Fishing for catfishes: using a model trained on Twitter data to predict author gender in Reddit posts

Marcell Kovacs
n. Mat: 0001030874
Alma Mater Studiorum – Università di Bologna
DIT Forlì
Specialized Translation
marcell.kovacs@studio.unibo.it

Abstract

The aim of this project is to create a model, in order to solve a binary classification problem, which is to identify the gender of the author of a post on the social media and content aggregator platform Reddit. This is done with the purpose of catching people impersonating either gender.
Reddit is a pseudonymous website, where users select a username, but provide no biographical data, and as such there is no way to determine the gender of the author of a post, unless they state it, and there is no way to obtain a gold standard on which to train a model.
The model was trained on previously collected Twitter data, obtained from the PAN2018 author profiling task. After this, the model was tested on Reddit data.
Because of the fact that the two datasets are different, the results are not optimal, but could constitute a starting point for the creation of more accurate models, and they still show a few interesting points.

1 Introduction and Background

Men and women are different, and it shows in their writing styles, as well. According to a number of studies, there are statistically significant differences in written texts produced by men and women. For instance, women used more words related to psychological and social processes, whereas men referred more to object properties and impersonal topics (Newman et al., 2008). A number of studies concern identifying the gender of posters on the internet. This process can be useful for social media websites, such as Twitter, where people interact with each other, due to the fact that there is a risk of people lying about their identities. Lying is often done with the intent of scamming people for a financial gain, or to compromise in some way or intentionally upset one or more victims. This phenomenon is called *catfishing* (Simmons and Lee, 2020).

The present study concentrates on Reddit, which is an online discussion forum, that self-labels as ”the front page of the internet” (Anderson, 2015). According to usage statistic, reddit is the 19th most popular website on the Internet and it had over 1.8 billion visits in August 2022 (SimilarWeb, 2022).

The primary purpose of Reddit is sharing content among subscribers of particular *subreddits*, which is a name given to the subforums of Reddit, each related to specific interests. These can be movies, music bands, politics, geography, memes, etc., and submissions can be either pictures, links to other websites, or text posts. A few of the subreddits are dedicated to finding friendships, or relationships, and to meet up with people in real life (Anderson, 2015). This can be dangerous, because users are completely anonymous. In my opinion, Reddit creates a perfect environment for people intending to lie about their identities.

A number of studies have been carried out on gender detection on Reddit, but they relied on informal surveys conducted on the users of the website, such as (Barthel et al., 2016) and a different study, carried out by Craig S. Finlay, used Computer Mediated Discourse Analysis (CMDA) on the collected data (Finlay, 2014). Few studies have been carried out using neural networks on Reddit and none with the purpose of being used on subreddits dedicated to finding relationships. The present study proposes the use of Convolutional Neural Networks (CNNs) for the purpose of detecting gender on Reddit.

Because of the fact that Reddit is a pseudonymous website, there is no way to obtain useful biographical data from the posts, or the user pages (Anderson, 2015), which is one of the reasons why Reddit itself was not used to train a model. One way to obtain useful data would be to conduct a voluntary survey on users who would like to participate in the study, but this was not done, because of the number of responses necessary, and because of the nature of this study. For the purposes of the present study a pre-compiled Twitter dataset was used on which to train models. The Twitter dataset comes from the PAN2018 Author Profiling task (PAN, 2018).

2 Corpora and Dataset description

The training dataset was obtained from the Author Profiling shared task of the PAN organization at CLEF 2018 (Conference and Labs of the Evaluation Forum) (PAN, 2018). The PAN2018 Author Profiling Twitter dataset contains over 6.5GB worth of tagged posts taken from Twitter, alongside a ground truth file containing the gender of the author of each tweet. It is balanced, and has data from 1,500 female Twitter users, and 1,500 male Twitter users. The truth .txt file contains randomly generated hexadecimal IDs for each separate author and states their gender as either ”male” or ”female”. The tweets themselves are found in .xml files, each titled with the ID of the author and containing one hundred tweets. There are a total of 3,000 .xml files (1,500 for males and 1,500 for females).

For the testing part of the study a separate Reddit corpus was built utilising the ”Bulk Downloader for Reddit (BDFR)” tool for Python (Parlakci, 2022). This tool is able to download submissions in bulk from Reddit.

