Efficient Text-based Propaganda Detection via Language Model Cascades

Lin Tian[1], Xiuzhen Zhang[1], Maria Myung-Hee Kim[2] and Jennifer Biggs[2]

[1]*RMIT University, Melbourne, Australia*
[2]*Defence Science and Technology Group, Australia*

Abstract
Identifying propaganda social media posts is an important task. We show how to leverage large language models for DIPROMATS challenge Task 1 – automated detection of English propaganda social media posts. We also demonstrate a more efficient way to utilise large language models, designed to speed up the inference time and maintain competitive performance. Our submission is ranked the first among all 34 runs for the task and achieved a normalised ICM score of 0.8202 and an F_1 score of 0.6784.

Keywords
Text-based Propaganda Detection, Automated Propaganda Detection, Cascades Model, GPT-J

1. Introduction

In this work, we explore different text-based solutions for the first task of DIPROMATS 2023 (Automatic Detection and Characterization of Propaganda Techniques from Diplomats) [1], as part of IberLEF 2023. This event aims to promote the research on developing Natural Language Processing (NLP) tools for propaganda detection on social media text data, esp. written in Spanish and English. In this work, we focused on the English task. Our approach utilises large language models with ensembling and cascades strategies for propaganda identification on social media solely based on the text content.

We focus on the DIPROMATS challenge Task 1 – automated detection of English propaganda social media posts. We design a system of cascades of language models for propaganda detection. We also demonstrate an efficient way to utilise large language models, designed to speed up the inference time and maintain competitive performance. Our submission is ranked the first among all 34 runs for the task and achieved a normalised ICM score of 0.8202 and an F_1 score of 0.6784

2. Related Work

A limited amount of relevant work on propaganda identification has been conducted in NLP research field. A highly related topic is information campaign detection. Early studies have

IberLEF 2023, September 2023, Jaén, Spain
s3795533@student.rmit.edu.au (L. Tian); xiuzhen.zhang@rmit.edu.au (X. Zhang); myung.kim@defence.gov.au (M. M. Kim); jennifer.biggs@defence.gov.au (J. Biggs)

Table 1
Overall statistics of English Training Data

	Total	Propaganda	Non-propaganda
# users	491	198	476
# tweets	8,048	1,974	6,434
avg. # word per tweet	37.03	38.93	36.44
avg. # URLs per tweet	0.99	0.78	1.05

focusd on extracting hand-engineered features from the textual contents of user posts [2, 3, 4, 3, 5, 6]. Signals such as writing style, sentiment as well as emotions have been explored [2, 4]. User online activities have also been applied [3, 5].

Recently, approaches combining user posts and online activities for troll detection have emerged [7, 8, 9, 10, 11, 12, 13]. Addawood et al. [8] identified 49 linguistic markers of deception and measured their usage by troll accounts. They showed that such deceptive language cues can help to accurately identify trolls. Im et al. [9] proposed a detection approach that relies on users' metadata and activity (e.g. number of shared links, retweets, mentions, etc.), and linguistic features to identify active trolls on Twitter.

There has been an increasing interest towards propaganda detection tasks in recent years. GPT-3 [14], released by OpenAI, is a recent language model with a massive number of parameters (175 billion) trained for text generation. The autoregressive language model achieved strong performance on several NLP tasks, even with limited or no fine-tuning by applying zero-shot or few-shot learning. Compared with the usage through the API call, we explored an open-source version of GPT-J (6 billion parameters) for the propaganda detection task of DIPROMATS 2023.

We designed the whole framework aiming not only for performance accuracy but also for robustness and model efficiency. When adopting the large language model as a backbone, we primarily focused on improving the overall performance and speeding up the inference time. Thus, we compared confidence-based cascades models and ensemble models for this task.

3. Dataset

The given English dataset for the propaganda identification task contains 8,048 posts; this training dataset includes propaganda labels. Among them, we randomly sampled 805 instances to use as a development dataset and used the rest as a training dataset.

Table 1 shows the overall statistics of the training data, which consists of 491 total unique users (via given usernames) with 8,048 posts. The average number of words per post is 37.03, and each post has an average of one hyperlink.

The dataset has an imbalanced distribution over the propaganda and non-propaganda labels, as shown in Figure 1. Note that for given countries, the imbalanced label distribution holds across all four countries with "European Union" the most imbalanced one.

The dataset includes four different types of online posts (tweet, quoted, retweet and reply), but most of them are original tweets as shown in Figure 2.

