Resume Gender Analysis in Humans and Machines

Aparimit Chandra*
aparimitchan@umass.edu

Long Le*
lnle@umass.edu

Hannah Lerner*
hmlerner@umass.edu

Abstract

Modern Natural Language Processing (NLP) techniques have become very adept at analyzing, categorizing, and performing various other tasks on texts. However, these algorithms are often reliant on datasets created by human annotators. This can be an issue when it comes to analyzing certain pieces of writing. In particular, resumes can be vulnerable to gender biases that people may hold. Additionally, resumes have increasingly been processed by automatic Machine Learning (ML) models where gender might also become a factor. In this paper, we explore the two following areas: (1) human bias in anonymized resume review, and (2) ML models' ability to predict gender in those resumes.

1 Introduction

Nowadays, resume screening is often the first step facing job-seekers in their search. A key thesis in (Derous and Ryan, 2019) is that nonjob-related factors (*e.g.* race) might lead to unethical discrimination in this screening stage. Specifically, the applicant's gender has been shown to be a major source of discrimination. For example, (Moss-Racusin et al., 2012) conducted an experiment where the same resume for a lab manager position is either named *John* or *Jennifer*. They presented the resumes to a group of scientists, and found that the female candidate was often deemed as less qualified. Anonymization has been proposed to combat this problem. (Muñoz, 2019), for instance, suggests a blind hiring practice in the Philippine recruitment process.

In the United States, many recruiters today also anonymize incoming resumes before review. While anonymization is an important component in a bias-free system, it still does not completely solve the problem of gender bias. For example, if an applicant is the president of *Women in CS* society at *Smith College*, then it is highly probable that this person is a female. On the one hand, this gender indicator would most likely affect the gender perception of a resume reviewer. On the other hand, we might not want to remove this indicator in the anonymization process as it, however revealing, is a legitimate and important piece of the person's resume. With this problem in mind, we design two annotation tasks to explore human bias in judging the quality of a resume in the presence of possibly strong gender indicators.

In addition to human bias, it is also important to understand how machines perceive gender. With the rapid adoption of automation, recruiters often rely on *resume search engines* to filter and search for candidates in a massive pool of applicants. (Chen et al., 2018)'s research on resume search engines revealed that there is unfairness associated with the inferred-gender of each resume. In this paper, we would like to determine how much ”genderness” is detectable by ML models. First, the supervised models are trained specifically to detect gender in anonymized resumes. Their performance is analyzed. Further, the trained feature importance is used to detect gender-indicating remains from the anonymization process. Second, the unsupervised models are used to cluster resumes. There, we seek to understand when gender is not a label, how much genderness is detectable in the clusters anyway.

The organization of this paper is as follows. In Section 2, we describe the resume collection and anonymization process. Section 3 describes the annotation experiment and its findings. Sections 4 and 5 describe the supervised and unsupervised models respectively. Section 6 is the concluding remark.

2 Dataset

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.

2.1 Generating Gender Ground Truth

An important challenge is to determine the ground-truth value for the gender of each resume. Because we did not have access to the resume metadata that HackUMass collected, we decided to use `https://gender-api.com/` to generate our gender ground truth value. The site predicts the gender of each resume based on the applicant's name.

The API was able to produce predictions with a fairly high degree of confidence for almost all resume names. However, there were some inaccuracies, especially with non-English names). We went through and manually gendered these resumes to remedy it. Sometimes, the Linkedin profiles of those people could be found so their gender was exactly labeled.

2.2 Preprocessing

Each resume came in as a pdf and went through the preprocessing steps shown in Figure 1. First, we went through each resume from our collection (approximately 1800 total resumes), and ran an Optical Character Recognition (OCR) program to capture the text and the approximate underlying structure of the resume.

2.3 Anonymization

After the resume texts were obtained, we took a simple approach to anonymization. Here, we removed all information before the word "Education" as we saw that it overwhelmingly was the first section after someone's name and personal information. Some resumes, however, might have the "Education" section at the end. We attempted to rectify this issue by identifying the resumes that were too short after anonymization, and removing those. There are more sophisticated methods for data anonymization. For example, Microsoft's Presidio API uses a host of techniques such as Named Entity Recognition and Regular Expressions, to remove sensitive information like names and locations. However, we found that this API failed in detecting non-English names and information associated with those names, for example links to Github.

