et al. (2022) focused on identifying hate speech and offensive content in Arabic social media platforms. By employing transfer learning techniques, they found that BERT (Devlin et al., 2018) and AraBERT (Antoun et al., 2020) yielded the highest accuracy rates, at 98% and 96%, respectively. Other significant contributions to the area

[1] http://propaganda.math.unipd.it/semeval2021task6/

of Arabic hate speech and offensive content detection include works by Zampieri et al. (2020) and Mubarak et al. (2020).

3 Task and Dataset

As discussed earlier we used the datasets released as a part of the ArAIEval shared task (Hasanain et al., 2023a). We participated in subtask 1A and 2A. They are defined as follows.

Subtask 1A: Given a multigenre (tweet and news paragraphs of the news articles) snippet, identify whether it contains content with persuasion technique. This is a binary classification task.

The data for Subtask 1A is composed of IDs, text, and labels. These labels are either 'true' or 'false', indicating whether the content contains a propagandistic technique. As observed in our analysis, there is a significant skew in the label distribution. As shown in Table 1, only 21% of the data is labeled as 'false,' while the remaining 79% carries a 'true' label. This imbalance in classes could introduce challenges during the training phase. Furthermore, we found that 64.9% of the data originates from paragraphs, while the remaining 35.1% is sourced from tweets.

Subtask 2A: Given a tweet, categorize whether it is disinformative. This is a binary classification task.

The data format for Subtask 2A is identical to that of Subtask 1A. Similar to Subtask 1A, this subtask also shows a skewed label distribution. Specifically, only 18.8% of the data is tagged as **disinfo**, while the remaining 79% carries the **no-disinfo** tag, as can be seen in Table 1. This imbalance in class distribution could present challenges during the model training process.

For our experiments, we used the same training, development, and test datasets as provided by the organizers. Details on the data distribution can be found in Table 1.

Evaluation Measures: The official evaluation metric for Subtask A is Micro-F1, while for Subtask B, it is Macro-F1.

4 Methodology

4.1 Pre-trained Models

Given that large-scale pre-trained Transformer models have achieved state-of-the-art performance

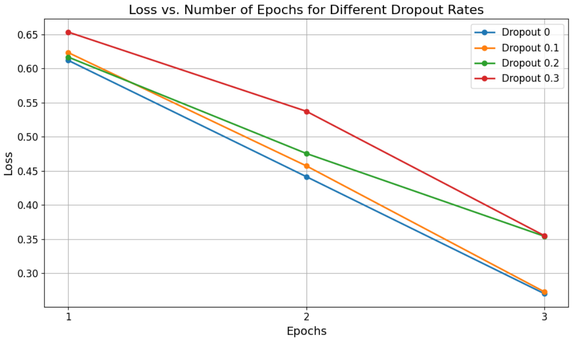

Figure 2: Loss per epoch with different dropout rate.

for several NLP tasks. Therefore, as deep learning algorithms, we used deep contextualized text representations based on such pre-trained transformer models. We used AraBERT (Antoun et al., 2020), MarBERT (Abdul-Mageed et al., 2021) and Qarib (Abdelali et al., 2021) due to their promising performance in other Arabic NLP tasks.

Consequently, text preprocessing was done using the AraBERT preprocessor with the default configuration. Hyperparameters were tuned and optimized through the use of randomized grid search. The chosen configuration for the task involved a maximum tokenization length of 128, a batch size of 16, running for a total of 3 epochs during training, with a learning rate set at 4e-5, and utilizing the AdamW optimizer. As a loss function, we used cross-entropy loss:

$$\text{CrossEntropyLoss} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \cdot \log(p_{ij})$$

where, N is the number of samples, C is the number of classes, y_{ij} is the ground truth label (1 if the sample i belongs to class j, 0 otherwise), and p_{ij} is the predicted probability of sample i belonging to class j.

