Gender Prediction Using Email Data for Algorithmic Fairness

Peihan Liu, Matrin J. Strauss

August 2021

Abstract

Gender identification is an essential subject of natural language processing. For example, authorship analysis can help people identify the authors' gender based on the texts and help people improve their writing to be more neutral. There are many well-known linguistic models to address this problem. Specifically, we use Naive Bayesian methods, including Bag-Of-Words (BOW) and Term frequency-inverse document frequency (TF-IDF), to predict gender and measure the "neutral degree" numerically. We also provide some concerns about the existing model and some potential directions to use the model for algorithmic fairness.

1 Introduction

Gender identification has been a topic of intense study with the rapid growth of technologies. Many essential models have been proposed to identify gender-based texts, such as BERT [1], RoBERTa [2], Word2Vec [3][4], GPT-3 [7], and many neural networks, including CNN, RNN, etc. Pre-trained models show impressive performance on gender prediction, as they reduce the need for training data. However, given enough data and computing resources, we use the classic Naive Bayes (NB) method [5] to identify gender using the Enron dataset [6], collected and prepared by the CALO project in 2001.

Moreover, we study the related algorithmic fairness problems. It is a relatively new area attracting much attention due to the growing importance of addressing social biases in machine learning. We analyze the usage of some interesting n-grams and punctuations statistically. Finally, we develop a tool for users to check their texts as to gendered languages based on our models and statistical results.

The material in this paper is organized as follows. In section two, we recall some basic definitions; in section three, we introduce the data we used and preprocess the data. Next, we present our results in section four, where we perform statistical analysis. The last section is about the future works and limitations of this paper.

2 Preliminary

The Naive Bayes model is based on the Bayes theorem stated as follows. The two techniques we use in our research are bag-of-words (BOW) NB and term frequency-inverse document frequency (TF-IDF) NB. We are going to illustrate them separately using the Bayes theorem.

Theorem 2.1 (Bayes Theorem)**.**

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

2.1 Bag-Of-Words (BOW) Naive Bayes

The conditional probability of the author who wrote the sentence " $word_1$ $word_2$ ... $word_n$ " being male is simply $P(male|word_1 \cap word_2 \cap ... \cap word_n)$. Hence, let A be the indicator variable of author's gender and B be the occurrence of words in a given sentence, i.e. the occurrence of $word_1 \cap word_2 \cap ... \cap word_n$, then we have the following,

$$\begin{aligned} & P(male|word_1 \cap word_2 \cap ... \cap word_n) \\ = & \frac{P(male \cap word_1 \cap word_2 \cap ... \cap word_n)}{P(word_1 \cap word_2 \cap ... \cap word_n)} \\ = & \frac{P(word_1 \cap word_2 \cap ... \cap word_n|male) \cdot P(male)}{P(word_1 \cap word_2 \cap ... \cap word_n)} \end{aligned}$$

based on the Bayes theorem. If we further assume the independence of the words, then we have a simpler form given by,

$$\begin{aligned} & P(male|word_1 \cap word_2 \cap ... \cap word_n) \\ = & \frac{P(word_1 \cap word_2 \cap ... \cap word_n|male) \cdot P(male)}{P(word_1 \cap word_2 \cap ... \cap word_n)} \\ = & \frac{P(word_1|male)P(word_2|male)...P(word_n|male)P(male)}{P(word_1)P(word_2)...P(word_n)} \\ = & \frac{\Pi_{i=1}^{n} P(word_i|male)}{\Pi_{i=1}^{n} P(word_i)}, \end{aligned}$$

where $P(word_i|male)$ is the probability of the occurrence of $word_i$ given that the author is male and $P(word_i)$ is the probability of the occurrence of $word_i$, i.e.

$$\begin{aligned} P(word_i|male) &= \frac{\#\text{ of } word_i \text{ in male class}}{Total\#\text{ of words in male class}} \\ P(word_i) &= \frac{\#\text{ of } word_i \text{ in all classes}}{Total\#\text{ of words in all classes}} \end{aligned}$$

However, BOW NB assumes the independence of words and does not include information on the grammar of the sentences nor on the ordering of the words.