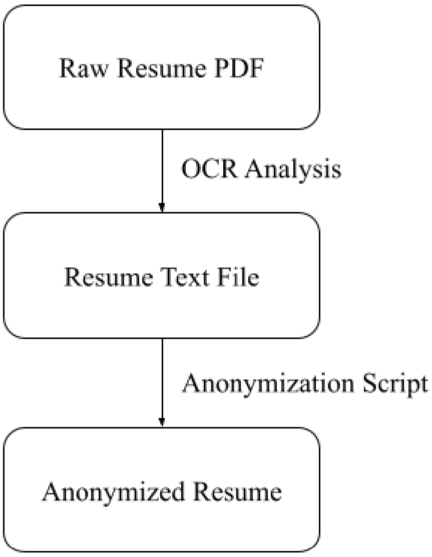


Figure 1: A diagram of the data processing pipeline that each resume pdf undergoes.

2.4 Data Loss & Statistics

We noticed that there were just under 400 female resumes out of 1800 ones. Thus, the aim was to obtain 1000 resumes of fairly balanced genders. Unfortunately during this process, we did not get 1000 total resumes due to an error in one of the scrapers 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 removed during the annotation process due to the lack of quality of the scraped resume. This left us with 726 usable resumes. We then went on to label each resume during our annotation process.

The statistics for those resumes are included in Table 1. Figure 2 is the word cloud for the 726 resumes used in our study. As can be seen, the most observable words in the tech resumes include *project*, *computer science*, *experience*.

Category	# of Resumes
Total Resumes	1892
Usable Resumes	726
Male	448
Female	340

Table 1: Statistics of the resume bank

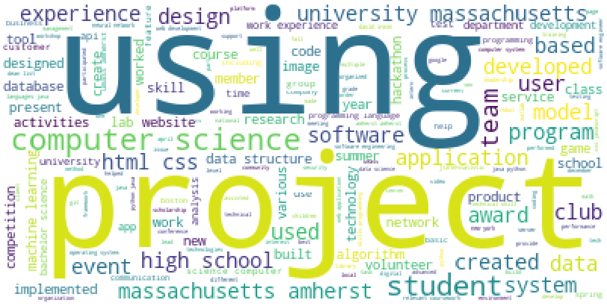


Figure 2: Word cloud for the resume collection.

2.5 Inclusivity During Data Gathering

While this method of gathering ground truth gender values is fairly accurate, it unfortunately leaves out people who may identify as non-binary. Nonetheless, it is difficult to determine non-binaryness without explicit metadata. Although HackUMass does collect such data, it had long been discarded. We had access to the HackHer resume database which did have some gender metadata, but we ultimately decided that mixing resumes targeted for two different hackathons might skew the later annotation task.

3 Human Bias Analysis

3.1 Annotation Experiment

We designed an experiment inspired by (Moss-Racusin et al., 2012) aimed at detecting the correlation between the applicant's perceived gender and quality. Each human annotator was assigned to read over the same set of HackUMass resumes 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?".

We have 4 annotators: two males and two females. As such, we can also explore if the annotator's own gender affects how they answered the questions. Before the experiment, we were looking to observe patterns 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

In this subsection, we are interested in the interplay between gender, both of annotator and applicant, and assigned resume quality. Due to the small number of annotators, the findings from each annotator is listed below.

- 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.
- 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 female annotators 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 (Cohen, 1960) as a quantitative measure of inter-annotator agreement, with a value of 0.4 to 0.5 as being moderate agreement and 0.1 to 0.2 as being poor. We had low agreement across all annotators on the gender task which leads us to believe that gender classification in anonymized resumes is a difficult task.

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

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

For example, there does not seem to exist any apparent syntactical differences between male and female resumes. Thus, keywords (*e.g. Women in CS*) become the most important tool in classifying gender. On the other hand, we had a significantly higher agreement for the resume quality task. This means that the annotators broadly agreed on what good and bad resumes look like. 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 2 and 3. We also calculated some group agreement metrics on the gender recognition task. The Fleiss's Kappa is 0.229 and the Krippendorff's alpha is 0.219, both corresponding to a slight agreement. This is consistent with the agreement rates in Table 2.

