Early Detection of Multilingual Troll Accounts on Twitter

Lin Miao[1,2], Mark Last[2], Marian Litvak[3]
[1]*Beijing Information Science and Technology University, Beijing, China*
[2]*Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[3]*Shamoon College of Engineering, Beer-Sheva, Israel*
miaol@post.bgu.ac.il, mlast@bgu.ac.il, litvak.marina@gmail.com

***Abstract*—Internet troll farms have recently been employed as a powerful and prevailing weapon of information warfare. Even though different tactics may be utilized by different groups of state-sponsored trolls, our goal is to leverage identified troll data for revealing new emerging trolls generating multilingual content. In this work, we adopt a model agnostic meta-learning framework making use of previously released troll farm datasets for the early detection of newly-emerged troll accounts from identified or unidentified troll farms. The detection earliness of various models is evaluated using variable amounts of the earliest tweets from the tested accounts. To evaluate the proposed meta-model, we compare it to several classification models based on different types of account features. Our experiments demonstrate the effectiveness of the meta-model requiring as few as ten tweets to detect a troll account with an average accuracy of 94%.**

***Index Terms*—Twitter, troll account detection, multilingual classification, meta-learning**

I. Introduction

Social media platforms allow easy and fast creation and exchange of user-generated content. Trolls are Internet users who attempt to manipulate opinion or sow discord by spreading disinformation, inflammatory and false information[1]. During the past several years, a large number of organizations utilize troll farms to distribute rumors, conspiracy, and speculation, in an attempt to manipulate public opinion on social media. As more and more countries start to weaponize opinion manipulation, social media has been flooded with troll accounts spreading fake news, propaganda, and misleading information. For example, Russia has been accused of using trolls on Twitter to engage in espionage, manipulation, and propaganda on social media[2]. According to research funded by the UK government, Russian internet trolls are spreading support for the invasion of Ukraine during the conflicts in 2022[3][4]. At the end of January 2019, Twitter started to delete thousands of troll accounts that may attribute to the government of Russia, Iran, and Venezuela. This emphasizes the importance of detecting trolls to protect the public from inappropriate content dissemination on social media. As such, the timely detection of troll accounts should stop opinion manipulation as early as possible. Considering that there are many languages used in social media, the detection of troll accounts needs to handle multilingual content.

Twitter has been releasing the archives of detected state-backed troll farms over time attributing to different countries such as Russia, Iran, Venezuela, etc. Motivated by the above-mentioned issues, in this paper, we aim at detecting multilingual new trolls as early as possible. "New trolls" refers to the new emerging undetected troll accounts, which could be associated with either known (previously identified) troll farms or new unknown troll farms. In order to tackle this problem, we attempt to leverage the available data of known troll farms for early detection of new emerging trolls. In terms of early detection, it arises the question of how many tweets will be sufficient to detect a troll account. Therefore, our goal is to utilize the available data of known trolls to build models that can detect new trolls with only a few tweets. In other words, given a new account represented by as few tweets as possible, the models should be able to determine whether it is a troll or not with a high degree of accuracy. Under all these constraints, we would like to capture the characteristics of the existing trolls and transfer the induced knowledge to other troll farms and languages. In this formulation, we view this problem as a meta-learning problem, which is to train a model on some learning tasks making it easy to adapt to solve new tasks with only small amount of examples [1]. Specifically, in this work, we employ meta-learning to make use of the previously collected trolls for the early detection of new emerging troll accounts. In addition, we investigate the effectiveness of different types of features for early detection of troll accounts.

Overall, the contributions of this paper are summarized as follows:

- Aiming at detecting troll accounts as early as possible, we apply meta-learning to build a meta-model, which can be adapted efficiently to new trolls.
- Considering the multilingual troll accounts, we utilize pre-trained multilingual representation produced by multilingual BERT to achieve troll detection in multilingual content.
- We explore several classification models with different

[1]https://en.wikipedia.org/wiki/Internet_troll

[2]https://mashable.com/2018/01/22/drawing-lines-of-contention-study-twitter-university-of-washington#oHRH4HW4wmqB

[3]https://www.theguardian.com/world/2022/may/01/troll-factory-spreading-russian-pro-war-lies-online-says-uk

[4]https://www.gov.uk/government/news/uk-exposes-sick-russian-troll-factory-plaguing-social-media-with-kremlin-propaganda

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

feature sets and evaluate their ability for early detection of new trolls.
- We release multiple non-troll datasets corresponding to seven different troll farms in multiple languages, which can contribute to the troll detection research.

