Transfer Learning of initial dataturks Model

Objective

Use the initial model trained on the dataturks dataset to and the dataset from Zeerak Waseem and transfer learning to build a model that can identify misogyny.

Description

This data comes from two of Zeerak Waseems's studies (see below for references). The data was originally split across four .json files comprised of Tweets and (some of) their associated meta-data, in addition to an annotation. These annotations are one of `*"racism"*`, `*"Sexism"*`, `*"sexism"*`, `*"Both"*`, `*"none"*`, or `*"Neither"*`. The dataset combined the neither and sexist files into one from Zeerak's dataset. They were collected using search terms to do with the Australian cooking show My Kitchen Rules (#mkr). Kat and Andre are the names the hosts of the TV show. There are some other terms but I need to find the email.

The dataturks dataset comes from a website where labeling is crowdsourced. These were labeled as harassing or not and not as misogynistic.

Four studies were complete,

1. The original dataturks dataset was studied
2. All the layers of neural net were fine-tuned
3. Last layer fine-tuned
4. Trained only on the sexism dataset from Zeerak

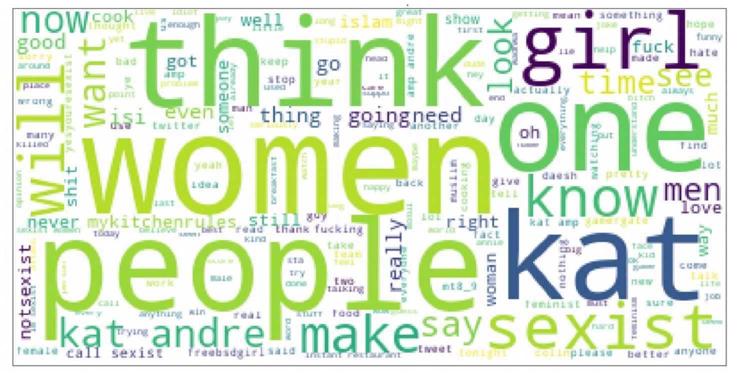
These were completed using gloVe word embeddings and a feedforward architecture. The normalization steps were to remove basic punctuation stopwords and certain hashtags such.

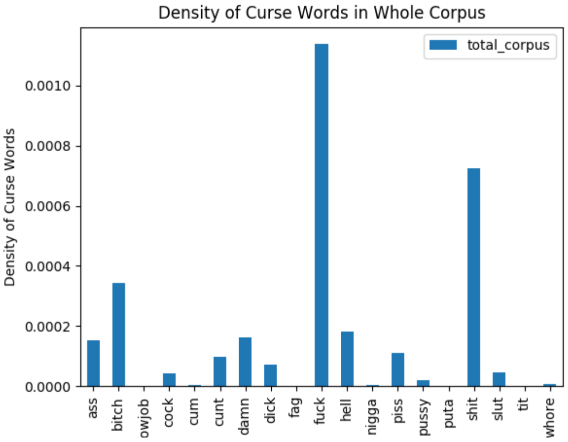
The original model trained only on the dataturks dataset, was able to reach 81% accuracy and thus we shall use this for the baseline.

Initial Corpus Exploration

Common words, curse words and non-lexical data

Here you can see a word cloud of the total corpus and densities of curse words in the same corpus. As you can see women is a common word in the corpus, along with girl and Kat. The most common curse words are fuck and shit, followed by bitch.