3.4 Accuracy vs Ground Truth

The annotators' accuracy compared to the ground truth on the gender task was quite low across the board. This makes us believe the gender vs quality depends more on the annotator's perception of the gender rather than the actual gender. This also is in line with our inter-annotator agreement findings. Even when the annotators had high agreement over resume quality, 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 perceived male resumes vs the quality of perceived female resumes as shown in 3.2. The gender classification accuracy metrics are available in Table 4.

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 4: Annotator metrics against the ground truth on the gender task.

3.5 Quality vs Ground Truth

In this subsection, we explore if there is a correlation between the quality judgement of the annotators vs the actual genders of the resumes. We wanted to see if there was an actual difference in qualities between male and female resumes.

- Male 1 classified female resumes, bad roughly 50% of the time and classified male resumes bad roughly 43% of the time.
- Male 2 classified female resumes, bad roughly 63% of the time and classified male resumes bad roughly 64% of the time.
- Female 1 classified female resumes, bad roughly 41% of the time and classified male resumes bad roughly 43% of the time.
- Female 2 classified female resumes, bad roughly 27% of the time and classified male(perceived) resumes bad roughly 34% of the time.

As can be observed, the difference in percentages for each annotator is less than 7%, which means the actual gender did not really affect the quality metric that much. Furthermore, the differences are much less significant than the ones encountered in the perceived gender vs quality analysis. This leads us to hypothesize that the perception affects the bias much more than the actual gender and that the actual gender has no correlation with the quality of resumes.

4 Supervised Learning Models

To quantify whether or not a machine is able to outperform humans on the gender recognition task, we ran multiple supervised classifiers, each based on different feature extraction methods.

We used 3 feature extraction models: N-gram, GloVe (Pennington et al., 2014) and BERT (Devlin et al., 2019). We used bi-gram to model a keyword-based classification approach. We consider this our baseline as most of the annotators reported that they also used a keyword-based approach while classifying so we expected this to be most similar to the annotators. The second feature extraction method we used was GloVe. Our motivation behind using GloVe was that it provides a continuous feature representation thus it would group similar resumes together in the feature space. Lastly, we used BERT as we hoped BERT would be able to additionally identify any syntactic or hidden patterns and see if those could be used to distinguish between male and female resumes. We then combined each of the embedding method with 3 different classifiers to highlight the strengths of each feature representation. For example, we expect continuous embeddings (GloVe and BERT) to work well with continuous optimizers in Support Vector Machine and Neural Network while a discrete representation such as Bag of Words would work better with Logistic Regression.

4.1 N-gram Embeddings

To build the N-gram embeddings, we utilized sklearn built-in features to generate the feature vectors. scikit-learn has built-in tools for gathering N-gram 1-hot vectors so the process was very straightforward to build these.

4.2 GloVe Embeddings

The GloVe embeddings were fairly straightforward to build, but were not as simple as the N-gram. To build these, we first had to download the GloVe embeddings online. In order to maintain reasonably fast computation times, we decided to use the 6b tokens 300 dimension vector file. Once the word embeddings were read in, we went through and cleaned each resume, and assigned every word in it to it's corresponding glove embedding (or a 300-dimensional vector of zeroes if the word did not exist in the GloVe embeddings file). From there, we took the mean across all words in the resume so that each resume produced a single 300-dimensional vector for processing.

4.3 BERT Embeddings

The BERT embeddings were extracted using the Sentence BERT (Reimers and Gurevych, 2019) model. We used the ”nli-bert-base-max-pooling” model to be exact. We hoped that the NLI fine tuning would make it more sensitive to the syntax structures as it was tuned for syntax similarity tasks. To generate the embeddings, we fed the whole document into the embedder which gave us a 1024 dimensional feature vector to be used in the classification model. For the coding implementation, we used the Sentence transformer library which is available in hugging face. The model had a max input document size length of 512 tokens. This was practically not an issue since nearly all of the resumes after anonymization had a document length less than 512.

We decided to test three supervised learning models for our experiment. Adding diversity to the models tested gave us a better look at what features were most important to the classifiers when determining gender.

