Master's Thesis Essay

Bjørn-Andreas Lamo

Spring 2021

1 Introduction

In this thesis we will discuss the viability and applicability of machine learning models to infer a Reddit users gender from their comment history. Gender is only one of the traits of the wider field of author profiling; other traits include age, location, education, etc. We will limit this paper to only consider gender in the profiling exercise because of its availability. Author profiling has been performed with statistical models, machine learning, and recently neural networks.[TODO cite] BERT is the current state of the art model for NLP problems, and has been for the past years.[1] Previous author profiling papers have achieved good results.[2][3][4]

This paper will deviate from most of the other previous work in that instead of making *one model* for all the classification, we will explore making *models per interest or topic.* By making the model topic dependent, the predictions are topic independent, as the predictions are only within the topic. With this approach the model avoids being biased based on keywords related to a topic.

Reddit is an ideal source for text data because the content is already sorted by topic, and it is easy to retrieve all or many of a users comments site wide. The users on Reddit are nominally anonymous, but some voluntary their gender when relevant to the discussion. We will take advantage of this to extract a large number of gendered users, and then fetch their comment history. Due to the nature of comments, one single comment can very short, and have few linguistic features. When profiling a user we will utilize as much as their comment history as possible.

2 Essay structure

Next we will present previous related works that motivates and validates this thesis. These articles will range from author profiling on social media, to more general linguistic gender identification.

In section 4 and 5 we will introduce the science of author profiling and Reddit as a social media platform. This will be just a brief overview of the subject.

And in section 6 and 7 we will discuss features we can extract from the dataset, and which models we can apply them to.

Lastly in section 8 we will lay forth the intended process of entire project. From acquiring the dataset, preprocessing the data, testing different models, and validating the result.

3 Previous works

3.1 Inferring gender of Reddit users

The previous work that most resembles this project is *Inferring gender of Reddit users* by Evgenij Vasilev.[5] We will both be classifying genders based on Reddit comment history. We're also both generating the dataset by exploiting gendered subreddits. Because of these similarities we should be able to fairly compare our and their resulting model. The two departing factors between our and their approach is that our model will be topic independent and we will use state-of-the-art BERT model for the training.

In the paper they achieved an 82% F1 score predicting genders with a Character-CNN model. The dataset was made from voluntarily gendered user flair on gender related subreddits. The actual data was extracted from a database containing all of Reddit from December 2015 to July 2017.[6] Several models was experimented with, such as LogReg, XGBoost, LSTM, Char-CNN. All of them scored around 80% F1. Which is not impressive considering the best model was only marginally better than a model that took significantly less time training. On a comment level the prediction dropped to about 50% F1 on all models, rather abysmal given the binary male and female classification. Indicative of the gender neutrality of a single comment.

3.2 Gender Differences in English Syntax

Britta Mondorf's Gender Differences in English Syntax[7] is a synthesis between theoretical linguistic theory and empirical corpus based studies. The text data is from the London Lund Corpus, which is transcribed spoken British English. The linguistic features that is considered is different adverbial clauses usage and position.

In the findings there is a clear preference for women to use postposed clauses, especially in conditional clauses. Whereas men favor preposed clauses. Speech intonation is also preserved in the corpus, and women scored high on postposed clauses produced under a separate intonation contour. The author goes on to explain that given the intonation such clauses express a lower commitment than presupposed information. And because preposed clauses have less local scope they are used to introduce new topics. Men's preference for preposed clauses might be because of status differences, since higher status individuals generally have the privilege to control the topics. Higher status individuals are also assumed to have good knowledge, and are therefore more sure in their believes, and thus have a greater degree of commitment in their truths; this may explain the significant men's preference to use concessive clauses, compared to women.

This article is interesting because there is a clear preference in syntax between genders. However the relevance for our study is not conclusive. As their corpus is based on in-person conversation from the 70's. Gendered syntax distribution might not be at all similar when writing anonymously on an internet forum. Still, it gives credence to the possibility of uncovering linguistic gender differences on Reddit.

3.3 Privacy on Reddit? Towards Large-scale User Classification

[TODO]

4 Author Profiling

Author profiling has been practiced for a long time. The chief concern finding the real author of a piece of text, often of literary works,[8][9] but also to catch [TODO forensic criminals].[10] [TODO utdyp]Other than singling out an individual author profiling also narrow down the authors traits. Traits like age & gender,[11] location, education, personality,[12] occupation[13], psychometric traits.[14] Every conceivable way to divide groups of people can be used in author profiling, though degree of success varies.[TODO find unsuccessful]

With the rise of social media many ordinary people has become the author of posts and comments on these sites. Albeit a much shorter piece of literature than a novel, but still in considerable quantity. A move from individual analysis to automated computer models was necessary with the increase in authors and numerous shorter texts. Gender classification based on posted text on the internet has been in academic interest for decades.[15]

