Hidetsune at SemEval-2024 Task 4: An Application of Machine Learning to Multilingual Propagandistic Memes Identification Using Machine Translation

Hidetsune Takahashi
Waseda University
takahashi78h@toki.waseda.jp

Abstract

In this system paper for SemEval-2024 Task 4 subtask 2b, I present my approach to identifying propagandistic memes in multiple languages. I firstly establish a baseline for English and then implement the model into other languages (Bulgarian, North Macedonian and Arabic) by using machine translation. Data from other subtasks (subtask 1, subtask 2a) are also used in addition to data for this subtask, and additional data from Kaggle are concatenated to these in order to enhance the model. The results show high reliability of my English baseline and a room for improvement of its implementation.

1 Introduction

SemEval 2024 Task 4 (Dimitrov et al., 2024) calls for classification of memes into different persuasion techniques in textual content only (subtask 1) or in textual and visual content (subtask 2a), and identifying whether or not memes are propagandistic (subtask 2b). I participate in subtask 2b, which is a binary classification problem between "propagandistic" and "non_propagandistic". Various memes are provided in English, Bulgarian, North Macedonian and Arabic to determine whether the memes are propagandistic or not.

My baseline is established so that it achieves fairly high accuracy in English. Although it adopts a classical machine learning method with training, the training data are adjusted by being concatenated with additional dataset. After setting up my baseline, the model is implemented into other languages using machine translation.

Participating in this task allows me to test the ability of my model, achieving a fairly high score for its simplicity. In future studies, the model can be strengthened well enough with appropriately adjusted training data. As for the other languages, on the other hand, the implementation of my English baseline does not necessarily show consistent reliability. Although my baseline works relatively well for Arabic to some extent, the scores go down drastically for Bulgarian and North Macedonian. One of the main reasons for this issue might be that accuracy of machine translation is not high enough, changing the original meanings of the memes and possibly making it more difficult for the English based model to identify propagandistic memes.

My code is available on GitHub [1].

2 Background

The subtask I participate in focuses on classification of memes. Given a json file that has an ID, an image name and text for each meme as input, the subtask requires assigning either "propagandistic" or "non_propagandistic" to the memes. In development phase, data in English were given and participants were allowed to test their solutions for the English data only. Participants were told that they would also have unlabeled test data in three non-English languages as the evaluation phase starts, meaning that no information was released about non-English languages in the development phase.

Propagandistic memes on social media have become a growing issue in the past few years. As more and more people use social media platforms, many types of information including propagandistic one is spread to a number of people (Bradshaw and Howard, 2019). These days, issues caused by propaganda on social media have become worse than most people might think, as a study by O'CONNOR and Weatherall (2019) gives a warning. In fact, these memes did change people's thoughts for elections influencing people's voting behaviors (Aral and Eckles, 2019). Therefore, it might become more and more important in the future for NLP to be able to recognize whether or not the information uses persuasion techniques with high accuracy.

[1] https://github.com/Hidetsune/SemEval2024_Task4.git

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 370–373

There are already many previous studies that adopt modern NLP techniques for propagandistic memes detection. A previous study by Abdullah et al. (2022) uses RoBERTa, which is the state-of-the-art pre-trained language model at the time, resulting in a F1 score of 60.2%. Another previous study by Sprenkamp et al. (2023) tries modern Large Language Models including GPT-3 and GPT-4, also resulting in reliable scores. Al-Omari et al. (2019) utilizes combinations of multiple deep learning models including BERT, BiLSTM and XGBoost with accuracy of around 0.67 in F1 score.

On the other hand of these previous studies, my methodology intends to test classical machine learning approach rather than state-of-the-art LLMs. During my methodology, SpaCy-v3 model (Kömeçoğlu, 2023) is trained with over 20000 memes in English, showing high reliability with accuracy of 0.71353 in F1 macro. Then, the trained model is implemented into non-English languages (Bulgarian, North Macedonian and Arabic) by using Google Translate (Nidhal, 2023).

My participation in this task reveals a high potential of applying classical NLP methods to detection of propagandistic memes with properly adjusted training data. The results reveal that a fairly high score can be achieved without state-of-the-art LLMs and complicated methods. Although the direct implementation of my English baseline into other languages has a room for improvement, the accuracy might go up easily with better machine translation models and careful consideration of differences in topics behind the memes. This paper introduces both strengths and weaknesses of my approach, guiding a direction to future application of classical machine learning to a modern issue of propagandistic memes that requires automatic binary classification.

