Deep & Machine Learning Approaches to Analyzing Gender Representations in Journalism

Stephanie Campa
scampa@stanford.edu

Maggie Davis
mfdavis@stanford.edu

Daniela Gonzalez
daniela7@stanford.edu

Abstract

Given the importance of journalism in society, it is concerning that there are still differences in the way that people of different genders are described in the news. We are using Natural Language Processing techniques to classify the gender of subjects described in news headlines as a method of gauging gender bias. We aim to identify the key language differences in how men and women are represented, and how these biases may translate to NLP. Our dataset consists of approximately 1800 news headlines from eleven different publications, with each headline manually labeled as being about a "man", a "woman", or neither/both. Our final model uses a convolutional neural network (CNN) with GloVe word embeddings and word classes, max-pooling, a fully-connected layer with dropout, and the softmax function to obtain an accuracy of 86.7% on this classification task.

1 Introduction

1.1 Motivation

Journalism plays a key role in modern society and is widely regarded as a crucial component of a functioning democracy. However, despite increasing social pressure in support of equality of the sexes, men and women continue to be represented differently in the news. Not only are men more frequently the subject of journalistic articles than women, but the way men and women are talked about in the news differs. For instance, women are more likely to be portrayed in the context of their families, whereas men are more often discussed in the context of their accomplishments. This is only one example of a broad and complex issue. In light of that, our goal is to employ Natural Language Processing to delve more deeply into this phenomenon.

1.2 Project Description

Our project aims to address the NLP task of classifying gender in news articles – not of authors, but rather of subjects. Given an input headline, our model will aim to classify the gender of the subject of the piece. The output will therefore be one of two classes: 'Woman' or 'Man.' Articles that are not about people or are about both men and women will not appear in our dataset, as we are not as interested in exploring the use of language in those cases. These classes are described in greater detail in the Data section below.

Although attempting to solve this NLP task is not visibly useful in itself (as a human reader can quickly and easily identify the gender of an article subject based on simple context clues, such as pronouns), we hope to capitalize on the biases that exist in NLP to learn more about the gender-based biases that are most prevalent in journalism. Our project poses the questions: In what ways are men and women talked about and represented differently through language in journalism? What are the key differences? How/when do these biases translate to NLP, and how/when do they not?

A note on gender identity: We recognize that the media may not always accurately portray subjects' gender identities. However, we have chosen to label headlines based on how the media identifies the subjects' genders, not how the subjects themselves identify, because this project aims to explore

how gender is portrayed in the media. We also recognize that gender is a spectrum that cannot be accurately and rigorously depicted with the three labels we have selected. Nevertheless, these labels were chosen based on the most prominent gender identities currently represented in the news.

2 Related Work

Although various researchers have attempted to classify the gender of authors with NLP and deep learning, few have tried to classify the gender of the subject. Nevertheless, these two tasks are related in several ways, and because of that, we drew from related work in this space when determining how to approach our problem. Mukherjee and Liu [1] examined the task of classifying the gender of blog authors and found that combining multiple features – such as part-of-speech tags and factor analysis, among others – was effective. They used supervised learning algorithms – namely, Naive Bayes and SVM – in order to focus their experiments on the features as opposed to the algorithms themselves. Beyond this, much additional research has been conducted in the space of gender classification, but since most of these tasks are more often related to authorship, their approaches focus on differences in writing style between men and women.

In addition to drawing from papers pertaining to gender classification, we also explored deep learning approaches to text classification more generally. Since our project deals specifically with headline classification, where the text inputs are short in length, we pulled from Kim [2]'s work utilizing CNNs for sentence classification. In this paper, Kim found that simple CNNs with static vectors and some hyperparameter tuning could achieve excellent results compared to the state-of-the-art at the time of the paper's publishing, despite CNNs' original invention for the purpose of computer vision. Kim found that a CNN model with pre-trained vectors from word2vec was able to achieve higher accuracy than prior approaches, particularly on an opinion polarity detection subtask. Although this task is highly different from ours, Kim's success in the sentence classification space more broadly motivated us to implement a CNN and experiment with pre-trained word vector embeddings, not only word2vec but also GloVe, in addition to some of Mukherjee and Liu's aforementioned feature combinations [1].

3 Approach

3.1 Baseline

For our baseline, we implemented a Naive Bayes classifier. This classifier takes in word counts as input (after stop word removal) and uses those counts to calculate probabilities. It then uses the Naive Bayes assumption to determine if there is a higher probability that the article is about a man or a woman based on the words that its headline contains, outputting the class with the highest probability as our prediction. The code for our Naive Bayes classifier was adapted from the programming portion of a CS229 assignment completed by one of our team members last quarter.

