How often do you think about the Roman Empire?

A study comparing X thoughts of Dutch people about the Roman Empire.

s1033509
Rabdoud University, Nijmegen, The Netherlands

ABSTRACT

This work provides a pipeline to investigate a reflection of the every-day thoughts of a subset of people about a historical event. More specifically, this study investigates the distribution of male and female X users talking about the Roman Empire. This, after a news story reporting that men think more about the Roman Empire than women. The author uses Dutch tweets from a publicly available scraped X dump. The tweets are classified using a fine-tuned BERT model by whether they talk about the Roman Empire. Then a publicly available fine-tuned Distilbert model is used to classify whether the tweets came from a male or female account. The distribution is reported, and it is found that indeed, Dutch men think more about the Roman Empire than women do. The code is available at https://github.com/freek1/txmm-project.

1 INTRODUCTION

Dutch news outlet Algemeen Dagblad recently caught wind of a remarkable TikTok trend: women asking their male friends/family how often they think about the Roman Empire [1]. The trend showed that the interviewed men think about the Roman Empire unexpectedly much, while highlighting that women do not think about it as often. This study aims to examine this effect by observing the correlation of Dutch people tweeting about the Roman Empire and their gender.

The author will extract the information and topic from a Dutch Twitter (X) dataset [10]. The tweets will be analysed using a pre-trained large language model (LLM), BERT (Bidirectional Encoder Representations from Transformers) [3]. I will fine-tune the model for sequence classification using a small set of keywords which I have decided describe the Roman Empire. The model will then classify all tweets from a different Dutch tweets dataset, which has data from tweets in 2020 [4]. From this set, the genders of the authors of the tweets will be estimated using their twitter names, their public twitter description and the tweet itself. This will be done using a DistilBERT [9] model (a distilled version of BERT) available on Huggingface, fine-tuned for classifying gender based on text [7]. More details on the approach will be explained in section 3.

The motivation for this topic is to analyze how interesting parts of history, like the Roman Empire in this case, are reflected in every-day thoughts of people today. The present-day proxy for our every-day thoughts is an unstructured text medium like X, where lots of data are available.

2 LITERATURE SECTION

To address the research question, it is necessary to distinguish the topic in natural language. Recently, the state-of-the-art has advanced incredibly quickly in the field of natural language processing (NLP), and thus in algorithms – specifically, deep learning approaches, such as pre-trained LLMs – understanding human produced text [8]. Many pre-trained LLMs have been presented, but the main kick-starter to the advances of deep networks in NLP was BERT [3]. The authors show that the pre-trained BERT can be fine-tuned with a single additional output layer to create state-of-the-art models for many tasks. After its inception, many iterations have been made, e.g. more robust pre-training parameters by Liu et al. (2019), with their RoBERTa model (Robustly Optimized BERT Pretraining Approach) [6]. In this study, the autor chooses to use BERT for its simplicity, small size, in combination with the time constraints of this project. BERT is trained on Wikipedia text and BookCorpus data [3]. This enabled the model to understand relations between words and their meaning. For the present study, this means that the model should be able to couple e.g. Rome to The Colosseum, Caesar, and the Roman Empire. In other words, it should be able to generalize text about historical periods well, and therefore BERT is a good model for the task of this study.

There exist methods for estimating a persons gender from the text they use. One example is the method of Cheng et al. (2011) [2], which uses feature selection and several classifiers to achieve around 85% accuracy. Other methods were picked up by some Dutch news outlets, such as TweetGenie (`tweetgenie.nl`), which estimates gender and age from the tweets of a given X account, also boasting 85% accuracy. Both methods do not provide source code, so in the interest of time this study will implement classification by a fine-tuned text classification model from Huggingface [7].

Both parts of this study – classifying tweets using LLMs and estimating gender from twitter names – come with large biases. These biases are acknowledged and will be discussed in the discussion at section 5.

3 APPROACH

The data from the dataset [10] are preprocessed in the following way. The usernames, hashtags and links are replaced by standard tokens and then the duplicate tweets and tweets that were posted more than twice by the same user are removed. This leaves 1,702,651 tweets.