As we focused on building a text-based model for this task, our work has not utilised the

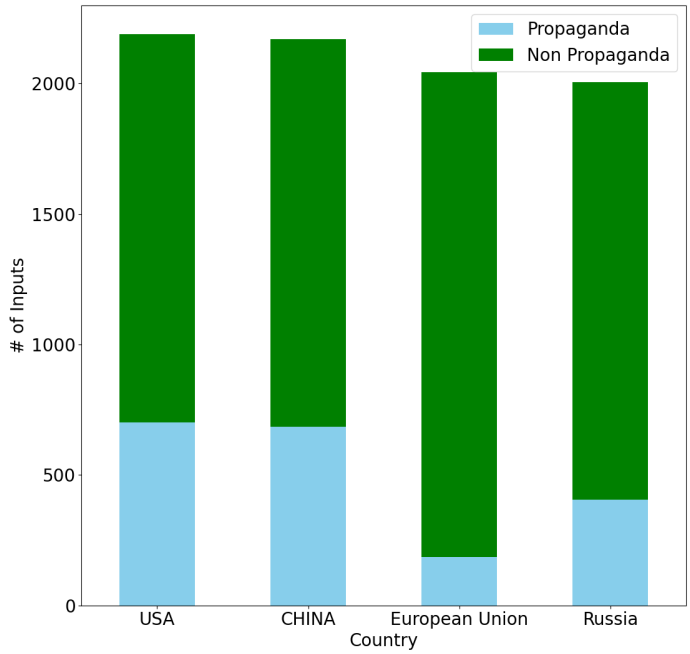

Figure 1: Country Distribution on English Training Data

given meta-features (e.g., username, number of likes, and retweet counts).

4. Methodology

4.1. Model Training

Compared to GPT-3 [14], we adopt the open-source GPT-J as our backbone model for all the experiments. The GPT-J model is a GPT-2-like causal language model trained on the Pile dataset [15]. As the Pile dataset is an English based dataset, our model can only handle English data. The subtask1 of the propaganda identification task can be framed as a binary text classification task.

4.2. Cascade Models

As shown in Figure 3, two GPT-J based models are included in our cascades. One GPT-J is fine-tuned with in-domain propaganda training data, the other troll-boosted GPT-J model is sequentially fine-tuned on the public twitter troll data first and then fine-tuned with in-domain propaganda training data.