A total of 1,429 submissions in .xml format were collected from the */r/r4r* subreddit[1]. It was chosen because of its relevance to the present study. According to its description, available at the link in the footnote, it is dedicated to people *”looking for platonic or non-platonic friends, gaming buddies, online friends, soulmates, travelmates, smoking buddies, activity partners, friends with benefits, or casual encounters”*. Its relevance is constituted in the fact that it is a forum where people can lie about their identities, and more importantly because in the title of each submission, the author has to tag their gender, and the gender of the person or people they are looking for. This makes it easy to separate the posts into presumed male and presumed female folders, on which to test the model. These tags were the following:

- M — Male
- F — Female
- T — Transgender
- R — Redditor / All

The posts are all titled according to the following scheme: *”<age>[<r4r>] <location> – <subject>”*, where *”<r4r>”* denotes gender and preference. For example: *”25 [M4F] New York - Looking for a friend”* would be a 25 year-old male looking for a female friend. For the purposes of this study only cisgender people were considered – that is male or female.

In order to build the corpus, six tags in the post titles were considered: M4M, M4F, M4R for males, and F4M, F4F, F4R for females.

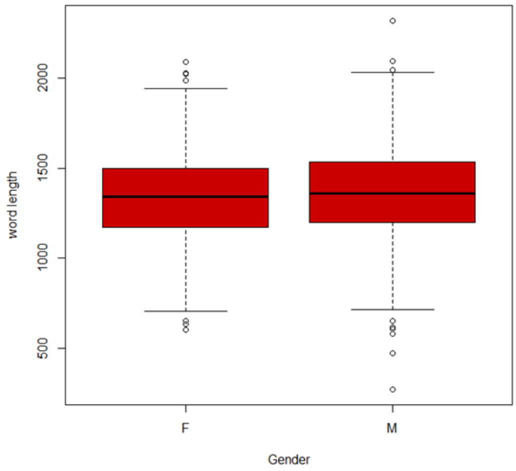
A total of 685 submissions by women were collected, and a total of 744 by men. These 1,429 collected submissions were separated into two folders (presumed male and presumed female) and were used to test the model. Reddit data is unbalanced and is skewed towards males, because there are more men posting than women. 64% of Reddit users are males according to previous demographical surveys (Dixon, 2022).

[1] https://www.reddit.com/r/r4r/

3 Methodology and Code

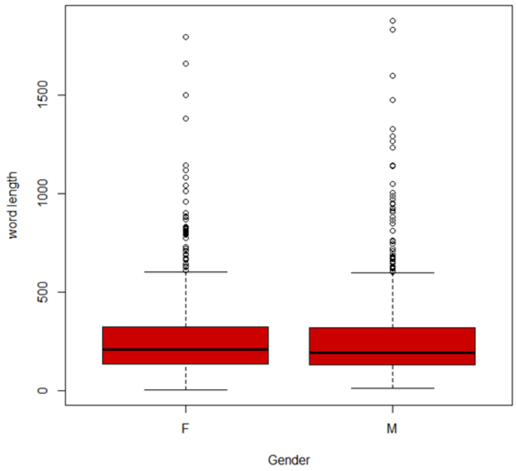
Firstly, code was written from scratch in order to extract the tweets from the PAN2018 dataset's .xml files. This code worked by extracting the tweets present in each .xml file, written by a single author, and compiling them into .txt files containing their tweets, and tagging them as either male or female, rather than every single tweet. This was done, because posts on Reddit are longer on average than Twitter. When the PAN2018 dataset was compiled in 2018, Twitter had a 140 character limit to its tweets, and according to data sampled by Twitter employee Isaac Hepworth, the average tweet's length was of 28 characters, or about 6 words, taking into account the fact that in English the average word length is approximately 5 characters (Panzarino, 2012). The following figure shows the distribution of the lengths of the texts in the Twitter corpus:

figure 1 – Average PAN2018 .txt length



This dataset is normally distributed (the *p*-value according to the Shapiro-Wilk normality test equals p = 0.06 for males and 0.6 for females), and the variances between the two genders are equal (Ansari-Bradley test *p*-value equals 0.56).
- Mean length for males: 1366.7 words
- Mean length for females: 1338.8 words
After this, code was written in order to extract the Reddit submissions from the .xml files, which were downloaded using the BDFR tool. The code looked for the previously outlined tags in the posts' titles, and compiled the corpus into presumed male and female folders, in which each .txt file contained one Reddit submission. The following figure shows the distribution of the length of the Reddit submissions:

figure 2 – Average Reddit post .txt length



According to the Shapiro-Wilk normality test the p-value equals 2.2×10^{-16}. This means that the Reddit dataset is not normally distributed, and that there are a lot of outliers, meaning there are posts much longer than the average.