Nonetheless, it might still better than word-embedding since the context of Enron is domain specific, which makes it harder to find corresponding vector from pre-trained word embedding models, such as Word2Vec.

2.2 TF-IDF Naive Bayes

Term frequency–inverse document frequency (TF-IDF) shows how important a word is to a dataset. It is often used as a weighting factor, which helps to adjust for the fact that some words appear more frequently. i.e. TF-IDF gives different measure. Similar to BOW, TF-IDF is only useful as a lexical level feature and can not capture semantics, but it values the rareness of words. It is especially useful when we need to catch a signal in a dataset. Also, TF-IDF is good at capturing text similarities, which is important for the algorithmic fairness problem we will discuss in the later part of this paper.

The definition of TF-IDF consist of two parts, TF and IDF, are given as follows,

$$TF(word) = \frac{\#\text{ of } word \text{ in a document}}{\#\text{ of all words in a document}}$$
$$IDF(word) = \frac{\#\text{ of all documents}}{\#\text{ of all documents containing word}}.$$

In practice, we usually use IDF's logarithm to make computations more reliable and the numbers more easily manipulated.

$$IDF(word) = \log \frac{\#\text{ of all documents}}{\#\text{ of all documents containing word}}$$

Similar to BOW, we have the following results based on the Bayes theorem,

$$P(male|word_1 \cap word_2 \cap ... \cap word_n)$$
$$= \frac{P(word_1|male)P(word_2|male)...P(word_n|male)P(male)}{P(word_1)P(word_2)...P(word_n)}.$$

However, we don't use the frequency to compute the probabilities. Instead, we use the definitions of TF and IDF, given by following,

$$P(word_i|male) = \frac{TF(word_i|male)IDF(word_i)}{\sum_{word_j} TF(word_j|male)IDF(word_j)}$$
$$= \frac{\frac{\#\text{ of } word_i \text{ in male class}}{\#\text{ of all words in male class}} \cdot \log \frac{\#\text{ of all messages}}{\#\text{ of all messages containing } word_i}}{\sum_{word_j} \frac{\#\text{ of } word_j \text{ in male class}}{\#\text{ of all words in male class}} \cdot \log \frac{\#\text{ of all messages}}{\#\text{ of all messages containing } word_j}}$$
$$P(word_i) = \frac{TF(word_i)IDF(word_i)}{\sum_{word_j} TF(word_j)IDF(word_j)}$$
$$= \frac{\frac{\#\text{ of } word_i \text{ in male class}}{\#\text{ of all words in male class}} \cdot \log \frac{\#\text{ of all messages}}{\#\text{ of all messages containing } word_i}}{\sum_{word_j} \frac{\#\text{ of } word_j \text{ in male class}}{\#\text{ of all words in male class}} \cdot \log \frac{\#\text{ of all messages}}{\#\text{ of all messages containing } word_j}}$$

3 Data

3.1 Dataset Description

We conduct extensive experiments on Enron email corpus, a publicly available real-world email corpus in English[6]. It is collected and prepared by the CALO project in 2001. There are around 1.7Gb messages from 114 users, among which 86 users are male and 28 of them are females. After discarding unlabelled emails, we have 24887 labeled emails, and among these, 16476 are from males and 8411 are from females.

3.2 Data Preprocessing

We use NTLK pacakge to preprocess the messages in the following ways.