3 System overview

The main strategy of my system is a classical machine learning method for English baseline and implementation of it into other languages using machine translation. A quick overview of my algorithm is as follows.

1. **Data preparation**: Using official datasets from all the subtasks and additional data on Kaggle[2], training data are prepared to have 11001 memes (including 11000th counting from 0) in English for both propagandistic and non-propagandistic at maximum, making up nearly 22000 memes in total.

2. **Training**: Train a SpaCy model (Kömeçoğlu, 2023) using the prepared training data. Both the training and test data are processed so that they do not have usernames that appear in the additional data and all the memes are lower cased for high efficiency to train the model.

3. **Implementation into non-English languages**: Translate non-English memes into English. This process enables my established baseline to perform in multiple languages, and machine translation is used in this step. After translation of the memes, the model is used in the same way as in English memes.

The system imports a prepared training CSV file as a pandas dataframe, where all the data from Task 4 (train and validation data of subtask 1, subtask 2a and subtask 2b) and additional data from Kaggle[2] are concatenated to compose a large training data with 20774 rows. Only first 20000 rows are used as for the additional data[2] due to its large data size, and all the rows that exceed the limitation of 11001 memes (for both "propagandistic" and "non_propagandistic") are eliminated from the concatenated dataset. Memes in the data are all in English for the purpose of establishing a baseline that guarantees fairly high accuracy in English. Then, SpaCy-v3 model (Kömeçoğlu, 2023) is trained using the resulting data, and test data with unlabeled memes in English is imported as a json file. After that, the trained model is used to assign either "propagandistic" or "non_propagandistic" to each unlabeled meme. As for other languages (Bulgarian, North Macedonian and Arabic), the memes are translated into English so that my English baseline can be implemented into them. Google Translate (Nidhal, 2023) is adopted in this part, and the same trained model is used similarly to the English baseline after translation.

My participation in this task allows for testing the classical NLP approach and simple implementation of it using machine translation. The English baseline, which uses classical machine learning methods, achieves its certain ability to identify propagandistic memes with a reliable score.

On the other hand, the scores go down as for non-English languages. This might be because of

[2] https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter

slight changes in meanings in the translation process, which turns out to be the biggest weakness of my approach. Classification of English memes is fairly difficult even for humans given a single utterance. For instance, "VOTE REPUBLICAN. THEY MAY NOT BE PERFECT. BUT THE OTHER SIDE IS INSANE." can be easily classified as propagandistic, but sentences like "CRY ALL YOU WANT..... HE'S DOING EXACTLY WHAT I HIRED HIM FOR....." might not be that easy for the classification because their nuance might depend on situations (on SNS or at work etc.) to some extent. Since propagandistic memes classification is difficult in this way, possible changes of original meanings by machine translation might have resulted in a serious issue for classification of test data, lowering the scores of other languages dramatically.

Another possible reason for the lowered performance in non-English languages is that there are large differences in topics. Words like "Trump" and "Russia" frequently appear in English memes, but "Bulgarian" is one of the most frequently used words in North Macedonian memes. Since Bulgaria and North Macedonia have some diplomatic issues (KAMBERI, 2023), memes that have basis on them ("THE BULGARIANS ENTER THE CONSTITUTION" for example) tend to appear frequently. This kind of differences in topics probably caused the lowered accuracy in non-English languages, revealing a new challenging problem of the approach to implement my English baseline into other languages.

4 Experimental setup

Before moving on to the actual training of the model, data preparation was an essential part of my solution. First of all, all the data of this task (including other subtasks) are imported as json files and converted into pandas dataframes. The dataframes have "text" and "propagandistic/non_propagandistic" as columns. Training data and validation data from both subtask 1 and subtask 2 are composed of "propagandistic" only, so they are concatenated as "large_data1", which is a dataframe with 8307 rows and all "propagandistic" memes. Also, data from Kaggle[2] are imported as all "non_propagandistic" dataframe and processed so that it has no usernames in "text" column. Validation data of subtask 2b are previously imported, but I decided to increase weight of training data of task 2b by twice and not to use validation data after trial submissions in development phase. Therefore, the resulting training data are composed of training data from subtask 2b (weight increased by twice), "large_data1" and additional data from Kaggle[2]. The training dataframe is adjusted so that it has 11001 memes for both propagandistic and non-propagandistic at maximum, being shuffled to be ready for training.