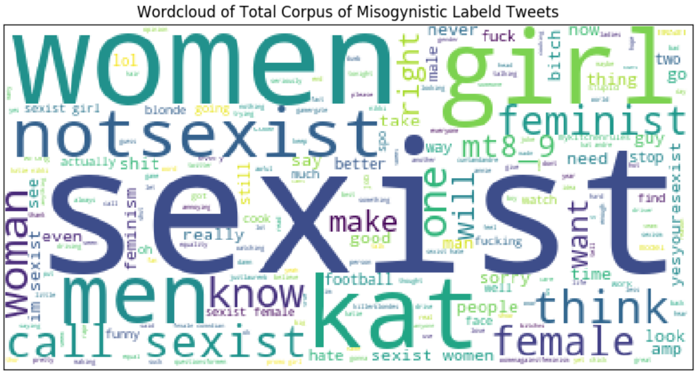


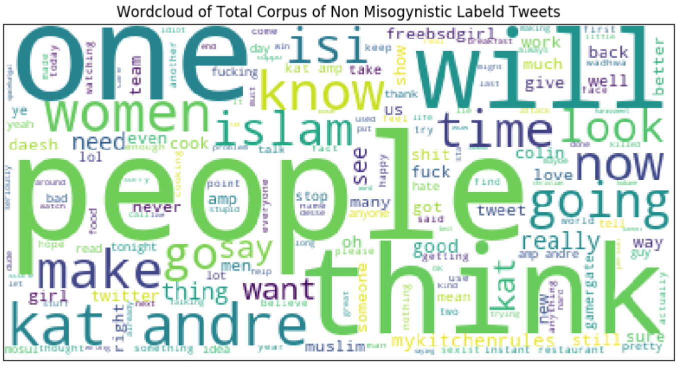


Here would normally go the non-lexical analysis, but there is non of that as this was pre-cleaned with the punctuation removed.

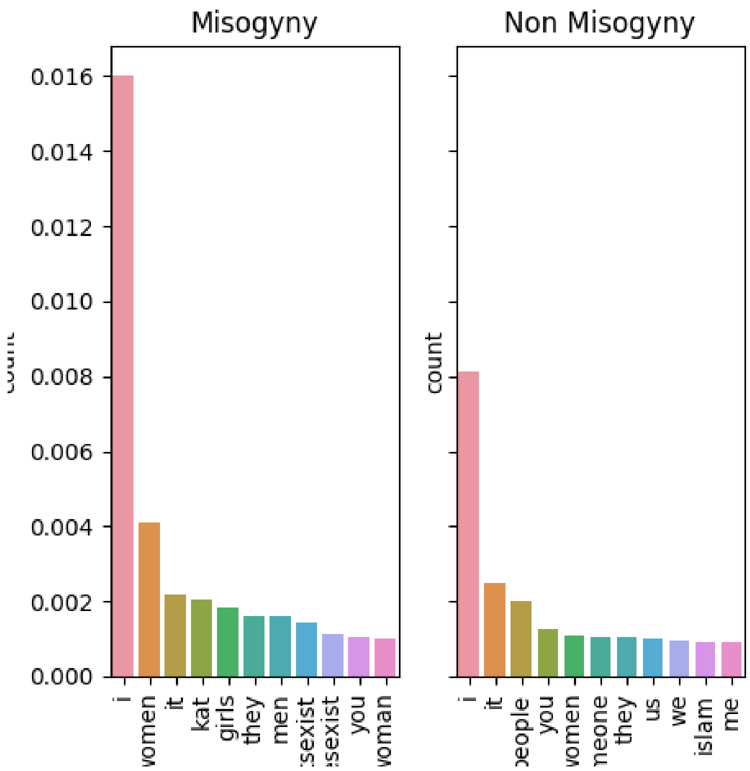
Initial Corpus Exploration split by Annotator Label

It is clear that terms like women, notsexist, bitch, occur more frequently in misogynistic labeld corpus than not.

