Resume Gender Analysis Progress Report

Aparimit Chandra* aparimitchan@umass.edu
Long Le* lnle@umass.edu
Hannah Lerner* hmlerner@umass.edu

1 Initial Steps And Progress So Far

In our project proposal, we provided the following template as tasks for our project:

1. **Creation of the Dataset**: We would have to first extract text from the PDF of the resumes so that we can do text analysis on the document. We will run a simple metadata extraction library like pypdf2 for a first pass and then run a OCR model on any resumes that pypdf2 was not able to extract text successfully.

2. **Bag-of-words search**: Once we have our resumes sorted, tokenized and cleaned, we will be able to pull the most commonly occurring words from the male resumes and the female resumes.

3. **Supervised learning**: We use the ground-truth gender as label and the content of each resume as features. The features are preprocessed to reasonably remove any gender indication (e.g. remove names, pronouns). Then, a supervised binary classifier can be built to label resume's gender. We would like to see if supervised classifier would be able to segment resume into two classes reasonably well. This would indicate some statistical differences in the content of resumes from different genders.

4. **Unsupervised Clustering**: We would like to cluster resumes into 2 groups based on the each resume's content. We have several ways to do this. For example, we can get a BERT embedding from each resume and run a k-mean clustering on them using k=2. We can also use other algorithms like latent semantic indexing or stochastic block models. Note that for each resume, we have the ground-truth label for the resume's gender. The goal here is to see whether resumes of the same gender would be in the same cluster, and thus we choose the number of clusters to equal to 2. A metric we can use is "node impurity" (see Decision tree). A higher purity metric means that the unsupervised clustering algorithm highly segregates resumes from different genders.

At this point in our work on the project, we have successfully created our dataset and performed a couple basic bag of words analysis tasks on the newly created dataset. We have also created confusion matrices between each annotator with each other as well as with the ground truth. Our bag of words analysis focuses on finding which words are most commonly associated with male resumes, and which words are most commonly associated with female resumes. We performed computations using the TF-IDF metric in an attempt to see how male resumes were similar to female resumes. For our final report, we plan to go more in depth with our analysis utilizing supervised and unsupervised learning algorithms to explore potential algorithmic bias that might occur between resumes. We will be using supervised learning methods as a baseline to explore how unsupervised methods perform.

2 Dataset

Before we build the resume dataset, we first had to figure out the gender of all the resumes and try to get a ground truth value out of it. We decided to use https://gender-api.com/ to help us figure out the gender of each resume. This site works by letting us feed in a first and last name for it to then output it's predicted gender for that name. We had to pay a small fee to use this site (approximately 10 USD) so we made sure to be very careful when setting up the calls to it. The api was able to predict the gender based on name with a fairly high degree of confidence for almost all resume names, but there were a few that slipped through the cracks (primarily non-english names). We went through and manually gendered these resumes to make up for it. Sometimes, the Linkedin profiles of these people could be found so their gender were exactly labeled. If we were not certain on a specific name, we marked the gender to be unknown.