- Lowercase: we change all capital letters to its lowercase.
  e.g. *'Eat'* $\rightarrow$ *'eat'*
- Tokenization: we split sentences to pieces, or in other words, we remove all whitespaces and punctuations.
  e.g. *'I like apples.'* $\rightarrow [I, like, apples]$
- Porter Stemmer: we normalize all the words by removing the common morphological and inflexional endings from words.
  e.g.: *'moved'/ 'move'/ 'moving'* $\rightarrow$ *'move'*
- Stop words: we discard some words that occur extremely frequently in any text but unnecessary. The full list of stopwords we used can be found in Appendix A.
  e.g. *is, to, we, ...*

However, we have some concerns regarding these preprocessing steps. For example, some punctuations or the frequency of common stop words might be good gender indicators. Hence, we analyze the occurrence of some punctuations and the frequency of common stop words statistically in section 4.

3.3 Data Inprocessing: n-gram

An n-gram is a contiguous sequence of n words from a given sentence, and an n-gram model is a probabilistic model for predicting the next word in a sentence of the form of an (n-1) - order Markov model [8][9]. For example, if the original sentence is 'I am on my way home now,' then different values of n will lead to different processed results shown below.

- unigram (or 1-gram): ['I', 'am', 'on', 'my', 'way', 'home', 'now']
- bigram (or 2-gram): ['I am', 'am on', 'on my', 'my way', 'way home', 'home now']

- trigram (or 3-gram): [‘I am on’, ‘am on my’, ‘on my way’, ‘my way home’, ‘way home now’]

We mainly work on the cases when n is less than or equal to 6, and we noticed that the difference between cases of various n is not significant. Hence, we set n to be 1 in this paper.

4 Results

4.1 Experiments on gender prediction

We include 70% of the email messages as the training data and the rest as the testing data; preprocess the data and use 1-gram. For comparison purposes, we present accuracies of BOW NB, TF-IDF NB, and two pre-trained NB models, Gaussian NB [10] and Bernoulli NB [11].

Table 1: Accuracies comparison

Method	Accuracy
BOW (Bag-Of-Words) Naive Bayesian:	0.67
TF-IDF Naive Bayesian	0.75
Gaussian Naive Bayesian (pretrained)	0.65
Bernouli Naive Bayesian (pretrained)	0.66

We also present the ROC curve shown below. The performance of TF-IDF NB is the best compared to BOW NB, GNB, and BNB. However, we observe some interesting results that might need further study. Given that the original data is imbalanced, we train these models with imbalanced and balanced training data, but the difference of accuracies is insignificant. Also, the accuracy of males and accuracy of females is different on average,

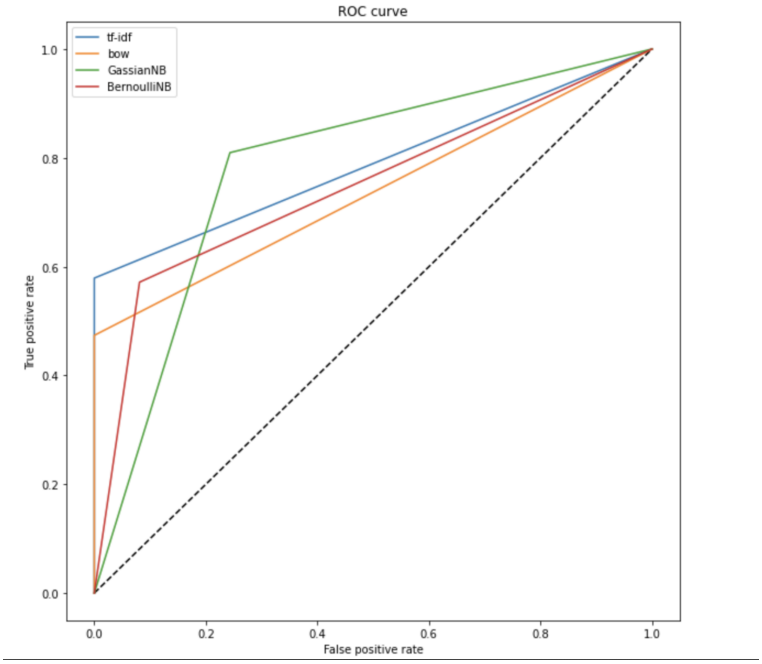
$$\frac{truthM\_predictedM}{truthM\_predictedM + truthF\_predictedM} \approx 60\%$$
$$\frac{truthF\_predictedF}{truthF\_predictedF + truthM\_predictedF} \approx 75\%.$$

