Development of gender classification using transformers

Saul Ramirez

Introduction

One of the main applications of Data Science is in marketing. The better you can understand your audience's demographic, the better you can sell to them. However, most users on the internet have become more skeptical about knowingly giving their information away, and typically avoid filling out their profile information. User gender information, or lack thereof, may produce heteroskedasticity further down our pipeline. As a result, this data should be imputed can be used in any recommendation system. For this project, I analyze a small dataset from Kaggle, consisting of 20,050 Twitter profiles to determine if gender classification is possible.

This project was adopted from the following tutorial:
https://www.analyticsvidhya.com/blog/2021/08/twitter-based-gender-classification-a-machine-learning-project/

Methods

As part of my implementation, I examine profile's handle, description, along with a sample tweet to predict the user's gender. I approach this problem is by using Naive Bayes classification, and XG Boost to get an initial estimate of the difficulty of the problem. Then I use a classification transformer and compare the accuracies at the end.

Although the topic of gender is very complicated and it is known that there are more than two genders, I will assume there are only two genders for simplicity. As there are only two classes, the statistical baseline of guessing correctly due to random chance is 50%, this is our accuracy baseline. The Kaggle dataset was a scraped dataset, out of the 20,500 profiles, only 11,194 instances belonged to Male and Female users; the rest of the profiles belonged to brands, or the gender was unknown. I only look at profiles from males and females and drop any instances with missing data. The text input is a combination of the profile handle, description, and text with commas in between. I used regex to clean punctuation and remove punctuation characters.

Table 1: Breakdown of data set classes

Female	5725
Male	5469
Brand	4328
Unknown	702

I noticed that many of the most common words were the same between the two genders and determined that this would be difficult. Words such as "Love", "Life", and "S" for sarcasm were a few of the examples.

Figure 1: Common words for woman (left). Common words for men (right).

The data was split into a 70-20-10, training-validation-test split. The validation and test spectra are relatively similar splits.

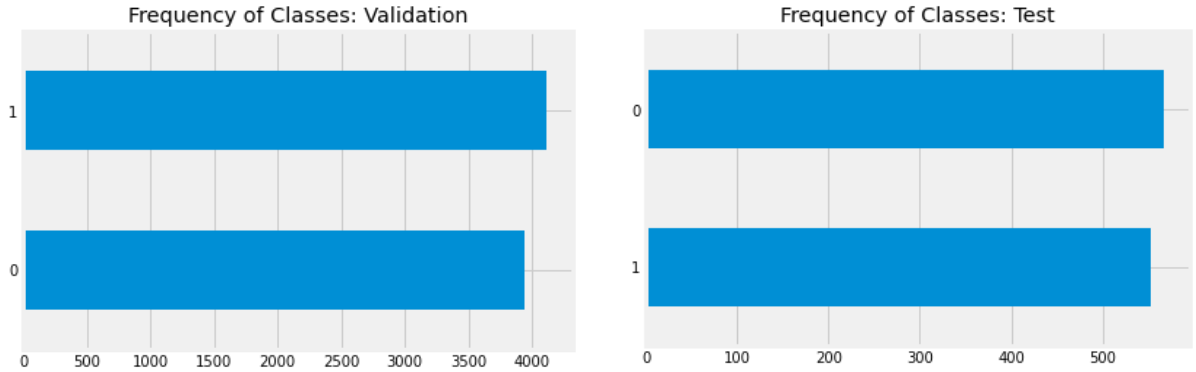

Figure 2: Splits for the validation and test data sets where 1 represents woman and 0 represents men.

Results

The first approach was using the naïve bayes and XG Boost. This method didn't work as well because I didn't remove stop words, and I only used a simple tokenizer instead of GLoVE. The results are shown in Table 2, even with XG Boost, the results are only slightly better than random chance. This gives us a good understanding that this could be a relatively challenging task. However, since the objective of this assignment is to apply transformers, I didn't focus too much on improving this approach.

Table 2: Splits for the validation and test data sets where 1 represents woman and 0 represents

Model	Accuracy	F1
Naïve Bayes Classifier	50.89%	0.42
XG Boost	54.02%	0.51

Next, I fine-tuned the BERT transformer for classification, without changing the dataset at all, I refer to this as the "Vanilla BERT". The results were significantly better than the simple machine learning approaches.