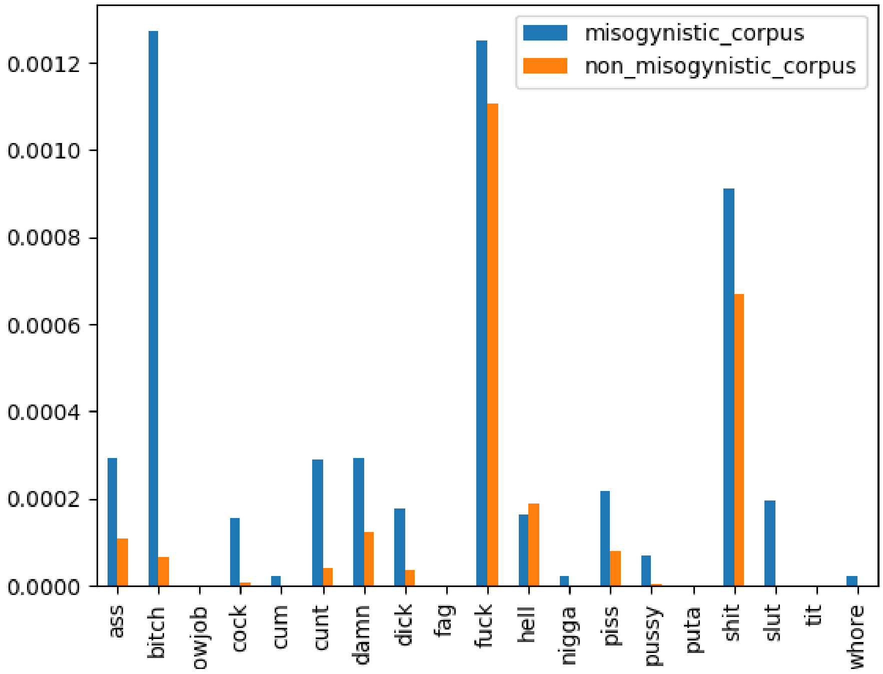




The top 10 most common nouns in the misogynistic and non misogynistic labeled tweets. Men appears more in the misogynistic but not in non misogynistic, implying that maybe there is often a comparative in misogynistic tweets. ie. women can’t drive as well as men.

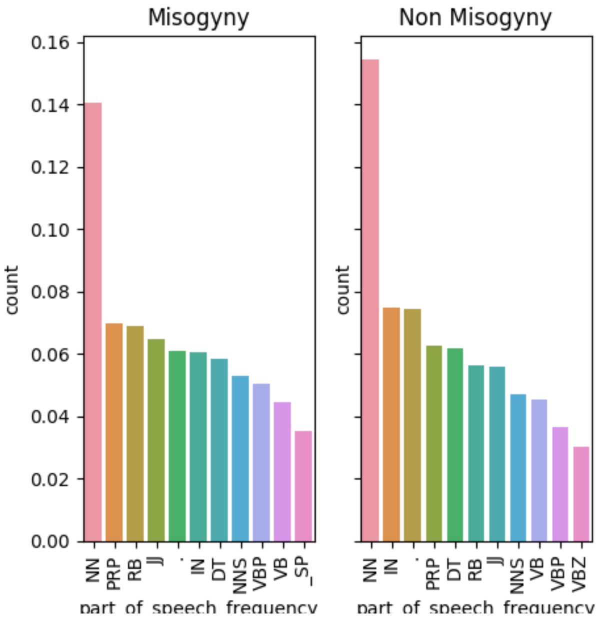


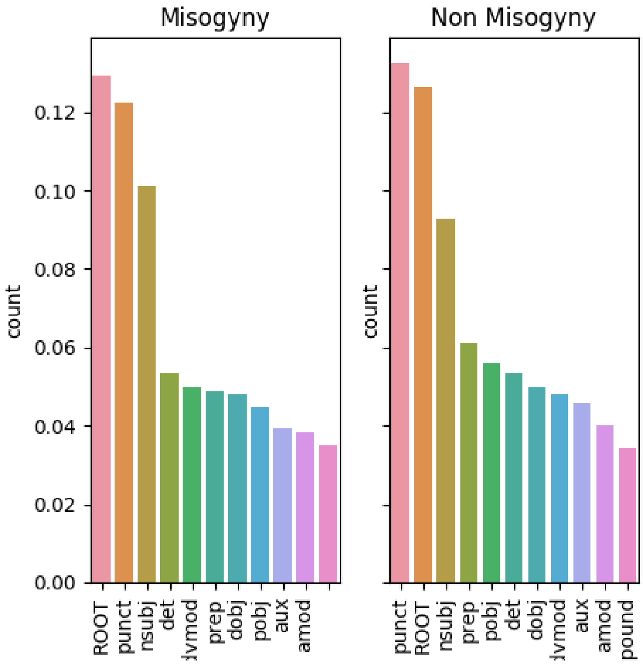
Misogynistic labels seems to have more curse words than non misogynistic, particularly the word bitch.