3.2 Intermediate

As an intermediate step, we implemented classification with support vector machines (SVM) with bag of words features, since this approach has achieved good results, particularly with the linear kernel since text is often linearly separable [3]. We used bag of words features as input to the SVM, and performed stop word removal on the input. We leveraged the SVC class of the sklearn.SVM module for performing our support vector classification.

3.3 Advanced

For our most advanced approach, we applied a feed-forward convolutional neural network (CNN) to our task. We first implemented this using bag-of-words embeddings and later experimented with word2vec, GloVe, word classes, part of speech features (F-measure, which is described below), and various combinations of these. Again, we performed stop word removal on the input. Ultimately, we found that the most successful feature embedding combination in terms of accuracy was factor analysis word classes with GloVe. After the embedding layer, our model has a 2-D convolution layer followed by a max-pooling layer and, finally, a fully-connected layer with dropout, using softmax as the loss function.

The code for this neural network was adapted from a codebase, which employs TensorFlow, in a GitHub repository [4] based on Kim's work [2]. Our modifications to the code include rewriting the way in which data was imported and parsed, fine-tuning hyperparameters, and experimenting with different feature embeddings and representations. We also simplified the code somewhat, as certain parts of the original codebase were not necessary for our purposes.

Factor	**Words**
Conversation	know, people, think, person, tell, feel, friends, talk, new, talking, mean, ask, understand, feelings, care, thinking, friend, relationship, realize, question, anxwer, saying
AtHome	woke, home, sleep, today, eat, tired, wake, watch, watched, dinner, ate, bed, day, house, tv, early, boring, yesterday, watching, sit
Family	years, family, mother, children, father, kids, parents, old, year, child, son, married, sister, dad, brother, moved, age, young, months, three, wife, living, college, four, high, give, died, six, baby, boy, spend, christmas
Time	friday, saturday, weekend, week, sunday, night, monday, tuesday, thursday, Wednesday, morning, tomorrow, tonight, evening, days, afternoon, weeks, hours, july, busy, meeting, hour, month, june
Work	work, working, job, trying, right, met, figure, meet, start, better, starting, try, worked, idea
PastActions	said, asked, told, looked, walked, called, talked, wanted, kept, took, sat, gave, knew, felt, turned, stopped, saw, ran, tried, picked, left, ended
Games	game, games, team, win, play, played, playing, won, seasno, beat, final, two, hit, first, video, second, run, star, third, shot, table, round, ten, chance, club, big, straight
Internet	site, email, page, please, website, web, post, link, check, blog, mail, information, free, send, comments, comment, using, internet, online, name, service, list, computer, add, thanks, update, message
Location	street, place, town, road, city, walking, trip, headed, front, car, beer, apartment, bus, area, park, building, walk, small, places, ride, driving, looking, local, sitting, drive, bar, bad, standing, floor, weather, beach, view

Figure 1: Selected word classes from Argamon, et al.

For additional features, we extracted word classes from factor analysis and F-measure from parts of speech. We utilized results of factor analysis performed by Argamon, et al. [7] to classify relevant words into classes. Some of these classes are shown in Figure 1. We then created a feature vector that represented the number of words from each class that were contained in a given headline. Additionally, we calculated the F-measure of each headline. F-measure explores a text's relative contextuality, or implicitness, versus its formality, or explicitness. It is used in Mukherjee, et al. to distinguish between tones of male and female writing [1]. A lower F-measure score is a result of increased usage of pronouns, verbs, adverbs, and interjections. This implies contextuality, and is associated with women's writing. A higher F-measure score is marked by a greater use of nouns, adjectives, prepositions, and articles. This score implies formality, and is correlated with men's writing. We implemented F-measure as a feature when training our neural network to see if characteristics of men's and women's writing are also found in headlines where men and women are subjects. We calculated F-measure as follows:

$$F\text{-measure} = 0.5 \times ((\text{freq}(nouns) + \text{freq}(adjectives) + \text{freq}(prepositions) + \text{freq}(articles)) \\ -(\text{freq}(pronouns) + \text{freq}(verbs) + \text{freq}(adverbs) + \text{freq}(interjections)) + 2)$$

4 Experiments

4.1 Data

Our dataset is comprised of approximately 1800 article headlines from eleven news sources: The Guardian, BBC, Buzzfeed, CNN, Daily Mail, Entertainment Weekly, The Huffington Post, ABC

News, MTV News, New York Magazine, and TIME. These news sources were selected because they are popular and regularly report about people. We split our dataset into approximately 80% for training, 10% for validation, and 10% for testing. To ensure relevancy, these headlines were obtained from the News API.