5 Reddit

Reddit is a popular site for sharing content and discussions. It was founded in 2005 as a basic link aggregator much akin to Digg at the time. Subsequently the ability to comment on posts was added, and as the site grew in popularity it was divided into topic specific communities called "subreddits". Users on Reddit are very good at policing what content belong on which subreddits, when a post is made other users can upvote/downvote and comment on the post. If the post receives more downvotes than upvotes early on it will be buried and other users won't see it. However if it's a high quality post in the appropriate subreddit, then it will rise to the top of the subreddit such that it is visible to more people. As time passes, the post will sink back down again from score decay. This ensures fresh content are always at the forefront. Reddit users are very good at policing which post belong in the subreddit by means of the voting system. The site shifted from a link aggregator to a forum with dedicated users discussing and sharing content. Today, Reddit is one of the most popular websites in the world,[TODO cite] and in 2020 alone there were 2 billion user comments on the site.[16]

The comment section on Reddit has a tree style. Making it easy to follow discussions in long reply chains. On most subreddits they are by default sorted most popular, ensuring most people see the best comments. As most people only see the top comments, early commentators have a disproportional advantage to garner more upvotes only by virtue of being early. Other subreddits sort by newest, resulting in many more top level comments, but not so deep reply chains.

Most Reddit users are from USA, and English is the predominant language throughout the site. There is a gender disparity on average over the site. A 2012 article suggested the user base was 74% male based on advertisement data.[TODO cite wiki] In 2021 a traditional [TODO statistics/consensus] was conducted that found of U.S. adults 23% men and 12% women used Reddit.[17] While this is the site average since subreddits are topic based, there are subreddits with higher ratio of women.

6 Textual Features

When profiling researchers can use every possible property with a text, even meta data, information not contained in the text itself. Such as when it was written, what media it was written in (handwritten, typewriter, digitally, etc), the context, what type (fiction, article, review, etc), and so on. We will only regard features that is in the text. Of these features there are:

Character-based features that is the usage of symbols such as periods, commas, parenthesis, hyphens, etc. Also the ratio between upper- and lowercase characters, and character length in the text.

Word-based features, vocabulary used, mean word length, misspelled words frequency, representing words as vectors.[18]

Sentence-based features has been used in classical statistical author profiling,[19] as the sentence length distribution is generally consistent between an authors work.[20]

Lastly there is **syntactical features**, of which we categories three of them: *part of speech*, *dependency features*, *tree features* [TODO https://www.aclweb.org/anthology/ 2108.pdf]

7 Models

8 Methodology

en tekst mange tekster

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Roberto López-Santillán, Manuel Montes-Y-Gómez, Luis Carlos González-Gurrola, Graciela Ramírez-Alonso, and Olanda Prieto-Ordaz. Richer document embeddings for author profiling tasks based on a

heuristic search. *Information Processing & Management*, 57(4):102227, 2020.

[3] Marco Polignano, Marco de Gemmis, and Giovanni Semeraro. Contextualized bert sentence embeddings for author profiling: The cost of performances. In *International Conference on Computational Science and Its Applications*, pages 135–149. Springer, 2020.

[4] Rosa María Ortega Mendoza, Anilú Franco Árcega, Manuel Montes y Gómez, et al. Author profiling on social media using new weighting schemes that emphasize personal information. *Computación y Sistemas*, 23(2):501–510, 2019.

[5] Evgenii Vasilev. Inferring gender of reddit users. masterthesis, Universität Koblenz-Landau, Universitätsbibliothek, 2018.

[6] Google bigquery reddit database. `https://console.cloud.google.com/bigquery?p=fh-bigquery&page=project`.

[7] Britta Mondorf. Gender differences in english syntax. *Journal of English Linguistics*, 30(2):158–180, 2002.

[8] James Shapiro. *Contested will: who wrote Shakespeare?* Simon and Schuster, 2011.

[9] Donald Ostrowski. 9. did mikhail sholokhov write the quiet don? In *Who Wrote That?*, pages 209–230. Cornell University Press, 2020.

[10] Dave Davies. Fbi profiler says linguistic work was pivotal in capture of unabomber. `https://www.npr.org/2017/08/22/545122205/fbi-profiler-says-linguistic-work-was-pivotal-in-capture-of-unabomber`.

[11] Francisco Rangel and Paolo Rosso. Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, 177, 2013.

[12] Rosa María Ortega Mendoza, Anilú Franco Árcega, Manuel Montes y Gómez, et al. Author profiling on social media using new weighting schemes that emphasize personal information. *Computación y Sistemas*, 23(2):501–510, 2019.

[13] Jiaqi Pan, Rishabh Bhardwaj, Wei Lu, Hai Leong Chieu, Xinghao Pan, and Ni Yi Puay. Twitter homophily: Network based prediction of user's occupation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2633–2638, 2019.

[14] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, 2007.

[15] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.

[16] Reddit's 2020 year in review. `https://redditblog.com/2020/12/08/reddits-2020-year-in-review/`.

[17] H. Tankovska. Reddit usage reach in the united states 2021, by gender. `https://www.statista.com/statistics/261765/share-of-us-internet-users-who-use-reddit-by-gender`.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[19] Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[20] G Udny Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390, 1939.