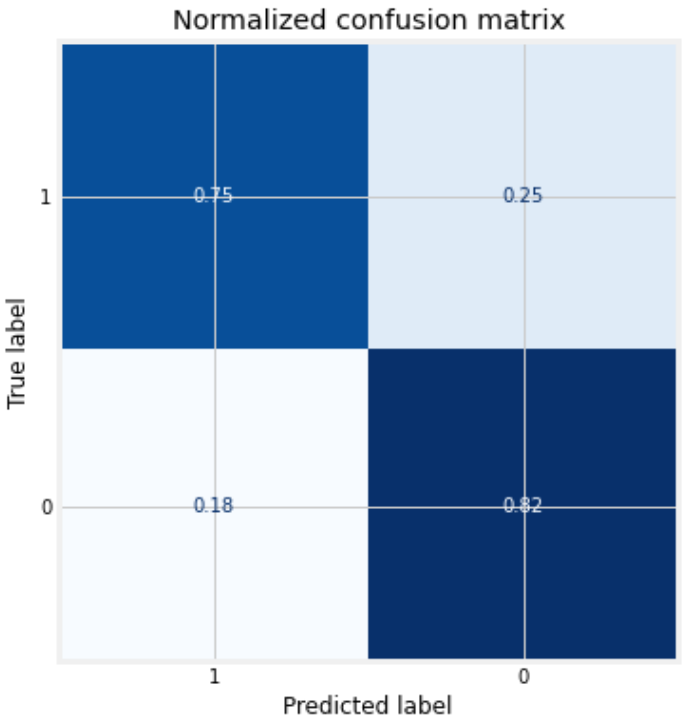

Figure 3: Fine-tuning BERT for gender classification.

I was interested in understanding if the information required to predict gender was mostly located in the profile description. Without using the handle or a sample tweet I retrained the BERT Transformer, as expected the results were not as good as the previous model but surprisingly good for having such low context. This approach had a F1 score of 0.69, which is still significantly better than the machine learning models.

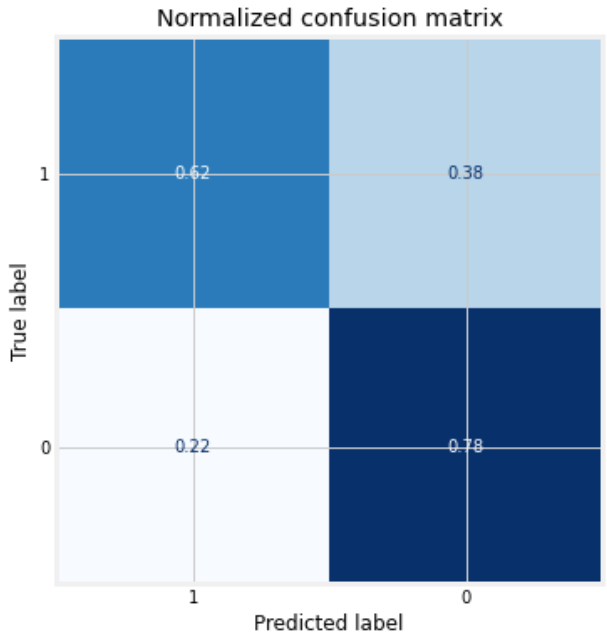

Figure 4: Splits for the validation and test data sets where 1 represents woman and 0 represents men.

Finally, I augmented the training data and fit a final model. To augment the data, I went through each of the text instances (handle, description, and tweet) and generated a number between 0 and 3 to determine if the data should be augmented with synonym replacement, random insert, random swap, random delete or backtranslation. Backtranslation was significantly slow, so I adjusted it to make it not occur as often. Once the data is augmented, the data is combined with the training set only as to keep the same validation and test sets.

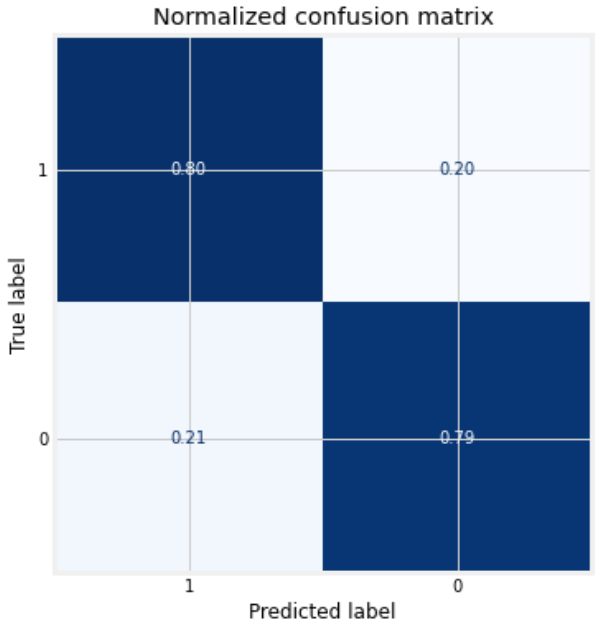

Figure 5: Splits for the validation and test data sets where 1 represents woman and 0 represents men.

