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(AP); 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 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.

2 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,[5][6] but also to catch criminals.[7] Other than singling out an individual author profiling also narrow down the authors traits. Traits like age & gender,[8] location, education, personality,[9] occupation[10], psychometric traits.[11] 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. Gender classification based on posted text on the internet has been in academic interest for decades[12].

3 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 post was added, and as the site grew in popularity the site 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.[13]

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.[14] While this is the site average since subreddits are topic based, there are subreddits with higher ratio of women.

4 Previous works

Predicting gender based on Reddit user comments has already been researched. Evgenij Vasilev[15] achieved an 82% F1 score predicting genders with a Character-CNN model. The dataset was made by exploiting the 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. Several models was experimented with, such as LogReg, XGBoost, LSTM, Char-CNN. All of them scored around 80% F1 score. Which is not impressive considering the best model was only marginally better than a model that took significantly less time training.

5 Textual features

6 Models

7 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*, volume abs/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*, volume 57. July 2020.

[3] Marco Polignano, Marco de Gemmis, and Giovanni Semeraro. *Contextualized BERT Sentence Embeddings for Author Profiling: The Cost of Performances*. Springer International Publishing, Cham, 2020.

[4] Rosa María Ortega Mendoza, Anilú Franco Árcega, and Manuel Montes y Gómez. *Author Profiling on Social Media using New Weighting Schemes that Emphasize Personal Information*, volume 23. June 2019.

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

[6] Donald Ostrowski. *9. Did Mikhail Sholokhov Write The Quiet Don?*, pages 209–230. Cornell University Press, 2020.

[7] Dave Davies. Fbi profiler says linguistic work was pivotal in capture of unabomber. *NPR*.

[8] Francisco Rangel and P. Rosso. Use of language and author profiling : Identification of gender and age. 2013.

[9] Rosa María Ortega Mendoza, Anilú Franco Árcega, and Manuel Montes y Gómez. Author Profiling on Social Media using New Weighting Schemes that Emphasize Personal Information. *Computación y Sistemas*, 23(2):501–510, June 2019.

[10] 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 *ACL*, 2019.

[11] 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.

[12] M. Koppel. *Automatically Categorizing Written Texts by Author Gender*, volume 17. November 2002.

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

[14] H. Tankovska. Reddit usage reach in the united states 2021, by gender. *statista*.

[15] E. Vasilev, C. Wagner, and F. Lemmerich. *Inferring Gender of Reddit Users*. Universität Koblenz-Landau, 2018.