4.4 Logistic Regression

Logistic regression (Cox, 1958) often is one of the first algorithms that comes to mind when performing binary data classification. It excels in pulling out the most key pieces of information from a chunk of data and is very fast to run. Because of this, we believed that it would make a good baseline analysis for resume analysis. We used scikit-learn's implementation to run logistic regression on the cleaned resumes.

4.5 Support Vector Machines (SVMs)

In an attempt to find patterns in more complex, non-linear settings, we also decided to run a support vector machine classifier (Cortes and Vapnik, 1995). Here too, we decided to use scikit-learn's implementation to run support vector classification on the cleaned resumes.

4.6 Multi-layer Perceptron Models (MLPs)

If female and male resumes really are very different, then we hoped that a complex nonlinear function approximator such as a multi-layer perceptron classifier would be able to recognize gender patterns in the data. Again, we decided to use scikit-learn's implementation to run multi-layer perceptron classification on the cleaned resumes.

4.7 Supervised Models' Result

The model outputs of each of the supervised learning algorithms can be seen in Figure 3. Note that

Model	Precision	Recall	F1	Accuracy	Kappa agreement with humans
N-gram Log Reg	0.671	0.725	0.697	0.646	0.0120
GloVe Log Reg	0.638	0.808	0.713	0.634	-0.0077
BERT Log Reg	0.591	0.789	0.676	0.575	-0.0100
N-gram SVM	0.560	0.992	0.716	0.559	0.0003
GloVe SVM	0.616	0.855	0.716	0.619	0.016
BERT SVM	0.584	0.877	0.701	0.581	-0.028
N-gram MLP	0.688	0.541	0.606	0.604	0.010
GloVe MLP	0.658	0.703	0.680	0.628	0.053
BERT MLP	0.580	0.774	0.664	0.556	0.0119

Figure 3: Supervised Model Results

Model	Average Gini impurity	Optimistic Accuracy
K-means	0.4843	0.571
Hierarchical clustering	0.4881	0.5406
Spectral clustering	0.4835	0.567
Gaussian Mixture Model	0.4843	0.573

Figure 4: Clustering Model Results

the SVM models typically have much higher recall than precision. This means that these models indiscriminately output male predictions for most resumes, leading to imprecise result. The accuracy of this model might mostly be due to the already prevalence of male resumes in our dataset. We can also see that the logistic regression model running on N-gram feature vectors performed the best overall for the supervised models, both in F_1 score and accuracy. We believe that the N-gram model performed the best because of certain identifying keywords found in the resumes that seem pretty decisive across genders. The performance of this model is slightly better than the average human accuracy of about 0.63. We believe this is due to several factors. First, the human annotators were instructed to spend a very short amount of time deciding the label for each resume (around 5 seconds per resume). This instruction is to simulate the real-world scenario where recruiters would also skim through the resumes. Due to the short duration of inspection, the human annotators might miss some gender indicators, leading to lower accuracy than a ML model that was fed the entire resume embedding. Other limitations of humans in memory, cognitive processing, and stamina might also contribute to worsened performance. Second, it is hard for an annotator with no technical background to develop a good heuristic online fast enough to yield good performance. For example, a reviewer with no background in Computer Science might not know beforehand that the *Grace Hopper Conference* is an all-female conference, and thus a very strong indicator of female. During the annotation process, the reviewer might encounter those *Grace Hopper* phrases frequently but without female ground-truth labels, it is still challenging for the reviewer to develop the association between the phrase and female, especially with limited memory, in a fast-paced review, amidst other words. Supervised models, on the other hand, might be able to learn those statistical correlations and co-occurrences better.

For the logistic regression model, We also include the most important words for each class prediction in Table 5. For binary classification, the feature importance is just the logistic 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 an 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

	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: Logistic Regression feature importance.

Women in Computer Science or **Women** in computing.

4.8 Humans and Supervised Machines agreement

We are also interested in the agreement rates between human annotators and supervised models to see if they utilize the same signal. First, we create a majority human gender prediction for each resume. We ignore the prediction of the F2 annotator as she was not technically grounded [1], and take the majority vote on the prediction of the remaining three annotators. We then compute the Cohen's Kappa agreement between human-majority prediction and each supervised model. The results are in Figure 3. The negative kappa values indicate disagreement. As seen by the fact that the values are nearly equal to 0 for all models, any agreement between the human annotators and the models was by chance.