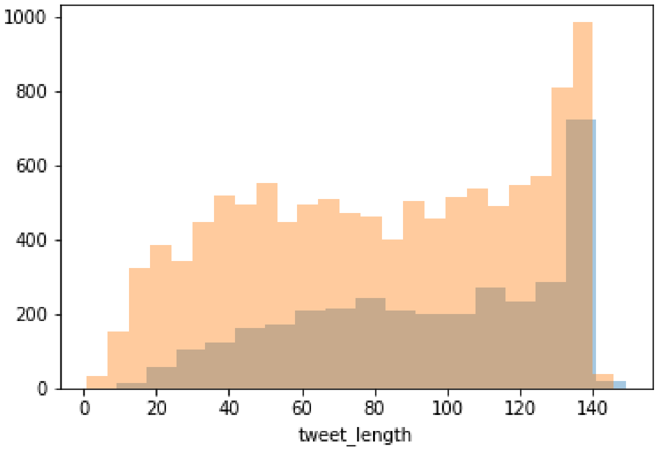
Now for some basic syntax frequency counts. The left hand graph is the part of speech count.

The right hand graph is a count of the the syntax relationships





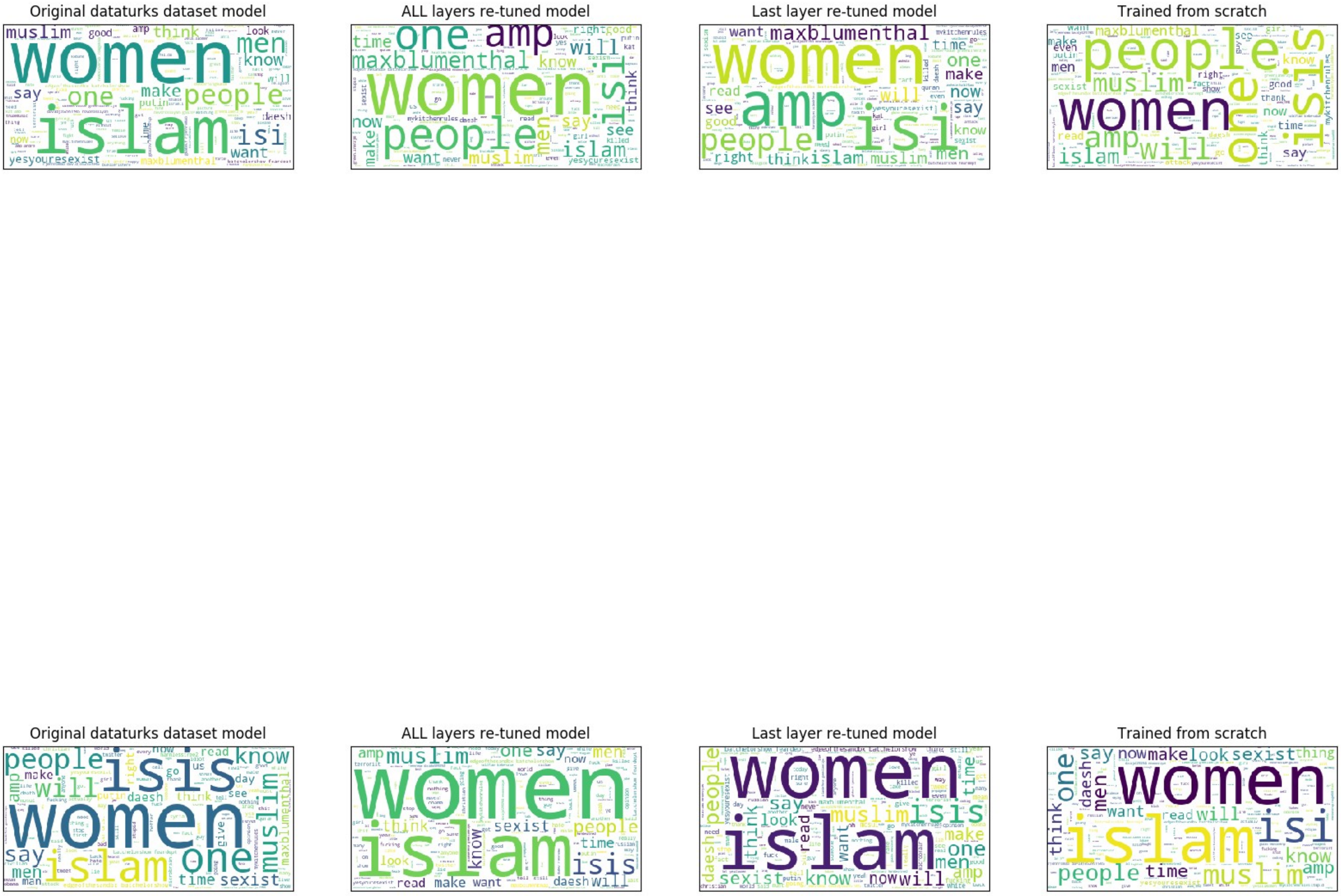
Orange is tweet length of non-misogynistic and blue is misogynistic. They have very similar shapes so the tweet length seems to be independent of the type of tweet it is. (it’s not normalized so ignore the y axis and just look at the distribution)