The results for the three transformer models are shown in Table 3. As we see the Vanilla BERT does the best. Before developing this approach, the best F1 score was 0.69 for the Vanilla BERT, and didn't receive the large boost until the twitter handle was added to the data. The results for the transformers are shown in Table 3.

Table 3: Splits for the validation and test data sets where 1 represents woman and 0 represents

Model	Accuracy	F1
Vanilla BERT	78.80%	0.79
BERT Description Only	68.73%	0.69
BERT Augmentation	78.56%	0.79

Discussion

To test how well this approach worked I build an API for the BERT Augmented Model. The BERT Augmented Model was selected since it had similar accuracy to the Vanilla BERT but had seen more training examples. I manually derived input for three viral figures on Twitter: Andrew Tate, the epitome of toxic masculinity, Alexandria Ocasio-Cortez, a United States Representative, and Elon Musk. The input was made in the same format as the training data: handle, profile description, sample tweet. The model was able to classify all three examples correctly, but what was interesting was that the model had such high confidence in its prediction given it had seen such few examples.

Table 4: Influential Twitter Users

User	Input	Label	Prediction	P
Andrew Tate	Cobratate, Light-Heavyweight Kickboxing World Champion. Escape the Matrix, Mastery is a funny thing. It's almost as if, on a long enough time, losing simply isn't an option. Such is the way of Wudan	Male	Male	0.995
Alexandria Ocasio-Cortez	AOC, US Representative, NY-14 (BX & Queens). In a modern, moral, & wealthy society, no American should be too poor to live. 💯% People-Funded, no lobbyist💰. She/her.,I see people are rushing out to fill up their cars for this hurricane at the gas station This wouldn't be an issue if they had electric cars.If the power is out for a week how are they going to get gas? We need to start planning ahead and moving forward	Female	Female	0.809
Elon Musk	elonmusk, , twitter deal temporarily on hold pending details supporting calculation that spam/fake accounts	Male	Male	0.996

It was interesting that with AOC, even though her pronouns, “She/her” were present in her tweet, the model had the lowest confidence of the three examples. Perhaps the model was looking heavily at the name of the user. Below is an example of a sample profile with variations of the name “Dan” adjusted to make the gender of the user ambiguous. We see that Dan and Danny, typically Male spellings were classified as Male with extremely high confidence. While Dani and Dany, Female spellings, were classified as Female with extremely high confidence as well. In these situations, it seemed like the description or tweet didn’t matter as much as had been previously shown in the results.

Table 5: Gender Bias of name spellings

Text	Prediction	P
Dani, Live, Laugh, Love, Merry Christmas Everyone.	Female	0.996
Dan, Live, Laugh, Love, Merry Christmas Everyone.	Male	0.986
Dany, Live, Laugh, Love, Merry Christmas Everyone.	Female	0.996
Danny, Live, Laugh, Love, Merry Christmas Everyone.	Male	0.889

Next, I tried making a generic lesbian profile to see if the change in pronouns would confuse the model, but it did not. However, by changing the handle name from BigGay to Big_Gay was all it took to get the model to change it’s mind from Female with 0.994 confidence to Male with 0.799 confidence- despite the rest of the text staying the same and being indicative of the correct classification.

Table 6: LGBTQ Example

Text	Prediction	P
BigGay, Cute, sweet and even funny, Staring at her and thinking, How did a girl like her end up with a girl like me.	Female	0.994
Big_Gay, Cute, sweet and even funny, Staring at her and thinking, How did a girl like her end up with a girl like me.	Male	0.799

This method was applied to my followers on Instagram as well with a similar format: handle, bio, image caption and it was able to correctly classify all 5 users. This shows that it's helpful across platforms.

Conclusion

The power of transformers is displayed with this task, as with very little data and feature engineering we were able to obtain almost 80% accuracy. It is surprising is that there is a distinction between the way that men and women present themselves online, and even more surprising is that these patterns can be distinguished using AI models. At first, I was concerned about the ethics of this problem and my approach since it seemed like I'm perpetuating gender bias to classify profiles online. The fact that I couldn't get higher F1-scores may shed light on the fact that gender may not be cleanly separable. Since both genders speak the same language, therefore the word distribution per class is probably very mixed. However, this is very useful in Marketing, as if we were talking about a identifying people who would be interested in female leggings, it would be inefficient to market to people who identify as Men.