We scraped, labeled, and parsed the data ourselves and – in order to complicate the task of gender classification – removed words that were obvious indicators of gender, including gender pronouns and names of key public figures, from the feature set. Since we were not in the possession of a dataset with gender annotations for news articles, we manually labeled the headlines to sort them into our two classes: 'Man' and 'Woman.' Again, if the subject of the headline was neither a man nor a woman, if the subject was a group of people that included both men and women, or if the subject was not a person at all, we omitted the headline from our dataset. Here are some examples:

- "Trump Wants to Start a Fourth of July Parade Tradition" - Man
- "Women Sue Yale Over a Fraternity Culture They Say Enables Harassment" - Woman
- "Senate Passes a Sweeping Land Conservation Bill" - Omitted from dataset

4.2 Evaluation Method

We evaluated the efficacy of our models primarily based on their test accuracy. We also conducted qualitative evaluation and error analysis by looking at examples of misclassified headlines. The metrics are available in our Results section, and our qualitative evaluation can be found in the Analysis section.

4.2.1 Experimental Details: Naive Bayes

For our Naive Bayes baseline, we implemented a Naive Bayes classifier as described in our Approach section. Specifically, we used word counts in headlines to determine the probabilities that certain words appear in certain classes. The model's input word counts are represented in the form of a vector for each headline, and the output is the class with the highest predicted probability for the respective input headline.

For our experiments surrounding Naive Bayes, we first implemented the classifier without removing any gender indicator words and then incorporated stop word removal to filter these out. After that, we computed the top indicative words for Naive Bayes to identify other potential stop words to exclude from the input data, such as names (e.g. 'Donald,' 'Theresa').

4.2.2 Experimental Details: Support Vector Machines

For our intermediate model, we implemented an SVM classifier as described in Approach. We used bag of words features, again with stop word removal, using vector representations of the word counts in each headline to create the matrices for training and test data. We tested with C=1.0 and $\gamma = \frac{1}{n\text{features}}$, and ran with both a linear kernel and Gaussian kernel.

4.2.3 Experimental Details: Convolutional Neural Networks

For our most advanced model, we utilized a feed-forward convolutional neural network (CNN), as described in our Approach. We experimented with many different changes to the feature embeddings. We loaded and utilized pre-trained word embeddings, adapting the loading code from a branch of the CNN GitHub repository [4]. We tried both GloVe embeddings developed by Pennington, et al. [5] and Word2Vec embeddings from Google News using the approach described by Mikolov, et al. [6]. We also experimented with word classes for factor analysis and F-measure for parts of speech. We ran our experiments using the settings suggested by Britz in the GitHub codebase: batch size of 64 and 200 epochs [4].

The most effective version of our model begins with an embedding layer from pre-trained 100-dimensional GloVe embeddings [5], which is followed by a 2-D convolution layer with 128 filters of sizes 3, 4, and 5 (that is, looking at 3, 4, or 5 words at a time), a max-pooling layer, and, finally, a fully-connected layer with dropout probability 0.5. The model uses softmax as the loss function. We experimented with several different hyperparameter settings, and the outcomes of these experiments are available in Results.

4.3 Results

Our training and testing datasets are randomly sampled, so the accuracy varies slightly across trials. For this reason, the accuracies for Naive Bayes and SVM are averages across 10 trials. Since our neural network takes longer to run, we did not compute averages for our more advanced method, but we found that the accuracies were relatively consistent across runs. The final accuracies for our models are as follows:

Model	**Accuracy**
Naive Bayes	73.9
SVM (Gaussian Kernel)	66.2
SVM (Linear Kernel)	72.5
CNN w/BOW	82.2
CNN w/word2vec	83.3
CNN w/GloVe	85.6
CNN w/Factor Analysis + BOW	75.0
CNN w/Factor Analysis, BOW, F-measure	76.1
CNN w/Factor Analysis + word2vec	82.8
CNN w/Factor Analysis, GloVe, F-measure	84.4
CNN w/Factor Analysis + GloVe	**86.7**

Figure 2: Model accuracy comparison

Unsurprisingly, our CNN models achieved higher accuracy than our Naive Bayes baseline and intermediate SVM. We anticipated that CNNs would perform well on our task because they have hierarchical architectures and are adept at keyphrase recognition, as opposed to other neural network architectures that are commonly used for NLP, such as RNNs, which are sequential [8]. CNNs are also good at identifying spatial patterns, which is useful for things like clustering words that may be similar to each other together. Both of these qualities are appropriate for our gender classification task.