After closely examining the weights in the cross-entropy loss function, we chose to assign four times the weight to the 'false' tag compared to the 'true' tag, resulting in a weight array of [1.0, 4.0] for the cross-entropy loss.

Additionally, we observed that the dataset is highly imbalanced. Incorporating a dropout layer improved the model's performance. To optimize this, we experimented with varying dropout rates and monitored the corresponding loss across different epochs, as illustrated in Figure 2.

Surprisingly, the models with lower dropout rates, which exhibited lower loss in the final epoch,

	Task 1A		Task 2A	
	Prop	**Non-Prop**	**Disinfo**	**No-Disinfo**
Train	1,918 (79%)	509 (21%)	2,656 (19.8%)	11,491 (81.2%)
Dev	202 (78%)	57 (22%)	397 (18.8%)	1,718 (81.2%)
Test	331 (65.8%)	172 (34.2%)	876 (23.8%)	2,853 (76.2%)
Total	2451	733	3929	15062

Table 1: Class label distribution for task 1A and 2A. Prop. – Contains propagandistic technique; Non-Prop – does not contain any propagandistic technique.

performed worse than those with slightly higher dropout rates. We suspect that the models may have overfitted when using lower dropout rates, resulting in subpar performance on the test set.

4.2 Large Language Models (LLMs)

For the LLMs, we investigate their performance in both in-context zero-shot and few-shot learning settings. This involves prompting and post-processing the output to extract the expected content. We utilized GPT-4 (OpenAI, 2023) in both zero-shot and few-shot settings for both subtasks. To ensure reproducibility, we set the temperature to zero for all settings. Note that for GPT-4, we used version 0314, which was released in June 2023. Our choice of this model was based on its accessibility. For the experiments, we employed the LLMeBench framework (Dalvi et al., 2023), following the prompts and instructions previously studied for Arabic in (Abdelali et al., 2023).

Model	**Dropout**	**Micro F1**		**Macro F1**	
		Dev	**Test**	**Dev**	**Test**
Submission			0.740		0.693
AraBERT	0	0.656	0.625	0.723	0.712
	0.1	0.772	0.704	0.725	0.714
	0.2	0.772	0.692	0.739	0.740
	0.3	n/a	n/a	0.743	0.713
MarBERT	0	0.810	**0.756**	0.707	0.696
	0.1	0.841	0.731	0.745	0.718
	0.2	0.818	0.746	0.769	0.731
	0.3	n/a	n/a	0.737	0.708

Table 2: Results with different dropout rates and submitted system for subtask 1A. n/a refers to the number was not ready at time of preparing the paper.

Model	**Dropout**	**Test**	
		Micro F1	**Macro F1**
Submission	0.2	0.893	0.845
Qarib	0	0.889	0.822
	0.1	0.898	0.844
	0.2	0.903	**0.869**
	0.3	0.897	0.849
MarBERT	0.1	0.898	0.843
	0.2	0.898	0.846
	0.3	0.899	0.849
AraBERT	0	0.802	0.794
	0.1	0.846	0.813
	0.2	0.893	0.846

Table 3: Model performance with different dropout rates and submitted system for subtask 2A (disinformative vs. not-disinformative).

5 Results and Discussion

5.1 Subtask 1A

For this shared task, we were given a dataset containing 504 text entries. We employed the model described in the previous section to predict various labels for each tweet. The final results released by the task organizers indicated that our model achieved a Micro F1 of 0.740 and a Macro F1 of 0.693. In Table 2, we present the performance metrics for our submitted system, comparing them with other models and various dropout rates.

Through our discovery, we realize that MarBERT performed extremely well compared to Arabert. This is expected as MarBERT is trained on tweets, which is very similar to the data provided. Nevertheless, we found it even more surprising that MarBERT's performance dropped after applying the dropout layer. This potentially indicates that the model might be undertrained and we might need to run a few more epochs.

