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 [TODO forensic criminals].[7] [TODO utdyp]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. 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.[12]

3 linguistic features

When profiling researchers can use every possible feature with a text, even meta features, 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 the linguistic features, that is only what 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 features, vocabulary used, mean word length, misspelled words frequency, representing words as vectors.[13]

Sentence based features has been used in classical statistical author profiling,[14] as the sentence length distribution is generally consistent between an authors work.[15]

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/E17-2108.pdf]

4 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.

5 Previous works

[TODO bad results with only one comment] Predicting gender based on Reddit user comments has already been researched. Evgenij Vasilev[18] 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.

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

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

[7] 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`

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

[9] 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.

[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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2633–2638, 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] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.

[13] 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.

[14] 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.

[15] 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.

[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-`

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