The keywords that are used to determine the positive train samples (i.e., tweets that are about the Roman Empire) are: `"romein"`, `"caesar"`, `"aurelius"`. Masking the 1.7 million tweets with these keywords only returns 209 positive matches. I also sample 209 tweets that are marked negative, thus creating a labelled dataset of 418 tweets with a perfect 50/50 class distribution. These labelled

TxMM'23-'24, Radboud University, Nijmegen, Netherlands

tweets are split into a train, validation and test set, of sizes 267, 67 and 84 items respectively (20% splits).

The tweets are tokenized and the model is fine-tuned with a batch size of 4, for 4 epochs on a Google Colab T4 GPU. Figure 1 shows the training loss, validation loss and validation accuracy over epochs. The model reaches good accuracy and loss values in a few epochs. Since the classes are perfectly balanced, the accuracy metric will suffice.

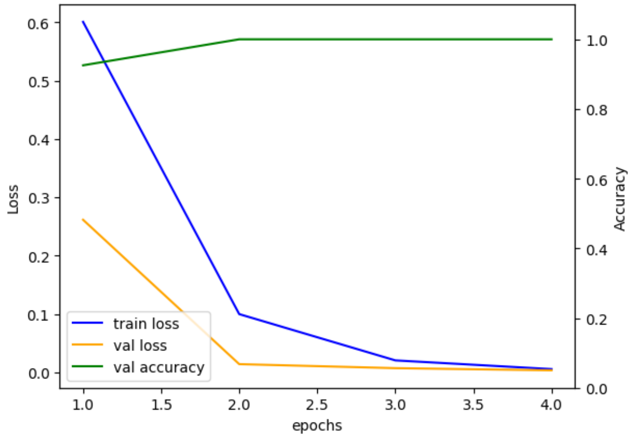

Figure 1: Fine-tuning of BERT model on tweets. The training and validation loss are plotted on the left, with the accuracy on the right (×100%).

The model is then tested on the test set of the tweets and it reports an accuracy of 100%. This is possible in this dataset, because the model is basically taught to identify keywords, which a model with the size of BERT should be able to do, though it does seem to generalize by empirical testing, see table 1.

Then, tweets from 2020 are loaded from the public Huggingface dataset `dutch_social` [4]. I use the "train split" portion of the provided dataset, because this gives enough data to work with, namely 162,805 dutch tweets. The relevant parts of the data contain the tweet, the screen name (username), description (of the user) and the content of the tweet. These tweets are classified by the fine-tuned BERT model from before, and we find that 856 tweets are labelled as positive. This means only 0.53% of tweets from this dataset are about the Roman Empire.

The gender of the posters of the positively marked tweets are then classified by a fine-tuned Distilbert model [7]. This model claims good results on an undisclosed dataset. I chose to use this model, since empirical testing showed promising results and the model was freely available on Huggingface. The classification of gender was done on a combined string in the form: `Name: <screen name>, Description: <user description>, Text: <tweet>`.

4 RESULTS AND ANALYSIS

Figure 2 shows the distribution of predicted probabilities of the tweets from the 2020 dataset being from a female. It shows that there are some tweets of which the model is quite certain of the gender (where the logits are close to 0 and 1). Tweets that the model is uncertain about classifying the gender, i.e., where the predicted probabilities are between 0.35 and 0.65, are not counted towards the final result. This is done, because empirical testing showed that tweets from organisations like e.g. weather channels or news channels are predicted with logits around 0.5. These predictions are plotted in orange in the histogram in figure 2.