The rest of the paper is organized as follows. We review related work in Section II. In Section III, we introduce our methodology. In Section IV - V, we explain the construction of the dataset and the experiments, respectively. We discuss the results of the experiments in Section VI. We conclude our paper in the last section.

II. Related Work

A. Troll Detection

As the opinion manipulation concern from troll farms rises, numerous studies are conducted to uncover the troll accounts on social media. In [2], they investigated the troll accounts believed to be controlled by the Internet Research Agency (IRA), and found Brexit-related content from 419 of these accounts. A large dataset was analyzed in [3], which contains 180,340 accounts active during the Russian-Ukrainian crisis of 2014 to discover a series of predictive features for the removal or shutdown of a suspicious account, and they find that lexical features are the most predictive, profile features have high predictive power, network features have moderate predictive power. In [4], the authors evaluate several machine learning methods for detecting abusive accounts with Arabic tweets based on content. Early detection of trolls was explored in [5] using a quantitative measure for identifying troll tweets with as few as 50 tweets.

However, very few studies focused on early detection in multilingual scenario.

B. Meta-learning

Meta-learning, known as "learning to learn", has been widely applied to solve a new task with only a few samples. Optimization-based meta-learning is aimed at learning an initialization for the parameters of a neural network model, such that the model can fast generalize from a small number of examples for the test task [6]. Meta-learning has been shown to be effective and beneficial for various machine learning tasks [1], [7]. Meta-learning has been also applied to plenty of natural language processing tasks, particularly in a multilingual scenario. Authors in [8] proposed a cross-lingual meta-learning architecture, and applied it for Natural Language Inference and Question Answering tasks, demonstrating the consistent effectiveness of meta-learning for 15 languages. In [9], the authors applied meta-learning to solve the user cold-start problem in recommendation system, which refers to recommend items to a new user whose preferences are not observed by the system. In [10], the authors aimed at solving item cold-start problems in Tweet recommendation. They adopted meta-learning taking the history of a user's items to output a model that can be applied to new test items arrive continuously.

In this work, we try to adopt meta-learning for efficient early troll detection from known to emerging troll farms in multilingual settings.

III. Methodology

The main goal of optimization-based meta-learning is to learn a good initialization of the parameters using labeled examples, which can be used as a good starting point to adapt to new examples. Inspired by the idea of meta-learning, we adopt it aiming to obtain a meta-model for further troll detection, which is able to be adapted to new trolls and achieve detection as early as possible. Proposed by [1], Model Agnostic Meta Learning (MAML) is a general optimization framework using gradient descent procedure to obtain a good initial model for a fast adaptation to new target examples. Therefore, we employed MAML for meta-learning in this work. Because of the multilingual setting of most troll farms, we utilize a pre-trained multilingual language model as the base model of our meta-learning. Therefore, the meta-learning procedure starts from a base model M with initial parameters θ. We consider each monitored account to be a training task. First, given a set of training tasks T, we update the model parameters using K data examples of each task by n gradient descent steps. Then the sum loss of all tasks is used to perform meta-update, such that the model parameters θ are updated. We set n as the number of steps for the inner update, α as the learning rate of the inner loop, and the β as the learning rate of the outer loop.

IV. Data Description

A. Troll Farms

Twitter has continuously published archives of accounts that are suspected to be associated with potential state-backed information operations, attempting to influence local and global politics. In this work, we selected several archives released by Twitter[5], the details are shown in Table I. These released archives of accounts are reported to potentially originate in certain countries. However, archives originating in the same country were disclosed and released at different times, or some archives originating in the same country were focused on different topics. For example, the accounts of Venezuela1 and Venezuela2 are all located in Venezuela. However, Venezuela1 is reported to focus on targeting domestic audiences, while Venezuela2 is reported to focus on divisive political themes. So we decided to treat each archive released by Twitter as representing a different troll farm.

To estimate the language distribution of each troll farm, we used Google's Compact Language Detector[6], which is reported to have the accuracy of 99.22%[7].

