Master's Thesis Essay

Bjørn-Andreas Lamo

Spring 2021

Abstract

With the rapid growth of social media people consume an increasing amount of content and news on these platforms. Compounding with the low threshold of reaching a large audience, social media has become a breeding ground for fake news. Evaluation of online post and the users behind them has become a real preventative effort. PAN at CLEF has for years been hosting events and competitions for author profiling. These task include verifying authorship, identifying bots and fake news. But most of the highest scoring submissions use simple statistical models such as logistic regression, random forest, and support vector machines.

We want to find deep linguistic features between different author groups. To do this we will make the classification topic independent and employ state of the art machine learning. If we limit our author profiling groups to only gender, then we can train the model on the rich dataset of Reddit comments as they are already divided into specific topics.

1 Introduction

In this thesis we will discuss the viability and applicability of machine learning models to predict 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. BERT is the current state of the art model for NLP problems, and has been for the past years.[1] It is a transformer-based machine learning model. Previous author profiling papers have achieved good results using BERT.[2][3]

We will use a BERT model paired with another neural network to capture linguistic features. The two models can either be co-joined or stacked on top of each other. In addition we will train *one model per topic*, instead of just training one model. By making the model topic dependent, the gender predictions are then topic independent. This reduces the significance of keywords from strongly gendered topics and forces the model to look for deeper linguistic features.

Reddit is an ideal data source for this thesis 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 disclose 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 be very short, and have few linguistic features. When profiling a user, we will utilize as much as their comment history as possible.

Note here that *gender* is not the same as sex. Gender roles are developed from the social and cultural interaction with an environment. In our dataset from Reddit we only have self-proclaimed binary gender, and will therefor make predictions in this space. This thesis is based on the assumption that people who identify as either male or female, have shared experiences within their group and that these experiences affect their language.

The resulting model will apply gender probability for a user, topic wise for all their comments, and from all the models deliver the average final gender probability.

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 the entire project. From acquiring the dataset, preprocessing the data, testing different models, and validating the result.

3 Previous works

Below are three previous works that are highly relevant for this thesis. The first one is a master thesis that is similar to this very thesis, and is the adversary to supersede. Secondly a paper on gender differences in syntax. If there are underlying linguistic differences between men and women, then we can use those to boost our model. And the thirdly paper demonstrate the feasibility of feature engineering even when using BERT. Promising that we can include linguistic features model with the BERT model.

3.1 Inferring gender of Reddit users

The previous work that most resembles this project is *Inferring gender of Reddit users* by Evgenij Vasilev.[4] 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.[5] They experimented with several models, such as LogReg, XGBoost, LSTM, Char-CNN. All of them scored around 80% F1, which is not impressive considering the best model were only marginally better than a model that took significantly less time training. LogReg scored 81.35 and trained for a little over 2 hours, while Char-CNN scored 82.33 and trained for over 100 hours. On the comment level prediction dropped to about 50% F1 on all models, rather abysmal given the binary male and female classification. This is indicative of the gender neutrality of a single comment.

3.2 Gender Differences in English Syntax

Britta Mondorf's *Gender Differences in English Syntax*[6] 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 are considered are 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 of the conclusion from this paper to our study is not certain. 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 none the less.

3.3 The Death of Feature Engineering?—BERT with Linguistic Features on SQuAD 2.0

With the advent of BERT[1], solving NLP problems with high performance by just providing word embeddings was possible. Given a pre-trained BERT model, word embeddings, and a dataset. It is possible to achieve state of the art performance within minutes of fine-tuning. That is the power of contextual embeddings.

But what if you want to train it to recognize sentence syntax such as noun phrases or verb phrases. BERT cannot compute a sentence with all words that are out of the embeddings vocab, so adding POS tags directly to the words would not work. It is possible to infuse the embeddings with POS tags embeddings, but that has not been found to improve the model significantly.[7]

Zhang suggested in *The Death of Feature Engineering?*[8] to concatenate the BERT model with a linguistic features model. In the paper he training on Stanford Question Answering Dataset; a benchmark corpus where given an informative paragraph the language model should be able to fill the missing blanks from a sentence describing some fact or information from the above paragraph. BERT has scored well on this dataset, but the author notes that with high dependency sentences BERT is not able to find the

relevant meaning.

To improve the model, the author extracts four linguistic features (named entity, POS tag, syntactic dependency, and stop words) from the input and sends them through a linear layer. The output from the linguistic linear layer is then concatenated the BERT output and are sent through a final linear layer to produce the output.

The author found that this significantly improved BERT_{base}, but slightly decreased BERT_{large} score. The author theorises that BERT_{large} is powerful enough, such that the additional model did not contribute anything.