Corpus Exploration of Predictions split by Annotator Label

Top layer is misogynistic and bottom layer is not. As you can see, islam is in the misogynistic corpus trained with original dataturks which had harassment labels. This makes sense as this is how it was previously labeled. The word vanishes from the word clouds as we move the the right of the top level, indicating that the model is able to transfer the learning to our new misogynistic labeled corpus.

Wordclouds of Misogynistic and Non-Misogynisitic Words in Corpus using Different Models



Original dataturks dataset model

ALL layers re-tuned model

Last layer re-tuned model

Trained from scratch

Original dataturks dataset model

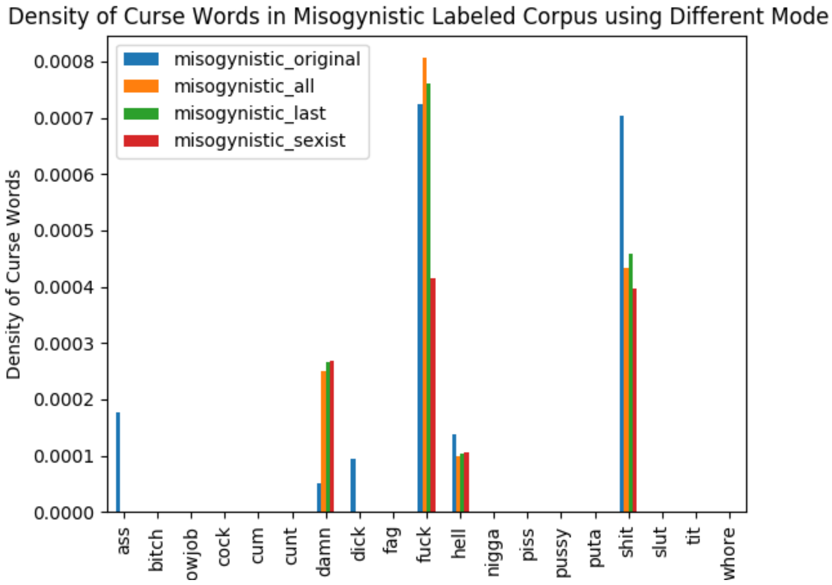
ALL layers re-tuned model

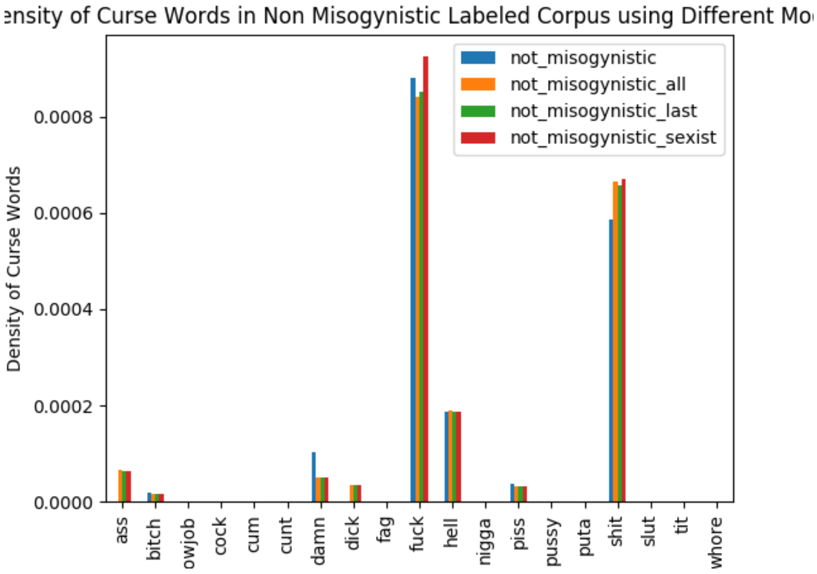
Last layer re-tuned model

Trained from scratch

The word women is a common word in all of the models, perhaps this means that there is some biasing on the word women. In other words, if the model is being told that all the tweets with the word women in it are misogynistic, its going to learn that rule.

The below graph shows the density of curse words in the predicted labeled dataset. As we can see non of the models follow a similar to curse words as the whole corpus did initially.