[5]https://transparency.twitter.com/en/reports/information-operations.html
[6]https://github.com/mikemccand/chromium-compact-language-detector
[7]https://dzone.com/articles/accuracy-and-performance

Dataset	Released Time	# Accounts	# Tweets	Major Language
IRA	Oct 2018	3,613	9,041,308	en(46%), ru(20%)
Iran1	Oct 2018	770	1,122,936	en(55%), ar(36%), es(1%)
Russia1	Jan 2019	416	765,247	en(75%), es(10%), pt(4%), fr(2%)
Iran2	Jan 2019	2,320	4,450,790	ar(35%), en(25%), fa(16%), es(5%)
Venezuela1	Jan 2019	1,196	8,962,498	en(60%), es(32%), pt(2%)
Venezuela2	Jan 2019	764	984,981	en(64%), es(27%), pt(2%)
Thailand	Sep 2020	926	21,385	th(86%), en(4%)

TABLE I
STATISTICS OF THE TROLL DATA PUBLISHED BY TWITTER

B. Normal Accounts

To build balanced datasets for troll detection, we constructed a matching non-troll dataset for each troll farm dataset listed in Table I. The normal accounts for each dataset are randomly selected from the same countries and use the same major languages. For each normal account, we collected 10% sample of tweets data using Twitter API[8].

V. EXPERIMENTS

For verifying the assumption that known troll accounts could help detecting new troll accounts, the accounts in each dataset were sorted by the account creation date. Then the accounts were split into training and testing sets by 7:3 according to their temporal order. In other words, the accounts in training sets were created at an earlier time than the accounts in testing sets.

As our research goal is to utilize all available data of known trolls for early detection of new trolls, while building the models, we used all the tweets of training accounts. To model early detection (to detect a troll account using a minimal amount of its tweets), the tweets of each tested account were sorted by the posted time. The models were applied on N earliest tweets from each tested account, where $N = 0, 1, 3, 5, 10, 20, 50$, in order to find the sufficient number of tweets needed to detect trolls.

Additionally, because we aim to build robust models on older trolls to detect new trolls, we chose to use IRA and Iran1 datasets as "known" trolls. IRA and Iran1 are the two earliest troll datasets published/detected by Twitter containing several mainly used languages, which makes them suitable to be used as "known" trolls for building models. Our meta-model was built on training sets of IRA and Iran1 datasets.

We compared our meta-model with other models applied to various feature categories. All experiments were performed using the same testing accounts corresponding to each dataset.

A. Features

In this work, we applied two types of information which can help characterize a twitter account: **profile**, and **textual content**.

[8]https://developer.twitter.com/en/docs/twitter-api

a) Profile Features: User profile contains the basic information of an account. In this work, we extract 12 profile features[9]. Because some user names were hashed in the troll archives released by Twitter, we did not consider the features which could be extracted from user names (e.g., length of the name, the number of different types of characters). Specifically, these features are mostly static or slowly changing, which means they are not expected to have significant and frequent changes over time. Besides, the profile features are recognized as language independent.

b) Textual Features: Tweet texts reflect the main topic of interest to users as well as the writing style of the users. BERT is well known of extracting high quality language features, as it is able to capture word and sentence level semantics from text data. Besides, each dataset contains tweets in several different languages. Therefore, we chose vectors produced by multilingual BERT [11] for representing tweet content, which was pre-trained on the top 104 languages with the largest Wikipedia corpus.

B. Classification Models

We implemented various classification models using different feature types. Motivated by the demand that the new troll farms lacking of information need to be detected as early as possible, we implemented transfer learning. Therefore, along with the comparisons of different feature types, we also evaluate the transfer learning ability of the classification models within feature type.

1) Profile Models: With the aim of early detection, we trained classifiers on profile features with the assumption that troll accounts can be detected immediately after their creation, even before its user posts tweets. It can be believed that using only profile features to detect troll accounts allows their detection as early as possible. We extracted the profile features of each account and applied Support Vector Machine (SVM) with rbf kernel as classifier following the traditional supervised learning process. We used profile features to build two types of baseline models: **Profile** trained and tested on train/test sets from the same dataset; **Profile_transfer** pre-trained on IRA and Iran2 and applied on a different corresponding dataset.

[9]Profile features: is location available, is description available, is profile url available, creation month, creation day, creation dayofweek, creation dayofyear, creation weekofyear, creation quarter, number of friends, number of followers, followers/friends.

2) Textual Models: Tweet texts reflect the main topic of interest to users as well as the writing style of the users. Textual models were applied on a raw text without feature engineering.