- Mean length for males: 268.3 words
- Mean length for females: 269.8 words

In order to train the model, the PAN2018 dataset was loaded and pre-processed on the Jupyter notebook, using the method outlined in the book *Natural Language Processing in Action* (Lane et al., 2019). Either a male or female label was appended to each file, and the dataset was shuffled. Two embeddings were tried: one model was trained using Google News' pre-trained word2vec vectors (Mikolov et al., 2013) and a second one using Fasttext embeddings (Joulin et al., 2016). The Fasttext one performed about 1-2% better in accuracy, thus the Google model was discarded.

The dataset was tokenised and vectorised using the Treebank Word Tokenizer[2] and the code outlined in the book. Hyperlinks were removed from the text using the *re.sub()*[3] function and the tokens were lowercased. A number of stopwords were excluded from the data (the following: 'the', 'in', 'of', 'is', 'a', 'to', 'an', 'be'). They were chosen because they contain exclusively grammatical information, which caused the model to be less accurate. A model was trained both with and without these stopwords. The one using these stopwords performed about 1% better in accuracy.

The following parameters were set for the neural network:

```
maxlen = 600
batch_size = 32
embedding_dims = 300
filters = 250
kernel_size = 3
hidden_dims = 250
epochs = 4
```

These values were hand-picked, until a reasonably accurate model was trained. "maxlen" was the only parameter which had a discernible effect on model accuracy, as well as the number of training epochs. The prepared data was split into a training, a validation, and a testing partition divided as such: 70% training, 15% validation, and 15% testing. The dropout layer was set to 0.5.

Please refer to the following link for the code: *https://github.com/mkovacs96/coli-author-profiling*

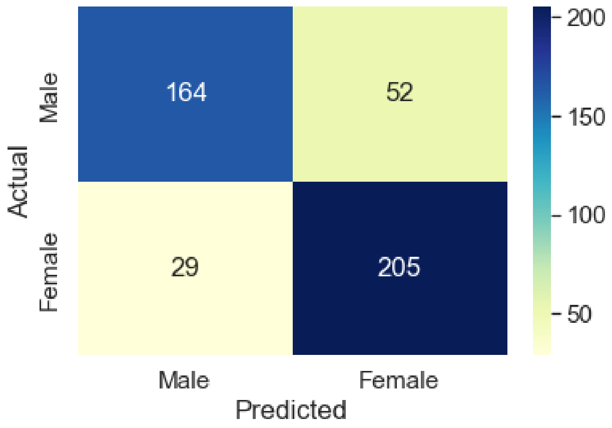
4 Results

It is very difficult to obtain a reasonably accurate model on Twitter, because the differences in writing styles between the two genders are subtle. The performance of models trained using the same parameters during re-training varied between 68% and 81% accuracy on the testing partition. This made it difficult to tweak the parameters and to notice areas of real improvement. A model which after training resulted accurate on the Twitter testing partition, did not necessarily perform better and with more accurate results on Reddit data. The final model described in this paper was chosen for two reasons: firstly, its accuracy of about 81% on the testing partition made it the most accurate out of all trained models, and secondly because it behaved in an interesting way on the Reddit dataset. The model's performance on Twitter data is in line with the results of the PAN2018 author profiling task, with an accuracy of about 80%. See the following figure for the confusion matrix and the table for the evaluation metrics:

[2] https://www.nltk.org/_modules/nltk/tokenize/treebank.html

[3] https://docs.python.org/3/library/re.html

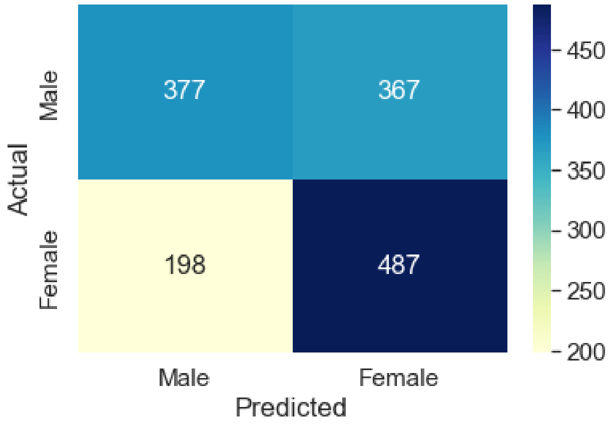
figure 3 – Confusion Matrix of the testing partition of the PAN2018 dataset