We created our own dataset for this project. It consists of 726 resumes that have been converted into text format. We were allowed access to the resumes from HackUMass. The specific resumes we used came from the HackUMass VI collection. Each resume came in as a pdf and went through certain preprocessing steps. First, we went through each resume from our collection (approximately 1800 total resumes), and then ran an OCR program over it to capture the text + the approximate underlying structure of the resume. We saved each of these resumes in a text file with the name being the id of the resume pre-supplied to us and the contents being the text captured from the OCR program. Each resume took approximately 1-3 seconds to modify to the desired format. We ran the OCR program on all of the HackUMass VI resumes supplied to us. We chose to take this route to gather resume data as the alternative of using PyPDF2 caused us major issues and general high loss rates in the resumes (many were not in a format compatible with the PyPDF2 library). This process was a little bit slower, but we feel like we were able to get generally more accurate readings from this method than when using PyPDF2. Once we had our collection of resumes, we went about anonymizing them so they could be used for annotation. Here we took a less elegant approach and simply removed all information before the word "Education" as we saw that it overwhelmingly was the first section after someone's name and personal information. We ran our anonymizing script on all the resumes output by the OCR program. After taking out the names/personal information from the resumes text files, we went through all the resumes and checked what gender each one was. In an attempt to balance out the dataset, we took all female resumes (which we found using the process described above) and (1000 - the number of female resumes) male resumes and put them together into a dataset that had in total 1000 resumes [Note: unfortunately during this process we did not get 1000 total resumes out due to an error in one of the scrapers we wrote and did not notice until we had almost completely finished annotating what we had. We decided that adding in the proper number of resumes would result in biased data as we would know that they are all going to be male and would alter the data negatively. Because of this we chose to not fix the number to be a proper 1000 resumes and instead have a dataset with 788 resumes. Another 62 were lost during the annotation process leaving us with 726 usable resumes.]. We then went on to label each resume during our annotation process.

3 Annotation

3.1 Annotation Experiment

We designed our annotation task to be repeated 4 times over by different annotators of different genders. Each annotator was assigned to read over the same set of resumes (as described in the previous section) and answer two primary questions: "Is this resume good or bad?" and "Do you think this resume was written by a man or a woman?". Our goal was to determine if there was any bias between annotators of different gender in how they answered the questions. We were looking to observe things such as male annotators marking more resumes to be male, female annotators marking more resumes to be female, annotators of either gender favoring their own gender when determining whether they thought the resume was good or bad (i.e. if they thought the resume was written by one gender was there any bias in how they thought of the content of the resume or alternatively, if they thought a resume was good, were they more or less likely to think it was written by a man or by a woman.).-

3.2 Gender vs Quality Metrics

Since we had four annotators (2 male and 2 female), we wanted to see if gender influenced your annotations and we also wanted to see if some inherent bias influences their judgement while judging the quality of a resume. We list out our findings from each of the annotators in this subsection:

- Male 1 classified resumes that he perceived to be female, bad roughly 48% of the time and classified male(perceived) resumes bad roughly 45% of the time.

- Male 2 classified resumes that he perceived to be female, bad roughly 69% of the time and classified male(perceived) resumes bad roughly 63% of the time.

Gender	M 1	M 2	F 1	F 2
M 1	1	0.295	0.312	0.176
M 2	0.295	1	0.266	0.157
F 1	0.312	0.266	1	0.183
F 2	0.176	0.157	0.183	1

Table 1: Cohen's Kappa agreement on the gender task (M for male and F for female).

- Female 1 classified resumes that she perceived to be female, bad roughly 28% of the time and classified male(perceived) resumes bad roughly 51% of the time.
- Female 2 classified resumes that she perceived to be female, bad roughly 16% of the time and classified male(perceived) resumes bad roughly 41% of the time.

From a quick analysis it appears that females were much less likely to classify a resume that they perceived to be female as bad whereas the males classification of good or bad had less of a correlation with the perceived gender. One thing to keep in mind though is that this could also be due to the fact that the annotators had a difference in opinion of what constitutes a good resume, we quantify this in the next section with the inter annotator agreement metrics

3.3 Inter-Annotator Agreement

We used Cohen's Kappa agreement as a quantitative measure of inter annotator agreement. with a value of 0.4 to 0.5 being moderate agreement as this can be seen as a sentiment analysis task and 0.1 to 0.2 being poor agreement. We had low agreement across all annotators on the gender task which leads us to believe that it is really hard to classify for gender resumes without relying on keywords as there were no apparent syntactical differences between male and female resumes. We had significantly higher agreement for the good and bad task which is more in line with our initial thoughts as it means the annotators broadly agreed on good and bad resumes. Female 2 had low agreement rates across the board for both tasks so we attach less importance to conclusions drawn from her annotations. The Kappa values are given in tables 1 and 2.