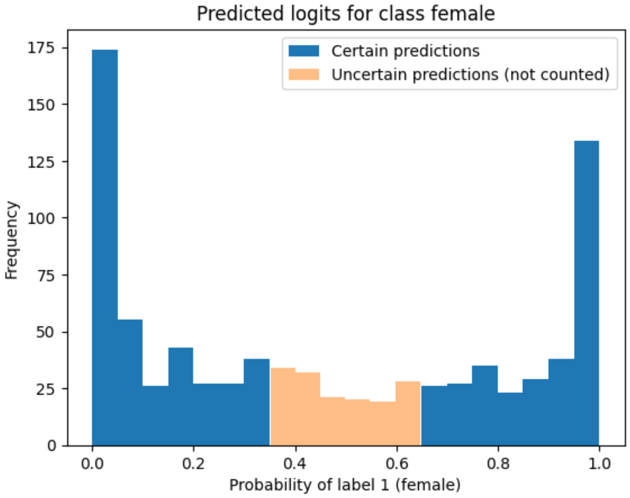

Figure 2: Predicted probabilities of the 2020 tweets.

The remaining X users (in blue) are classified with a threshold at 0.5 (or equivalently, 0.35 and 0.6) and this results in the distribution of male and female X users with tweets about the Roman Empire shown in figure 3. We see that there are 390 male users (55.55%) and 312 female users (44.44%) tweeting about the Roman Empire.

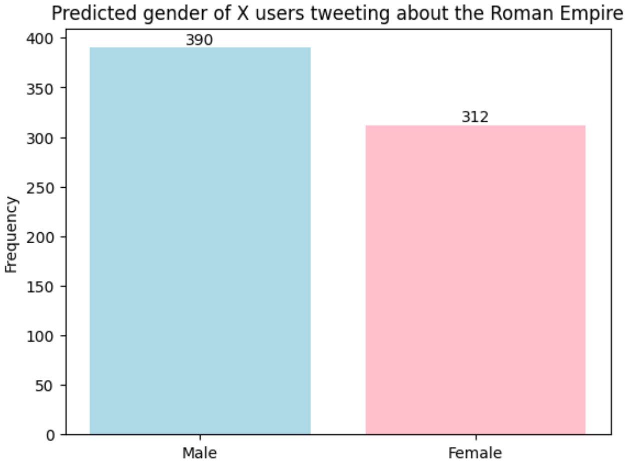

Figure 3: Final classification results of tweets about the Roman Empire.

These results are in line with expectations according to the news article from the Algemeen Dagblad. It noted that men think about the Roman Empire more frequently than women [1]. This research

Manual test text	Predicted label
"Die romeinen waren zo cool eigenlijk vroeger."	1
"Rare jongens, die Italianen met hun aquaducten enzo vroeger."	1
"Mama heeft lekker gekookt vandaag."	0
"Pax Romana van Marcus"	1

Table 1: Empirical test sentences for the trained BERT model on classifying Roman Empire topic.

shows that men tweet about the Roman Empire more than women in the 2020 tweets dataset.

5 DISCUSSION

A first limitation of this study is that not everyone thinking about the Roman Empire will tweet on X about it. Therefore the absolute numbers in this study are approximations, resting on the assumption that everyday thoughts are shared to some truth on X. A second limitation is that it is assumed that both genders share their thoughts equally often, and that about the same amount of males and females use X.

The threshold of the gender selection is set at 0.5, which can be moved to change the distribution of males and females. Since quite a few X users in the dataset are classified in the middle and thus relatively uncertain, the choice can be made to exclude the uncertain users. This would still lead to approximately the same distribution, looking at the amount of low probabilities (around 0) versus high probabilities (around 1) in figure 2. Therefore, the conclusion of this study still holds; there are more males who tweet about the Roman Empire than females in my dataset.

Since the fine-tuned Distilbert model used for gender classification did not report its training data and there are no credentials posted acknowledging the acclaimed performance, the reported performance of the model had to be assumed. To validate the choice of model, I did test some names from the dataset which I had observed to be clear examples of male or female usernames and the model predicted these few examples correctly. I also tested this with some tweets from e.g. a news outlet account, and the gender classifier correctly reported a prediction of around 0.5, indicating it is neither a male or a female. This was the reason for not including these tweets in the counting of males and females.