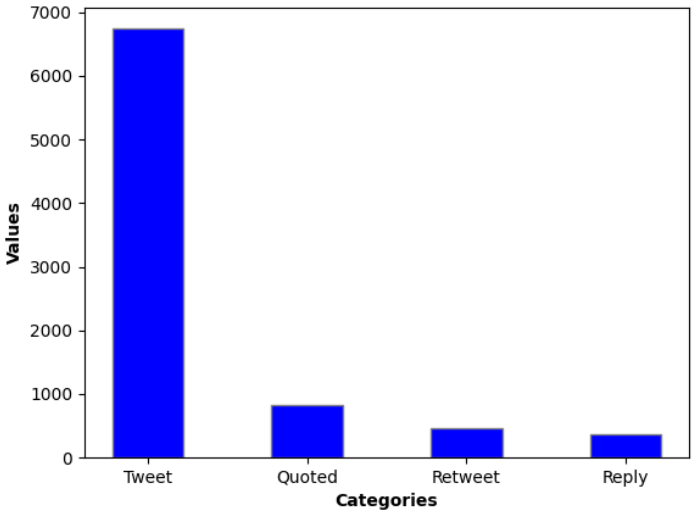

Figure 2: Data type distribution on English Training Data

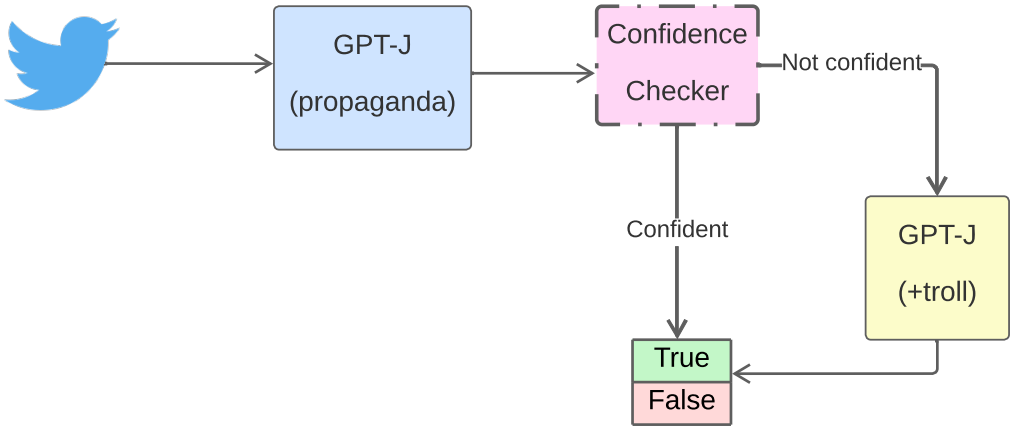

Figure 3: Illustration of our cascades troll-boost classification model

The confidence checker is working as the confidence-score based filter to distinguish the hard samples from easy ones. We use a threshold on the confidence score to determine when to exit from the cascade. The confidence threshold is one of the hyper-parameters in our settings. The final confidence threshold is picked depending on the best performance on our development set.

To highlight the practical benefit of cascades, it saves the computation cost and improves the inference speed compared to ensemble models. Based on our experiments, the cascades models

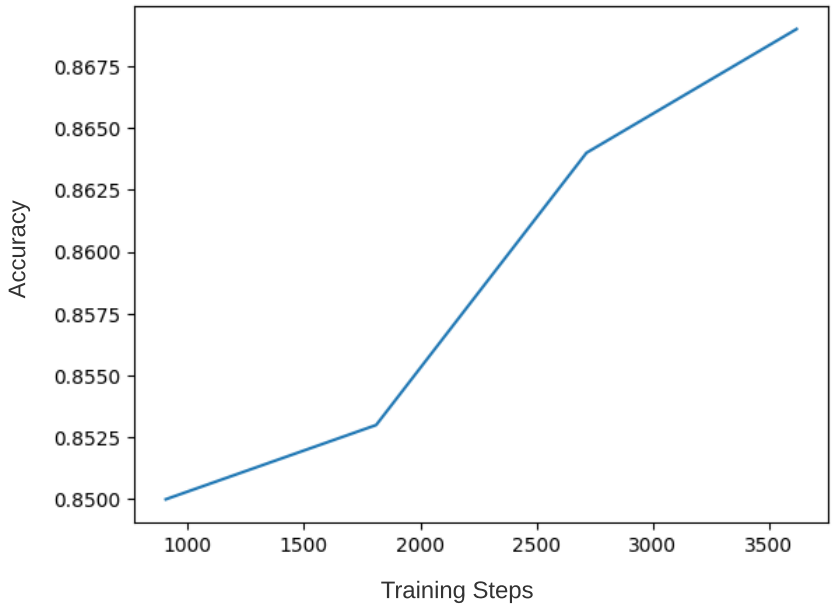

Figure 4: Training classification performance

yield the best performance on the development dataset.

4.3. Label Smoothing

One of the common problems of large language models is their over-confidence on prediction tasks. Label smoothing prevents the network from becoming over-confident and has been used in many state-of-the-art models, including image classification, language translation and speech recognition. Label smoothing is a simple yet effective regularisation tool operating on the labels.

The intuition behind label smoothing is not letting the model learn that a specific input results in a specific output only.

Instead of using one-hot encoded vectors ([0,1] in this case), we introduce noise distribution $u(y|x)$. Our new ground truth label for data (x_i, y_i) would be

$$\begin{aligned} p'\left(y \mid x_i\right) &= (1-\varepsilon)p\left(y \mid x_i\right) + \varepsilon u\left(y \mid x_i\right) \\ &= \begin{cases} 1-\varepsilon+\varepsilon u\left(y \mid x_i\right) & \text{if } y = y_i \\ \varepsilon u\left(y \mid x_i\right) & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

where ε is a weight factor, $\varepsilon \in [0, 1]$ and note that $\sum_{y=1}^{K} p'\left(y \mid x_i\right) = 1$.

By applying this trick, the model becomes less confident with extremely confident labels. This is exactly what we wanted to avoid. As our cascades models are selected purely based on the confidence score, it leads to estimate better on easy and hard sample selections.

Table 2
Hyper-parameters for model training

Name	Value
# epoch	4
batch size	4 per device
learning rate	[1e-5, 1e-4]
temperature	0
training steps	[3,620, 3850]
confidence threshold	[0.85,0.92]

Table 3
Performance over the development set

Model	F1	True	Fasle
GPT-2	0.59	0.60	0.83
GPT-J	0.62	0.63	0.90
Ensemble GPT-J	0.65	0.66	0.94
Cascades GPT-J	0.69	0.70	0.95

5. Experiments

5.1. Settings

For the GPT-J model, we used the huggingface version [16] for our experiments. To fine-tune the second GPT-J model further, we used the public Twitter Dataset [17]. This dataset contains different groups of users banned by Twitter since October 2018 for engaging in state-sponsored information operations.