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,[9][10] but also in criminal forensic analysis.[11] Other than finding one specific author, author profiling can also be used to narrow down on the authors traits. Traits like age & gender,[12] location, education, personality,[13] occupation[14], psychometric traits.[15] Every conceivable way to divide groups of people can be used in author profiling, though degree of success varies.

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.[16]

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, and in 2020 alone there were 2 billion user comments on the site.[17]

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. In 2021 a traditional phone survey was conducted and found that of U.S. adults 23% men and 12% women used Reddit.[18] 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. More recently, representing words as vectors has lead to great success.[19] With word vectors each word is a vector in a multidimensional space, words that are similar have a similar vector, and it is possible to add or subtract word vectors to find relations between words. A common is example: $King - Man + Woman = Queen$

- **Sentence-based features**
  Historically, sentence length has been used in statistical author profiling.[20] Sentence length distribution has been found to be generally consistent over an authors literary work.[21] However sentence length has not been found to be a good feature in determining gender, as the average sentence length varies over both age and gender.[22]

- **Syntactical features**

  - **Part-of-Speech feature** is the relative frequency of a part-of-speech. Part-of-speech is the word class for a given word, such as *noun*, *verb*, *adverb*, *adjective*, etc. The number of part-of-speech depends on the specific notation used. For instance it is possible to subdivide noun into proper, or plural; or verbs to base, present, or tense; etc. The Penn Treebank project has 34 distinguished part-of-speech tags.
  - **Dependency tree feature** is how the sentences part-of-speech clauses are arranged in a certain sequence. There are many possible permutations to express the same sentiment in a sentence. In Mondorf's paper above (section 3.2) we saw a clear gendered preference in the position of the adverbial clause in a sentence. That means the dependency trees are different between male and female sentences, but they hold the same semantic meaning.
  - **Tree features** are the features of the dependency tree. Which are the width and height of the tree, and the density of the branches.

7 Models

7.1 Neural networks

Neural networks are loosely based on biological brain neuron activations. Briefly it explained a neural network receives some input, passes it to the first layer of neurons and applies weights per neuron. An activation function calculates if each neuron is activated or not. The activated neurons are carried over to the next neuron layer, weights applied, activation decided, etc. After the last layer we compare the final resulting output with the expected output, the discrepancy between the two is used to modify the weights so it will hopefully do a better job next forward pass in the model. This is an example of a fully connected, or dense, neural network. It requires

many neurons per layer and many layers of neurons to be able to learn features from the input. But because every neuron is connected to every neuron in the next layer, the number of weights that needs to be calculated during training increases exponentially with the size of the model, and it quickly becomes unfeasible to train. Instead of having a model with all fully connected layers, modern neural networks mostly use the fully connected layer as an output layer, to scale down the hidden neuron size down to the output size.

The two most prevalent neural network architectures are convolutional neural networks(**CNN**) and recurrent neural networks(**RNN**). CNNs are great at finding finding features in images. While RNNs are more suitable for sequential input such as natural language.

Below is a brief overview over these two neural networks:

- **Convolutional neural networks**
  Given a two dimensional image CNNs works by applying a convolution, or filters, bit wise to the image. The convolution layers are stacked in the model. Each convolution can contain many filters, and each filter finds some feature. For instance the filter on lowest level can exclusively find vertical edges. On higher levels the convolutions can piece together many previous detected features, and detect complex shapes, such as a human face or a car. CNNs have found great success in computer vision tasks by increasing the model depth, i.e. increasing the number of convolutional layers.[23]

- **Recurrent neural networks**
  An RNN does element wise computation of a sequence, but it also remember the previous state and also accounts for that. Conceptually this is great for natural language, where the words in a sentence depend on the previous words in order to make sense. It is also possible to send in the input sequence reversed to capture even more inter-dependencies within the sequence, this is called a bidirectional RNN. There are two popular variations of RNN, sometimes referred to as their own distinct models, these are gated recurrent unit(**GRU**)[24] and long short-term memory(**LSTM**)[25]. These variations are more complex gated cells attached to the RNN. In a vanilla RNN the cell is just the computation of the current input and the previous state. With GRU and LSTM, the cells can choose how much of the previous state to ignore, and how much will be carried over to the next hidden state. LSTM also has an internal cell state that is passed on independently of the forgotten hidden state, which theoretically gives it an edge with long sequence

dependency. But in practice, LSTM and GRU perform about the same, with the bonus that GRU is faster to train.[26]

7.2 BERT

Bidirectional Encoder Representations from Transformers(**BERT**)[1] is the state of the art model for NLP tasks. Even though BERT performs very good, it is not known *why* it performs so good. The model is comprised of a stack of encoders, the two original models BERT_{base} has 12 encoders, while BERT_{large} has 24 encoders. These two models are pre-trained by Google on data extracted from BooksCorpus and English Wikipedia. The language model is based on the assumption that if the model can correctly predict what the masked words should be in a sentence, then the model understands language. This is how BERT is trained, by eventually predicting the correct words masked in a sentence.