On a similar note, the gender classification model was fed a particular string combination of the name, description and tweet. Due to the time constraints of this project, no time was spent evaluating the best method of 'prompting' the gender classification model. The author of the model also did not provide this information, so future research should investigate this for optimal performance. The actual fine-tuning the author did namely has a large impact on this. This is seen in the pre-training of open source LLMs, where performance differs greatly when using a differing prompting template than was trained on, e.g.: `"[INST]Tell me about brutalism[\INST]"`, used in Mistral models [5], versus `"### Input:\n Tell me about brutalism\n ### Response:"`, used in OpenChat(-based) models [11]. If the template used in pre-training does not match the template in inference, the performance drops. Because the authors of the gender classification model do not provide the training template they used as well as not providing the training data, there is no way to infer the template they used.

The pre-training data of the LLM used for classifying the Roman Empire is important, since the connection between e.g. aquaduct and Roman Empire needs to be made. If the model is not sufficiently pre-trained on the right knowledge, these topics would not be classified in the same class.

It should also be noted that the author does not pertain the existence of maximally two genders. The data with which the gender classification model is trained simply distinguishes only between male and female.

Finally, this study does not investigate the reason of thinking about the Roman Empire. This may be stereotypical, like 'men think about swords and big fights like in the movie 500' or simply because one lives e.g. in Nijmegen, where there are many structural Roman remains visible. Future research could further analyse the tweets which are currently reported in a binary class 'about the Roman Empire', to a more sophisticated selection of reasons of talking about the Roman Empire.

6 CONCLUSION

This study presented a pipeline to inspect a reflection of every-day thoughts of people about a historical event by analysing tweets. As an example, the tweets of Dutch people from 2020 [4] are analysed with regards to the Roman Empire. This is done by fine-tuning pre-trained LLMs to classify the tweets. The topic classification is done by creating a dataset using keywords and fine-tuning a BERT model. Then the gender of the X users is classified on their name, twitter biography, and tweet by a publicly available fine-tuned DistilBERT model [7]. The more certain subset of gender predictions (with prediction probabilities less than .35 and more than .65) are used and the results are interpreted. It is found that Dutch men think more about the Roman Empire than Dutch women from the 2020 tweets dataset.

It should however be stated that this is a lossy reflection of every-day thoughts of people, since not everyone thinking about the Roman Empire will tweet about it. Also, the threshold for gender classification could be changed. A third limitation of this study is the lack of disclosure of information about the gender classification model [7]. It is noted that the pre-training data of the topic classifying LLM is important and finally some directions for future work are described.

ACKNOWLEDGMENTS

I would like to thank the Teacher and the TA's for their help brainstorming in the hotslots.

REFERENCES

[1] Fleur Broeders. 2023. Nieuwe (voor velen onbegrijpelijke) trend: hoe vaak denk jíj aan het Romeinse Rijk? *Algemeen Dagblad* (September

2023).

[2] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. 2011. Author gender identification from text. *Digital investigation* 8, 1 (2011), 78–88.

[3] 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* (2018).

[4] Aakash Gupta. 2020. Dutch social media collection. https://doi.org/10.5072/FK2/MTPTL7

[5] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[6] 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. *arXiv preprint arXiv:1907.11692* (2019).

[7] padmajabfrl. 2022. Gender-Classification. *Huggingface* (2022). https://huggingface.co/padmajabfrl/Gender-Classification/

[8] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63, 10 (2020), 1872–1897.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[10] Erik Tjong Kim Sang and Antal van den Bosch. 2013. Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal* 3, 12/2013 (2013), 121–134.

[11] Guan Wang, Sijie Cheng, Qiying Yu, and Changling Liu. 2023. *Open-LLMs: Less is More for Open-source Models.* https://doi.org/10.5281/zenodo.8105775

7 WORK REPORT

My research process was searching for state-of-the-art ways to classify text into arbitrary classes, which landed on fine-tuning a transformer model. This process was made easy by the Huggingface library and its abundance of open-source models and datasets. The hotslot meetings helped me guide my research more sharply and helped narrow the topic and motivation down. I struggled a bit with finding a suitable way of estimating genders from text, but finally found the current method (a fine-tuned transformer), which I think is surprisingly fitting for my research.