5 Unsupervised Learning Models

We are also interested in the question: how much "genderness" is captured through clustering and other unsupervised techniques when gender is not an explicit label. Another motivation behind using unsupervised algorithms was the fact that in a real-world setting, we typically do not have any label, much less legally usable label for gender. Also, a resume search engine would have most likely not been trained to predict the gender of a resume. Thus, a general clustering algorithm would be much more similar, in mechanism and operation, to a search engine that is supposed to return the "good resumes" for example.

Here, we use two classes of algorithms: hard and soft clustering. In hard clustering models, the machine is asked to assign each resume to one of two clusters. A variety of common clustering algorithms including K-means (Hartigan and Wong, 1979), hierarchial (Zepeda-Mendoza and Resendis-Antonio, 2013) and spectral clustering (Ng et al., 2001), and Gaussian Mixture Model (Reynolds, 2009), were used. Note that all of those algorithms take the number of clusters in as a user-provided hyper-parameter. Thus, they allowed us to specify the number of clusters as 2 in order to attempt to discriminate between male and female resumes. There are other clustering algorithms such as the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) that do not require specifying the number of clusters. In the soft clustering case, a resume is decomposed as a combination of two latent topics. We used two popular topic modeling algorithms, Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), for this analysis.

5.1 Resume Feature Representation

In topic modeling literature, a collection of documents is often transformed to a matrix where each row is a document, and each column is the word count for the unique word in the lexicon in each document. Such a matrix is called a term-frequency matrix. A fine-grained improvement on that idea is the Term-frequency-inverse document frequency (TF-IDF) (Sammut and Webb, 2010). Using TF-IDF can avoid undesirable behavior where common words like "the" dominate the term-frequency matrix.

We mainly used TF-IDF to represent resumes in the topic modeling algorithms. For the clustering algorithms, we used continuous embedding GloVe as we found that it consistently outperformed BERT.

5.2 Resume Similarity

Without explicit gender labels, it might be very difficult to distinguish between male and female resumes. To investigate the difficulty, we used TF-IDF to embed each resume then calculated the cosine-similarity between any two resumes. We compute the average similarity between document

[1] The annotator works on some field other than Computer Science

classes in Table 6. 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. Indeed, there is not a lot of distinction between male and female resumes.

Figures 5 and 6 are the word clouds for female and male resumes respectively. Again, most words across these two classes are very similar and non-gender-specific. The gender indicators in Table 5 do not seem to show up in these word clouds at all.

Similarity	male	female
male	0.0525	0.04926
female	0.04926	0.0532

Table 6: TF-IDF Similarities between resume classes.

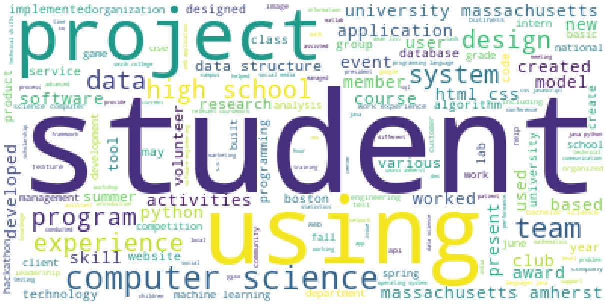


Figure 5: Word cloud for the female resumes.

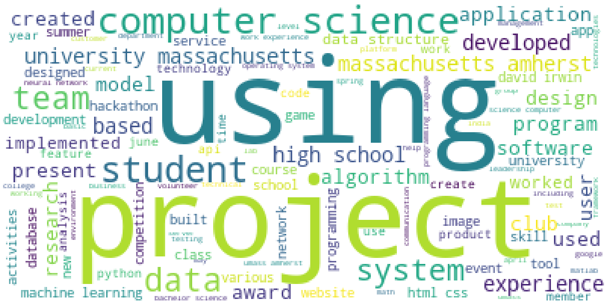


Figure 6: Word cloud for the male resumes.