- Fine-tuned BERT: We used BERT multilingual base model (cased), to utilize textual content features for fine-tuning BERT on our datasets. We fine-tuned BERT for each dataset denoted as **Bert_finetuned**. In the training process, we used the Adam optimizer and Cross Entropy loss. Empirically, the initial learning rate was set to 1e-5, and the batch size was set to 32. We also applied transfer learning technique with BERT, pre-trained on IRA and Iran1 and fine-tuned on each corresponding dataset. We denote this model by **Bert_transfer**.
- BERT-based Meta-Model: Based on the hypothesis that a meta-model trained on some previous trolls can be efficiently adapted to new emerging trolls, we chose the earliest two datasets (IRA, Iran1) in Table I as the known trolls to build **MetaModel**.
  We used multilingual BERT [11] as the base model to be trained on these two training sets. Each training set contains two classes. Specifically, we have training sets {IRA_troll, normal} and {Iran1_troll, normal}. For each task, there is a support set used to learn to solve this task, and query set used to evaluate the performance on this task. We used uniform distribution for task sampling, the Adam optimizer with batch size 32, and employed 5 epochs for the training procedure. We set one update step, sample K =100, learning rate $\alpha = 1e-4$ and $\beta = 1e-5$.

C. Evaluation

Except for the profile models, which is independent from the number of representing tweets, the other models were evaluated on the earliest N tweets of each account in testing sets, with $N = 1, 3, 5, 10, 20, 50$, respectively. The models are evaluated on the testing accounts of each dataset. The accuracy scores are reported in Table II. In other words, we conducted zero-shot evaluation of *MetaModel* and *Bert_transfer*, which means evaluate the models directly on the testing sets.

VI. Results and Discussions

The results of all the models are shown in Table II. The 2nd column "# tweets" displays the number of tweets representing each account in the testing set. For the columns: "Profile" and "Bert_finetuned", each row shows the result of the model trained and tested on different parts of the same dataset using corresponding features. For the columns "Bert_transfer", and "Metamodel", where the models were built on the training sets of IRA and Iran1, each row shows the results of the models tested on the corresponding dataset.

We can see that the simplest and earliest approach for troll detection—the models using profile features—yields low performance. It can be explained by a huge challenge to distinguish troll accounts with a small number of shallow features without considering textual content. However, for some troll farm datasets, the profile features show the potential in detection achieving relatively high accuracy, above 80%. But when using model *Profile_transfer* on the other troll farms except the training troll farms, the results are getting worse, which indicates incompatibility of the profile features extracted from certain troll farms to other troll farms.

Contrary to the models using only profile features, the other models were tested using different amounts of tweets per account. In general, the results match the intuitive assumption that more tweets provide higher accuracy.

Compared with profile features, the textual features lead to a significant improvement in accuracy with more tweets. Furthermore, the deep learning models using BERT outperform all the profile models. As such, we suggest that tweet content is more efficient than profile features or behavioral features for troll detection, although training in the former case takes more time.

As shown in Table II, in general, the *MetaModel* outperforms the other baselines. The only exception is that Thailand dataset, where its accuracy is lower than the behavioral model. *MetaModel* was built on IRA and Iran1 to learn a good model which can be efficiently adapted to the new troll farms. Similar to *MetaModel*, *Bert_transfer* model was trained on both IRA and Iran1. Differently, *Bert_transfer* aims to capture the knowledge of the observed examples, while *MetaModel* aims to enhance the ability to adapt. We found that *MetaModel* performs better on new troll farms than *Bert_transfer*, which approves the superiority of meta-learning, where the model is more capable to adapt to new tasks. Also, the results suggest that the *MetaModel* has the best performance for early troll detection. For most cases, the *MetaModel* is able to achieve the highest accuracy by using ten tweets of each account to determine whether it is troll or not. It even only needs five tweets on Iran1 and Venezuela2. In addition, from the overall results, the average accuracy of detecting trolls from different troll farms with ten tweets is 94%.

Compared with *Bert_finetuned*, *Bert_transfer* shows better performance on IRA, Iran1, and Iran2. It supports the intuition that learning on new target tasks benefits from a previous knowledge learned from existing datasets. On the contrary, *Bert_transfer* did not outperform *Bert_finetuned* on Thailand, Venezuela1, and Venezuela2. It can be explained by the fact that these troll farms are relatively different from the training troll farms (at least in a language level) resulting in worse performance. Nevertheless, the *MetaModel* obtains better results than *Bert_finetuned* and *Bert_transfer* on Thailand, Venezuela1, and Venezuela2 datasets. It demonstrates that meta-learning is better at adapting to new troll farms.