Quality	M 1	M 2	F 1	F 2
M 1	1	0.473	0.563	0.176
M 2	0.473	1	0.412	0.187
F 1	0.563	0.412	1	0.259
F 2	0.227	0.187	0.259	1

Table 2: Cohen's Kappa agreement on the quality task (M for male and F for female).

Gender	Precision	Recall	F1	Accuracy
M 1	0.68	0.75	0.72	0.66
M 2	0.63	0.93	0.75	0.66
F 1	0.64	0.70	0.67	0.61
F 2	0.64	0.66	0.65	0.59

Table 3: Annotator metrics against the ground truth on the gender task.

3.4 Accuracy vs Ground Truth

The annotators accuracy vs the ground truth on the gender tasks were quite low across the board leading us to believe it is not really easy to distinguish between male and female resumes without keywords. This makes us believe the gender vs quality depends more on perception of the gender rather than the actual gender. This also is in line with our inter annotator agreement findings as even when the annotators had agreement over good and bad meaning they were using similar reasoning to decide between good and bad resumes, there was a significant difference in how they judged the quality of male resumes vs the quality of female resumes. The metrics are available in table 3.

4 NLP Algorithms

4.1 Document Similarities

Using the TF-IDF (Term frequency -inverse document frequency) algorithm, we transform each of the 788 resumes into a vector of $\mathbb{R}^L$, where L is the vocabulary size ($L = 19231$ in our case). Then, we can calculate the cosine-similarity between any two resumes. We compute the average similarity between document class as in Table 4.

Similarity	male	female
male	0.0525	0.04926
female	0.04926	0.0532

Table 4: TF-IDF Similarities between resume classes.

	Most important words
male	experience, development, new, courses, history,biology, work, time, darthmouth, honors, gpa, **captain**, city, medical, computer,myanmar, control, theory, westwood,first, senior, **hindi**, developed, **latin**, general,**soccer**, ii, **eagle**, leadership, ect
female	**women**, advanced, **smith**, regional, social, house, cs, sciences, student, members, study, acton, dec, worcester, minor, **girls**, word, mit, introduction, national, coursework,basic, winchester, sat, quality, application, apr,html, dominion, **china**

Table 5: Log-Reg feature importance.

As seen, male and female resumes are 4.9% similar while resumes from the same class (either all males or all females) are about 5.2-5.3% similar. As such, there is not a lot of distinction between male and female resumes, at least in terms of TF-IDF features.

TF-IDF is used here since it is a common technique in information retrieval to assign an importance value to each word in each document.

4.2 Feature Importance Using Log-Reg

We obtain a bag-of-word (BOW) representation for each resume, and fit a binary classification Logistic Regression model (Log-Reg) on them using the male-female ground-truth labels. The average test accuracy for five-fold validation is about 64%, which is very close to the best human's accuracy.

We also include the most important words for each class in Table 5. For binary classification, the feature importance is just the Log-Reg coefficients. A high positive coefficient means that the word is pushing the model towards a ”male” prediction while a very negative coefficient means ”female”. We bolded some of the interesting words. For example, **Smith** College is a all-women college in Massachusetts. **Eagle** Scout is a rank in the Boy Scouts of America organization. The word **women** is the most important predictor for the ”female” classification of the Log-Reg model. This matches with our observation through manual annotation that a lot of females are part of organizations such as **Women** in Computer Science or **Women** in computing.

There are a lot of other methods to obtain feature importance for this binary classification task as well. Instead of a Log-Reg model, one can also fit a attention-based algorithm (such as BERT or LSTM with attention (Lai et al., 2019)), and use the attention values as feature importance measure. Another approach is to use the Shapley value from Game Theory. These more complex and time-intensive approaches will be explored in future iterations of this project.

References

Vivian Lai, Zheng Cai, and Chenhao Tan. 2019. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 486–495, Hong Kong, China. Association for Computational Linguistics.