Furthermore, inspired by [18], we included one more public dataset to improve the models' robustness and mitigate the performance variance. To make the data similar to the task input sequence, we sampled the most recent 20 tweets for each user then we assigned positive labels for troll-users' tweets and negative labels for non-troll users' tweets.

Models that achieved the best performance on our development dataset are used. Hyper-parameters are shown in Table 2.

5.2. Results

To evaluate the performance of the models, the official results are based on normalised ICM [19] and F1 scores.

Figure 4 shows the training process of our cascade model. The best performance, 80.90 F1 score, is achieved when our cascades models are fully used as our final model to do the predictions on the test dataset with positive propaganda class identification on 67.84 and non-propaganda class on 93.97.

In addition, we compared with pre-trained language models, GPT-2 [20] and GPT-J, and also ensembles of GPT-J. The comparison results are included in Table 3. The performance scores

are reported based on our development dataset. As shown in the results, the GPT-J performed better than the base GPT-2 model. The cascades models yield better results over the ensemble models.

6. Conclusion

In this paper, we propose a text-based propaganda classifier with simple cascades models. We show the effectiveness of using the large language models as the backbone and simple confidence-based cascades models for quicker inference. The utilisation of cascades model further shows the benefits of filtering out the hard samples over the label smoothed confidence scores and achieving the best performance in the propaganda detection task 1 in English. It further proves that troll detection and propaganda identification are two closely related tasks.

In future works, we plan to explore the given user meta-features as well as user reactions towards the source posts to improve the performance and apply this technique to other classification tasks such as hate-speech or racism identification on social media.

References

[1] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de-Albornoz, Iván Gonzalo-Verdugo, Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 71 (2023).

[2] J. De-La-Peña-Sordo, I. Santos, I. Pastor-López, P. G. Bringas, Filtering trolling comments through collective classification, in: International Conference on Network and System Security, Springer, 2013, pp. 707–713.

[3] I. O. Dlala, D. Attiaoui, A. Martin, B. B. Yaghlane, Trolls identification within an uncertain framework, in: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, IEEE, 2014, pp. 1011–1015.

[4] C. W. Seah, H. L. Chieu, K. M. A. Chai, L.-N. Teow, L. W. Yeong, Troll detection by domain-adapting sentiment analysis, in: 2015 18th International Conference on Information Fusion (Fusion), IEEE, 2015, pp. 792–799.

[5] J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec, Antisocial behavior in online discussion communities, in: Proceedings of the international aaai conference on web and social media, volume 9, 2015, pp. 61–70.

[6] T. Mihaylov, I. Koychev, G. Georgiev, P. Nakov, Exposing paid opinion manipulation trolls, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, 2015, pp. 443–450.

[7] A. Atanasov, G. D. F. Morales, P. Nakov, Predicting the role of political trolls in social media, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019, pp. 1023–1034.

[8] A. Addawood, A. Badawy, K. Lerman, E. Ferrara, Linguistic cues to deception: Identifying political trolls on social media, in: Proceedings of the international AAAI conference on web and social media, volume 13, 2019, pp. 15–25.

[9] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, E. Gilbert, Still out there: Modeling and identifying russian troll accounts on twitter, in: 12th ACM Conference on Web Science, 2020, pp. 1–10.

[10] H. Shafiei, A. Dadlani, Detection of fickle trolls in large-scale online social networks, Journal of big Data 9 (2022) 1–21.

[11] J. Stewart, M. Dawson, How the modification of personality traits leave one vulnerable to manipulation in social engineering, International Journal of Information Privacy, Security and Integrity 3 (2018) 187–208.

[12] A. Badawy, K. Lerman, E. Ferrara, Who falls for online political manipulation?, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 162–168.

[13] R. Dutt, A. Deb, E. Ferrara, "senator, we sell ads": Analysis of the 2016 russian facebook ads campaign, in: International conference on intelligent information technologies, Springer, 2018, pp. 151–168.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[15] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al., The pile: An 800gb dataset of diverse text for language modeling, arXiv preprint arXiv:2101.00027 (2020).

[16] Eleutherai/gpt-j-6b, https://huggingface.co/EleutherAI/gpt-j-6b, 2023. Accessed: 2023-04-30.

[17] Twitter transparency dataset, https://transparency.twitter.com/en/reports/moderation-research.html, 2023. Accessed: 2023-04-30.

[18] L. Tian, X. Zhang, J. H. Lau, Metatroll: Few-shot detection of state-sponsored trolls with transformer adapters, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1743–1753.

[19] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.