This is making me think that something is up with the models.

Evaluation of Models

We want to make sure we don’t label anything wrong. That way we build trust and make sure we don’t create any echo chambers online, so we want bottom RH corner to be min. Seems that model trained from scratch does that best.

Evaluation of model with ALL layers fine-tuned:

```
acc: 87.42%, loss: 0.34
Confusion matrix:
[[2181  139]
 [ 236  426]]
Marco F1:0.807606
Micro F1:0.874245
Weighted F1:0.870562
```

Evaluation of model with LAST layer fine-tuned:

```
acc: 87.22%, loss: 0.34
Confusion matrix:
[[2194  126]
 [ 255  407]]
Marco F1:0.800640
Micro F1:0.872233
Weighted F1:0.867065
```

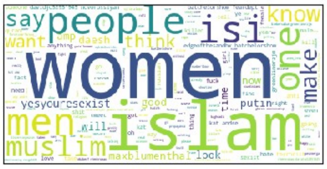
Evaluation of model trained from scratch:

```
acc: 88.46%, loss: 0.29
Confusion matrix:
[[2216  104]
 [ 240  422]]
Marco F1:0.819205
Micro F1:0.884641
Weighted F1:0.879681
```

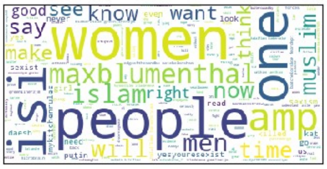
An interesting start but we still have some way to go.

What did we get Wrong?

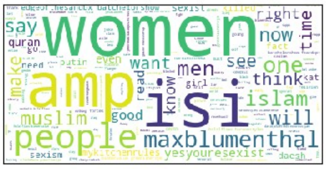
Here are some word clouds of the incorrectly labeled tweets across the four models.