Model name	Accuracy	F1	Precision	Recall	Loss
fasttext_stop_cnn_model.json	81.8	0.80	0.76	0.85	0.4513

As expected, during further testing, models behaved poorly on the Reddit dataset, because the source texts are very different from what one would find on Twitter. The character limit of posts on Reddit is of 10,000 characters, which means that submissions are much more coherent and grammatically cohesive, and there are no hashtags (for example #hashtag) or tags (for example @username). Another difference lies in the fact that the Twitter corpus has a general nature, whereas the */r/r4r* corpus is linked to dating and socialising only. This means that the posts are all written in a very similar and specific way. The models' accuracy varied between 50% (pure guessing) and 61% on the Reddit corpus. The following figure shows the confusion matrix of the best-performing model during testing on Reddit data:

figure 4 – Confusion Matrix for Reddit instances



Model name	Accuracy	F1	Precision	Recall	Loss
fasttext_stop_cnn_model.json	60.1	0.57	0.51	0.65	0.7295

Interestingly enough, while the model struggled to identify posts written by presumed males, with an almost perfect split in the middle, it managed to identify posts written by presumed females with an accuracy of 71%. Let us look at the possible implications of this in the next section.

5 Conclusions

Training a model on Twitter data with the purpose of predicting gender on Reddit does not provide reasonably accurate results, and as such if one intends to train a model on Reddit data, or from other anonymous websites, it would be a better idea to rely on surveys to obtain demographical information, rather than training the models on other websites, because their purposes are dissimilar. However, the model managed to identify presumably female submissions posted to the subreddit */r/r4r*, with a significantly higher accuracy. There are a few reasons as to why this might be:

Firstly, it might be pure chance that the model obtained this result. Secondly, the reason might be that women who post to this specific subreddit might write in a way that is more similar to Twitter, whereas men might not. Thirdly, it might mean that people posting as men on */r/r4r* could be more dishonest about their identities. More research could be done in the field of people's writing styles on different websites, as well.

Summing up, as expected the model performed poorly on the Reddit dataset, however its behaviour on presumably female submissions was unexpected.

References

Anderson, K. E. (2015). Ask me anything: What is Reddit? *Library Hi Tech News*, 32(5):8–11.

Barthel, M., Stocking, G., Holcomb, J., and Mitchell, A. (2016). Seven-in-Ten Reddit Users Get News on the Site. *Pew Research Center*, page 44.

Dixon, S. (2022). Distribution of Reddit users worldwide as of January 2022, by gender. `https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/`. Last checked on Sep 17, 2022.

Finlay, S. C. (2014). Age and Gender in Reddit Commenting and Success. *Journal of Information Science Theory and Practice*, 2(3):18–28.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759.*

Lane, H., Howard, C., and Hapke, H. M. (2019). *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python.* Manning Publications Co, Shelter Island, NY.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv:1310.4546.*

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3):211–236.

PAN (2018). Author profiling 2018. `https://pan.webis.de/clef18/pan18-web/author-profiling.html`. Last checked on Sep 17, 2022.

Panzarino, M. (2012). Interesting fact: more tweets posted are 28 characters than any other length [updated]. `https://thenextweb.com/news/interesting-fact-most-tweets-posted-are-approximately-30-characters-long`. Last checked on Sep 17, 2022.

Parlakci, A. (2022). GitHub, Bulk Downloader for Reddit. `https://github.com/aliparlakci/bulk-downloader-for-reddit`. Last checked on Sep 17, 2022.

SimilarWeb (2022). Reddit.com Traffic Analytics and Market Share. `https://www.similarweb.com/website/reddit.com/`. Last checked on Sep 17, 2022.

Simmons, M. and Lee, J. S. (2020). Catfishing: A Look into Online Dating and Impersonation. In Meiselwitz, G., editor, *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis*, volume 12194, pages 349–358. Springer International Publishing, Cham.