	Shot	Micro F1	Macro F1
Task 1A	0-shot	0.600	0.600
	5-shot	0.614	0.614
Task 2A	0-shot	0.759	0.707
	5-shot	0.852	0.804

Table 4: Results on the test set with zero- and few-shot learning using GPT-4.

5.2 Subtask 2A

For this shared task, we are provided with 3729 entries of text. The model described in the previous section was used to predict various labels for each tweet. The final results released by the task organizers have shown that the model that we have scored 0.7396 in Micro F1 and 0.74 in Macro F1. In Table 3 we have displayed some of our attempts, and after more experiments we are able to achieve higher result.

We noticed that in task2A that qarib outperformed MarBERT, despite both trained using a variety of tweets. This could be the result of better/bigger training set or the result of longer training duration. To discover why, further investigation and experimentation have to be made.

In Table 4, we report the results on the test sets for both tasks with zero and 5-shots learning using GPT-4. It appears that the performances are significantly lower than fine-tuned models. We see an improvement with 5-shots, which was also observed in prior studies (Abdelali et al., 2023). However, such performances are still lower than fine-tuned models. Further studies are required to understand their capabilities as prompt engineering is the key factor to achive a desired results with LLMs.

6 Conclusion and Future work

In this paper, we report on our participation in the ArAIEval 2023 shared task, which focuses on propaganda and disinformation detection. We experimented with various transformer-based models and fine-tuned them for our specific tasks. Despite challenges such as imbalanced data, we optimized our models and achieved commendable results. Our submitted system ranked 9th and 10th in subtasks 1A and 2A, respectively, on the leaderboard. In the future, our research will take advantage of the latest Large Language Models (LLMs) such as Llama, Alpaca, Bloom and more. We plan to do more experiment with data augmentation.

Limitations

Our study primarily focused on fine-tuned transformer-based models and zero-shot and few-shot learning with GPT-4. Given that the dataset is heavily skewed towards certain classes, our study did not address these aspects. However, this will be the focus of a future study.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2023. Benchmarking arabic ai with large language models.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Preslav Nakov, and Giovanni Da San Martino. 2022b. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *In IPM*, 56(5):1849–1864.

Zakaria Boulouard, Mariya Ouaissa, Mariyam Ouaissa, Moez Krichen, Mutiq Almutiq, and Gasmi Karim. 2022. Detecting hateful and offensive speech in arabic social media using transfer learning. *Applied Sciences*, 12:12823.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th Workshop on Semantic Evaluation*, SemEval '20, pages 1377–1414.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *EMNLP-IJCNLP*, pages 5636–5646.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2023. LLMeBench: A flexible framework for accelerating llms benchmarking. *arXiv:2308.04945*.

W.B. Demilie and A.O. Salau. 2022. Detection of fake news and hate speech for ethiopian languages: a systematic review of the approaches. *J Big Data*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *ACL-IJCNLP*, ACL-IJCNLP '21, pages 6603–6617, Online. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *SemEval*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *CSUR*, 51(4):1–30.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat. 2023a. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Maram Hasanain, Ahmed El-Shangiti, Rabindra Nath Nandi, Preslav Nakov, and Firoj Alam. 2023b. QCRI at SemEval-2023 task 3: News genre, framing and persuasion techniques detection using multilingual models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1237–1244, Toronto, Canada. Association for Computational Linguistics.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, 17(2):1–16.

Clyde R. Miller. 1939. The Techniques of Propaganda. pages 27–29.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and reasoning of deleted tweets before they are posted. *arXiv preprint arXiv:2305.04927*.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of osact4 arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, pages 48–52.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, LREC '20, pages 6086–6093.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937. Association for Computational Linguistics.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5597–5606. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *NAACL-HLT*, pages 1415–1420.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin.

2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *SemEval*, pages 1425–1447.

Lu Zhou, Wenbo Wang, and Keke Chen. 2016. Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 603–612, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.