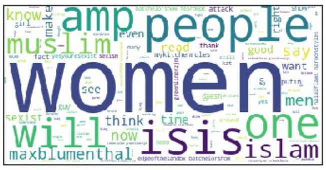
Original dataturks dataset model



ALL layers re-tuned model



Last layer re-tuned model

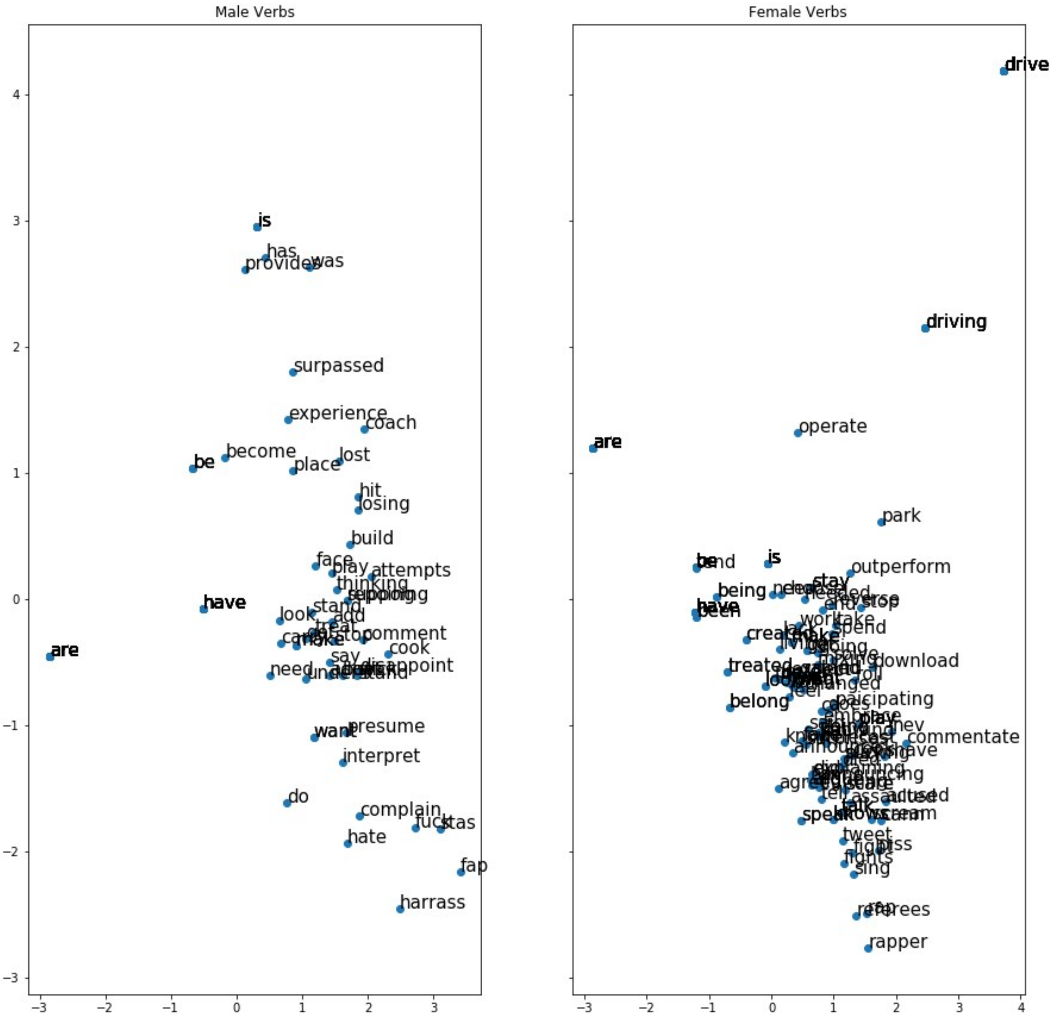


Trained from scratch

Women seems to be a feature that has been learnt. Not good!

A little Extra

This plot shows the verbs linked to the words man or men and women or woman in the whole corpus. As you can see, there is some clusternig/differents. This could be helpful for clustering or quantifying the degree to which language about women/men is different.



Male Verbs

Female Verbs

TO DO still

Proper error Analysis

- Compare to other model such as random forest and analyse the differeces between the labelings
- ROC-AUC graph
- Sensitivity
- Measure bias

Further Syntactic Study

Other bits and pieces such as dialect detection, influence of user profiles of classificaiton