8 Methodology

This is the preliminary outline of the research process we will endeavour.

8.1 Data collection

8.1.1 Reddit API

We can crawl Reddit by making API requests. As a denial of service precaution, that is that too many people make requests and overload the servers; there is a limit of 60 API requests per minutes. Also the user comment history is limited to the 1000 latest comments submitted.

8.1.2 Pushshift API

Alternatively instead of querying Reddit directly, there is a big-data storage and analytic project called Pushshift that copies submissions done to Reddit to its own servers. This service was created to facilitate big-data research on Reddit that the official API did not support. The limit on API requests is 200 per minute, and it allows us to extract a users entire comment history, with some caveat. Because Pushshift is a third-party clone of Reddit, it is possible that some data was not included between the cloning updates and would be missing. Also submission edits and precise scores may be inaccurate due to query limitations, but we can disregard these. On the whole Pushshift should be more complete because it provides a larger

comment history. However the official 1000 limit might be good, as it encourages more diversity in users, and not over reliance on a few users with huge histories that may be statistical outliers.

8.1.3 Users and comments

Gendered users are collected from the top level commentators on r/AskWomen and r/AskMen. Top level comments are the ones that respond directly to a post. Other comments in the thread are replying to other commentators, and do not need to follow the gendered response rule. We can filter out a minimum required upvotes on both posts and comments to ensure quality. On Ask[Something] subreddits some users will reply on the top level even though they are not the target demographic. When they commit this discrepancy it is customary to start the post with "I'm not a ... but-". This is not a common occurrence on AskWomen and AskMen, but we can naively filter those and evaluate if this is a problem.

When we have a list as self identifying men and women, we can do an API query for the user comment history. The comments will be sorted into subreddit and appropriate gender. Each subreddit a dataset of male and female comments. The datasets are going to be imbalanced. As noted above (section 5), the subreddits are topic-based, women and men have different interest, hence the distribution inequality.

8.2 Models

As previously stated, we will BERT as the main classification model, and we will experiment with attaching linguistic features neural networks. BERT naturally learns some syntactic knowledge, but it is not similar to annotated resources.[27] We are going to analyse it and hopefully extend it with the linguistic features model.

8.3 Evaluation

The performance measurement to evaluate the models will be F1-score. It is an industry standard to measure model performance, as the harmonic mean between accuracy and precision is a good presentation of scope and the quality of the predictions the model made. We will use Vasilevs[4] 82.33 F1-score with the Char-CNN model as our baseline. Unfortunately he does not list the test dataset. He does host a list of the entire Reddit user dataset used for training and testing. We'll try to inquire about the test dataset, if

that is unsuccessful, then a random proportional sampling from the entire dataset will have to do.

In addition to testing on our own test split of our harvested dataset. We can also test on the Reddit part of RtGender corpus.[28] They have collected 19,010 male users and 11,116 female users from various subreddits. It is possible there are some overlap of users for our dataset and this corpus, but we can filter those out to ensure the model has not seen the comment history before. It would be interesting to test on this dataset because we only harvest users from two subreddits. If we achieve the same score on the RtGender Reddit users, then we have proved the model is sturdy on Reddit users.

The finished model will have one model per topic. This is fine for Reddit as the comments are clearly divide into topics. But it would certainly limit the wider applicability outside of the site. *Is* a possibility, *but* with vector representation of a paragraph we can find the cosine similarity with all of the topics. In other words we can find which topic a social media post would fall under.

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

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

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

[7] Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. Syntax-infused transformer and bert models for machine translation and natural language understanding. *arXiv preprint arXiv:1911.06156*, 2019.

[8] Yue Zhang and Jiawei Li. The death of feature engineering?——bert with linguistic features on squad 2.0. Technical report, Technical Report CS224n, Stanford University, 2019.

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

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

[11] 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`.

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Mayur Rustagi, R Rajendra Prasath, Sumit Goswami, and Sudeshna Sarkar. Learning age and gender of blogger from stylistic variation. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 205–212. Springer, 2009.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[26] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[27] Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. Perturbed masking: Parameter-free probing for analyzing and interpreting bert. *arXiv preprint arXiv:2004.14786*, 2020.

[28] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.