4.2 Fairness

We also study the punctuations and common words that might be gender indicators. The statistical tool we employed is the Welch t-test. We only need to assume that male and female populations are normally distributed. Compared to the regular t-test, we do not need to require the variances to be the same.

The test statistics t is computed as follows,

$$t = \frac{\triangle \bar{X}}{s_{\triangle \bar{X}}} = \frac{\bar{X}_M - \bar{X}_F}{\sqrt{s^2_{\bar{X}_M} + s^2_{\bar{X}_F}}},$$



where $s^2_{\bar{X}_M}$ and $s^2_{\bar{X}_F}$ are given by

$$s^2_{\bar{X}_M} = \frac{s_M}{N_M}$$
$$s^2_{\bar{X}_F} = \frac{s_F}{N_F},$$

and $\bar{X}_M$ and $\bar{X}_F$ are the sample means of male and female, $s_{\bar{X}_M}$ and $s_{\bar{X}_F}$ are their standard errors, s_M and s_F are their sample standard deviations, and N_M and N_F are their sample size respectively. Note that the Welch–Satterthwaite equation approximates the degree of freedom ι,

$$\iota = \frac{(\frac{s^2_M}{N_M} + \frac{s^2_F}{N_F})^2}{\frac{s^4_M}{N^2_M \iota_M} + \frac{s^4_F}{N^4_F \iota_F}}$$

, where $\iota_M = N_M - 1$ and $\iota_F = N_F - 1$

We apply the Welch t-test on common punctuations and words and found that some punctuations and words are good gender indicators. The complete lists are in Appendix B.

5 Acknowledgement

I thank Professor Martin J. Strauss for leading me to think about this problem and for all the valuable discussions and feedback. This work is supported by

University of Michigan REU program.

Appendix A

Stopwords: ourselves hers between yourself but again there about once during out very having with they own an be some for do its yours such into of most itself other off is s am or who as from him each the themselves until below are we these your his through don nor me were her more himself this down should our their while above both up to ours had she all no when at any before them same and been have in will on does yourselves then that because what over why so can did not now under he you herself has just where too only myself which those i after few whom t being if theirs my against a by doing it how further was here than

Appendix B

Punctuations: . : ! # $ - ? ;

Common words: me my we our yourself he him she her they them which whom that these those am is are be have had a an the and but as while of at for with about through after to from in on over further then there how all more some such no not so will should now

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, `https://arxiv.org/abs/1907.11692`, 2019

[3] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space, `https://arxiv.org/abs/1301.3781`, 2013

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, Distributed representations of words and phrases and their compositionality, Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013

[5] Harry Zhang, The Optimality of Naive Bayes, Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, 2014

[6] Bryan Klimt and Yiming Yang, The Enron Corpus: A New Dataset for Email Classification Research, Proceedings of 15th European Conference on Machine Learning, online available at `https://www.cs.cmu.edu/~enron/`, 2004.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, Language Models are Few-Shot Learners, Proceedings of Advances in Neural Information Processing Systems 33, 2020

[8] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, Robert L. Mercer, Class-based n-gram models of natural language, Computational Linguistics, 1992

[9] William B. Cavnar, John M. Trenkle, N-Gram-Based Text Categorization, Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994

[10] George H. John, Pat Langley, Estimating continuous distributions in Bayesian classifiers, Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995

[11] Andrew McCallum, Kamal, Nigam, A Comparison of Event Models for Naive Bayes Text Classification, Proceedings in Workshop on Learning for Text