The reason why the additional data from Kaggle[2] are chosen is that they are less likely to be propagandistic compared to many other existing datasets. The data are from customer support on Twitter including AppleSupport and AmazonHelp, so the topic there should be something related to their products or services. There are many other existing datasets extracted from social media platforms, but it takes too much time end effort to manually assign propagandistic and non-propagandistic to each utterances. For the purpose of getting non-propagandistic utterances only, it might be one of the easiest and most realistic approach to find an existing dataset whose topic is clearly unrelated to politics and diplomacy as included in my solution.

After these data preparation steps, the model (Kömeçoğlu, 2023) is trained with the training dataset, and test data with unlabeled memes are imported as a json file. The memes in non-English languages are translated into English as stated in the previous section. The trained model assigns either "propagandistic" or "non_propagandistic" to each meme. The training data and test data are cleaned with new line removal and lower casing prior to use of them.

LANGUAGE	F1 macro	F1 micro	Ranking
English	0.71353	0.79000	13/20
Bulgarian	0.32670	0.33000	14/14
North Macedonian	0.38942	0.46000	13/14
Arabic	0.52825	0.54375	9/14

Table 1: Task scores for multiple languages

5 Results

Table 1 shows official results of my solutions. They show fairly high accuracy of my English baseline with nearly 0.8 in F1 micro. It can be said that my methodology for English baseline using classical machine learning works fairly well with a thoughtful training data adjustment.

As for the non-English languages, the scores go

down due to the potential reasons as stated in prior sections. Even so, Arabic has relatively high accuracy, which is lower than English by around 0.18 in F1 macro but higher than Bulgarian by around 0.20. This difference might have been caused by accuracy of machine translation mainly. Arabic is a widely used language with a total of about 372.7 million native speakers in the world.[3] There is a possibility that Google Translate (Nidhal, 2023), which I use for machine translation, has higher accuracy for widely used languages including Arabic, maintaining the original meanings and nuances of the memes fairly correctly.

6 Conclusions

To summarise, my methodology firstly focuses on establishing a baseline that guarantee fairly high accuracy for English memes. After that, the baseline is implemented into non-English languages by translating the memes into English using machine translation.

The results show high reliability of my English baseline. The methodology for the baseline has its basis on classical machine learning, but my participation in this task reveals its fundamental abilities to deal with complicated classification task with properly adjusted training data.

On the other hand, the results also show that the simple application of my English baseline has a room for improvement. The scores of non-English languages dramatically dropped although Arabic has relatively reasonable accuracy compared to Bulgarian and North Macedonian. There can be many possible reasons for this including the accuracy of machine translation and changes in topics between memes in different languages.

In future studies, it might be worthwhile to enhance the model with many more memes for my English baseline. Collecting propagandistic memes might be a time consuming task, but non-propagandistic memes, on the other hand, can be easily found and used by utilizing existing datasets whose topics clearly have nothing to do with politics and diplomacy. As for non-English languages, higher accuracy might be achieved by using better machine translation models and enhancing the baseline model with specific political or diplomatic topics in the countries.

[3] https://www.worlddata.info/languages/arabic.php

References

Malak Abdullah, Ola Altiti, and Rasha Obiedat. 2022. Detecting propaganda techniques in english news articles using pre-trained transformers. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 301–308.

Hani Al-Omari, Malak Abdullah, Ola AlTiti, and Samira Shaikh. 2019. JUSTDeep at NLP4IF 2019 task 1: Propaganda detection using ensemble deep learning models. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 113–118, Hong Kong, China. Association for Computational Linguistics.

Sinan Aral and Dean Eckles. 2019. Protecting elections from social media manipulation. *Science*, 365(6456):858–861.

Samantha Bradshaw and Philip N Howard. 2019. The global disinformation order: 2019 global inventory of organised social media manipulation.

Dimitar Dimitrov, Giovanni Da San Martino, Preslav Nakov, Firoj Alam, Maram Hasanain, Abul Hasnat, and Fabrizio Silvestri. 2024. SemEval-2024 Task 4: MULTILINGUAL DETECTION OF PERSUASION TECHNIQUES IN MEMES. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

Donika KAMBERI. 2023. An overview of the dispute between north macedonia and bulgaria through the optic of international law. *JUSTICIA–International Journal of Legal Sciences*, 11(19-20):69–75.

Başak Kömeçoğlu, Buluz. 2023. Emotion Classification with SpaCy v3 Comet.

Baccouri Nidhal. 2023. [The article referred to for machine translation]. *pypi*.

CAILIN O'CONNOR and James Owen Weatherall. 2019. The social media propaganda problem is worse than you think. *Issues in Science and Technology*, 36(1):30–32.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.