5.3 Clustering Models' Result

We need to devise a mechanism to evaluate the goodness of clusters in terms of distinguishing gender. One quantitative measure is the Gini impurity. For a cluster, the Gini impurity measures how "diverse" the cluster is in terms of ground-truth gender labels. A cluster with a high mix of female and male resumes would have a higher Gini value (*i.e.* more "impure"). A cluster that consists solely of female resumes would have a Gini value of 0 and is deemed highly gender discriminative. Mathematically, the Gini value is the probability of picking two points of different classes in the same cluster. The goodness of a clustering algorithm is the average Gini impurities of its two clusters. The lower the Gini average, the better the model is at segregating male and female resumes.

Another measure we used is the "optimistic accuracy". The clustering algorithms can be used as a crude binary classifier. However, recall that in a supervised model, the machine is asked point-blank to give a prediction, male or female, to a resume. In an unsupervised cluster, we can either assign all resumes in cluster 1 to "male", or equally valid to "female". In calculating the "optimistic accuracy", we will assign the cluster in whatever way that maximizes accuracy. That is to say, if assigning cluster 1 to female turns out to give a 45% accuracy on ground-truth labels, we would back-pedal and assign to male to get a 55% accuracy.

The result for the clustering models is given in Table 4. As seen, the Gini values are consistent with the optimistic accuracy *i.e.* lower Gini value (more discriminative) corresponds to higher accuracy. The Gaussian Mixture Model performs the best although the performance does not vary significantly across different algorithms. It is noticeable that even the optimistic estimates for the accuracy of the clustering models are much lower than those from their supervised counterparts in Table 3.

5.4 Topic Modeling Algorithms' Result

LSA works by performing Singular value decomposition (SVD) on the tf-idf feature matrix. This is a linear reduction technique that projects both words and resumes to a common latent space of topics. [2] A popular measure used to select the best number of latent topics is the topic coherence score (Stevens et al., 2012). The score measures how semantically similar the words in each topic are. A plot of the coherence score versus the number of topics is given in Figure 7. As seen, the best number of topics is 5.

In LSA, the k topics are obtained by looking at the k right or left most principal eigenvectors. We are interested in how the two most salient topics are correlated with binary genders. Figure 8 displays

[2] This idea of projection to a common space is very powerful. For example, for collaborative filtering in recommendation systems, we can project both users and movies to a common space of genre.

the 10 most important words for each of the 4 most probable topics according to the LSA algorithm. As seen, the first top two topics do not have any gender indicators. LSA seemed to capture other types of clustering. For instance, topic 2-4 are from high-school students with words like “ap”, “school”, “club”, “national”. Topic 4 most likely consists of web enthusiasts with words such as “react”, “js”, “api”.

LDA is another topic modeling algorithm, from probabilistic graphical models. There we also do not see any traces of gender emerging from the latent topics. Figures 9 and 10 display the most relevant terms for topics 1 and 2 in LDA. The appearing terms are fairly generic to Computer Science without any obvious gender indication.

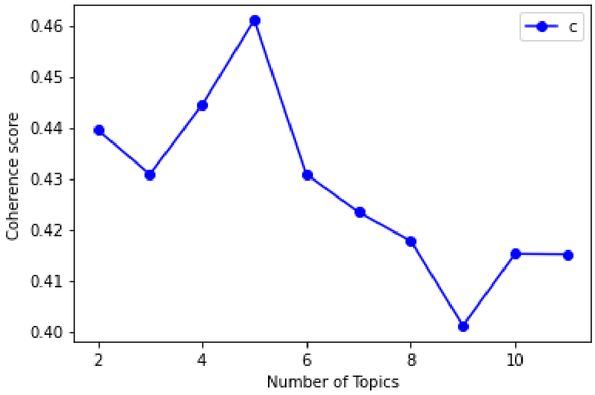
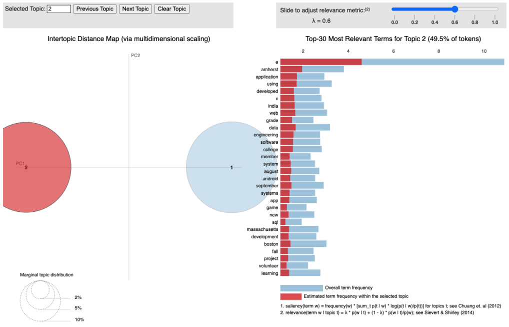


Figure 7: LSA Coherence Score vs Number of topics.