VII. Conclusions

In this paper, we addressed the problem of early detection of multilingual trolls on Twitter. We applied the optimization-based meta-learning algorithm MAML for the early detection of multilingual troll accounts from various troll farms. Using rich-labeled data from previously detected trolls, we learned a well-functioning meta-model, which can be efficiently adapted to detect new troll farms. The experiments demonstrated that

Datasets	# tweets	profile	Bert_finetuned	profile_transfer	Bert_transfer	MetaModel
IRA	n=1	0.69	0.56	0.69	0.57	0.72
	n=3		0.75		0.80	0.89
	n=5		0.79		0.91	0.93
	n=10		0.82		0.95	0.96
	n=20		0.85		0.96	0.96
	n=50		0.86		0.96	0.96
Iran1	n=1	0.62	0.78	0.53	0.78	0.86
	n=3		0.80		0.95	0.96
	n=5		0.82		0.98	0.98
	n=10		0.84		0.98	0.98
	n=20		0.84		0.98	0.98
	n=50		0.85		0.98	0.98
Iran2	n=1	0.53	0.66	0.52	0.72	0.82
	n=3		0.65		0.86	0.93
	n=5		0.75		0.89	0.95
	n=10		0.82		0.91	0.96
	n=20		0.86		0.91	0.96
	n=50		0.86		0.91	0.96
Russia1	n=1	0.81	0.75	0.68	0.69	0.77
	n=3		0.76		0.85	0.89
	n=5		0.83		0.88	0.93
	n=10		0.92		0.89	0.96
	n=20		0.92		0.89	0.96
	n=50		0.92		0.89	0.96
Thailand	n=1	0.75	0.64	0.48	0.66	0.77
	n=3		0.70		0.64	0.84
	n=5		0.77		0.64	0.84
	n=10		0.82		0.64	0.87
	n=20		0.85		0.64	0.87
	n=50		0.85		0.64	0.87
Venezuela1	n=1	0.68	0.43	0.59	0.49	0.69
	n=3		0.66		0.48	0.82
	n=5		0.69		0.49	0.86
	n=10		0.71		0.50	0.86
	n=20		0.77		0.50	0.86
	n=50		0.77		0.50	0.86
Venezuela2	n=1	0.85	0.91	0.31	0.77	0.83
	n=3		0.96		0.89	0.95
	n=5		0.97		0.91	0.95
	n=10		0.98		0.91	0.98
	n=20		0.98		0.91	0.98
	n=50		0.98		0.91	0.98

TABLE II
EXPERIMENTAL ACCURACY RESULTS OF ALL MODELS

the troll accounts in different troll farms can be identified by our meta-model using as few as ten tweets with an average accuracy of 94%. Utilizing the pre-trained multilingual language model, our meta-model is shown to effectively detect troll accounts in multilingual settings. In future work, the combination of different types of features may be explored for early troll detection. Future work may also include validation of the MetaModel on other newly released troll data from Twitter.

REFERENCES

[1] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." International conference on machine learning. PMLR, 2017.

[2] Llewellyn, Clare, et al. "For whom the bell trolls: Shifting troll behaviour in the Twitter Brexit debate." JCMS: Journal of Common Market Studies 57.5 (2019): 1148-1164.

[3] Volkova, Svitlana, and Eric Bell. "Account deletion prediction on RuNet: A case study of suspicious Twitter accounts active during the Russian-Ukrainian crisis." Proceedings of the Second Workshop on Computational Approaches to Deception Detection. 2016.

[4] Abozinadah, Ehab A., Alex V. Mbaziira, and J. Jones. "Detection of abusive accounts with Arabic tweets." Int. J. Knowl. Eng.-IACSIT 1.2 (2015): 113-119.

[5] Monakhov, Sergei. "Early detection of internet trolls: Introducing an algorithm based on word pairs/single words multiple repetition ratio." PloS one 15.8 (2020): e0236832.

[6] Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." Advances in neural information processing systems 30 (2017).

[7] Rusu, Andrei A., et al. "Meta-learning with latent embedding optimization." arXiv preprint arXiv:1807.05960 (2018).

[8] Nooralahzadeh, Farhad, et al. "Zero-Shot Cross-Lingual Transfer with Meta Learning." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[9] Vartak, Manasi, et al. "A meta-learning perspective on cold-start recommendations for items." Advances in neural information processing systems 30 (2017).

[10] Bharadhwaj, Homanga. "Meta-learning for user cold-start recommendation." 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.

[11] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of NAACL-HLT. 2019.