Selected Topic: 2

Slide to adjust relevance metric:(2) λ = 0.6

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9: LDA most significant words for cluster 1.

6 Conclusion and Future Work

In this paper, we first investigate how gender influences the perceived quality of a resume in anonymized screening. Based on the analysis on a very small number of annotators, we have found that

1. The annotator’s own gender is a factor in quality judgement.

2. The perceived gender of the applicant, not the actual gender, also influences the assigned quality score.

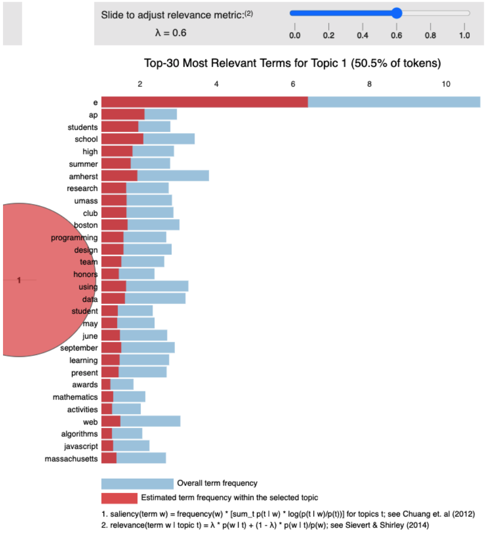
In the future, we would like to scale this experiment in Mechanical Turks so that more scientific and rigorous analysis can be done. For instance, the two-sample t-test for the difference in means can be run to determine if *there is a statistically significant difference between the quality of resumes from the two genders (using the actual gender)*(*). We did not run any statistical tests since the number of samples of 4 was too small. In principle, we expect the average female applicant to be as qualified as her male counterpart. However, since the quality score is a subjective measure assigned by human annotators, who themselves are biased, question (*) does not actually have clear-cut answer a priori.

We also found that supervised models that rely on keyword matching perform best on the gender recognition task. Those models provide us useful information about which (key)words are gender-revealing. There, the question on how to deal with the gender indicators remains. This is the question of whether to remove *President of Women in CS* phrase or not that we have discussed in the Introduction. In practice, although it is very unlikely that a supervised model trained specifically to look for gender would be deployed, those gender indicators can still bias the human reviewers. One caveat is that we found the performance of supervised models extremely variable depending on the train-test split. This is most likely because the number of the resumes we used is too small. In the future, we would like to replicate the result on much a larger scale.

For unsupervised models, we found that gender is not a salient feature in clustering or topic modeling. In fact, most tech resumes are very homogeneous with common words. When these models are not trained to look for gender-related keywords, it is hard for them to display visible gender bias.

	Most important words
topic 1	e, using, amherst, web, data, application, boston, india, developed, software
topic 2	ap, school, grade, india, application, club, high, physics, developed, web
topic 3	ap, grade, e, india, physics, acton, sat, smith, history, national
topic 4	ap, web, learning, acton, neural, react, js, application, research, api

Figure 8: LSA most important words. Note that there is some artifacts from the imperfect OCR preprocessor. For example, the word “e” is actually a bullet point in the resume pdf.


Slide to adjust relevance metric:(2)
λ = 0.6

1. saliency(term w) = frequency(w) * [sum_t p(t I w) * log(p(t I w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w I topic t) = λ * p(w I t) + (1 - λ) * p(w I t)/p(w); see Sievert & Shirley (2014)

Figure 10: LDA most significant words for cluster 2.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, page 1–14, New York, NY, USA. Association for Computing Machinery.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

Eva Derous and Ann Marie Ryan. 2019. When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal*, 29(2):113–130.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

J. A. Hartigan and M. A. Wong. 1979. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108.

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science faculty’s subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*.

Analiza Muñoz. 2019. The acceptability of government agencies on anonymized competency-based recruitment and selection process.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Douglas Reynolds. 2009. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea. Association for Computational Linguistics.

Marie Lisandra Zepeda-Mendoza and Osbaldo Resendis-Antonio. 2013. *Hierarchical Agglomerative Clustering*, pages 886–887. Springer New York, New York, NY.