Although Naive Bayes did not outperform our CNN models, it did perform relatively well for a baseline. This was somewhat surprising, but mostly expected, since Naive Bayes can easily latch onto words – whether they are indicative of bias or simply specific current events – and use those as clues for classification. Moreover, the Naive Bayes assumption takes into account the probability that an article will be about a man or a woman. Since articles about men are more prevalent (despite our efforts to balance our dataset), this makes it possible for Naive Bayes to perform well.

Our model performed better with GloVe embeddings than word2vec embeddings. This is consistent with the findings of Pennington, et al., who found that GloVe outperformed word2vec, even when controlling for vector length, context window size, corpus, and vocabulary size [5]. We hypothesize that context may be more immediately useful for our classification task than word meaning, since news headlines have a relatively standardized vocabulary and tone, and less ambiguity in word choice and meaning.

We observed that building a vector of word class scores (factor analysis) for each headline improved the accuracy of our model when paired with GloVe and word2vec. This likely occurred because certain classes of words as a whole were likely to be correlated with men or women, even though individual words in each class may appear less frequently in headlines. For instance, we expect words in the "business" class to be correlated with male subjects, and words in the "family" class to be correlated with female subjects. We also calculated the F-measure of each headline, which is a measure of a text's contextuality versus formality based on the frequency of various parts of speech in the headline. The F-measure was previously used in other papers to assess the tone of a piece of writing and to aid in classifying whether it was written by a man or a woman. The task of classifying the subject of a headline is a bit different from this, and we found that extracting F-measure did not significantly improve our accuracy. We can infer that this may not have made a big difference in our accuracy because relative to longer form texts, headlines do not have a lot of variation in the parts of speech they use because they are written to be short and eye-catching.

Embedding Dimension*	Num Filters	Dropout Probability	L2 Reg Lambda	BOW Accuracy	FA + GloVe Accuracy
128	128	0.5	0.0	71.6	**86.7**
64	128	0.5	0.0	76.7	86.7
256	128	0.5	0.0	78.9	86.7
128	64	0.5	0.0	77.2	85.0
128	256	0.5	0.0	75.0	86.1
128	128	0.2	0.0	78.3	85.0
128	128	0.3	0.0	81.1	85.0
128	128	0.4	0.0	74.4	86.1
128	128	0.5	1.0	**82.2**	86.1
128	128	0.5	5.0	80.0	85.6

Figure 3: Hyperparameter tuning for CNN models. *Note that the embedding dimenson does not effect the FA + GloVe column, since the GloVe word embeddings are pre-trained to a set dimension.

In order to maximize our neural network accuracies, we conducted hyperparameter tuning for our most basic deep learning approach (BOW), starting with the default values from the codebase (available in row of one Figure 3), and then for the advanced approach that was most successful with those default values, factor analysis with GloVe word embeddings. Ultimately, we found that the default parameters of 128 filters, dropout probability of 0.5, and L_2 regularization lambda value of 0.0 achieved the highest accuracy for our CNN with factor analysis combined with GloVe embeddings. For our bag of words approach, the highest accuracy was achieved by an embedding dimension of 128, 128 filters, dropout probability of 0.2, and L_2 regularization lambda value of 1.0. The complete results of our hyperparameter tuning experiments are available in Figure 3 above.

5 Analysis

5.1 Error Analysis

Headline	Prediction	Label
"Maria Ressa: Philippine Journalist Critical of Duerte, Arrested for Libel"	Man	Woman
"Lori Loughlin, Felicity Huffman Indicted in Sweeping College Bribery Scheme. Here Are the Biggest Allegations and What to Know"	Man	Woman
"Child Genius: Nishi Stuns Viewers in Mathematics Round"	Man	Woman
"Christian Dior Launches Latest 'Sisterhood' Slogans"	Woman	Man
"Ex-Boyfriend Arrested in Death of Woman Whose Body was Found in Suitcase"	Woman	Man

Figure 4: Misclassified headlines

We can gain insights about what causes our model to classify headlines by analyzing examples that were incorrectly classified, such as those in Figure 4 above. For instance, four headlines naming the actresses who were involved in the recent college admissions bribery scandal, Lori Loughlin and Felicity Huffman, were incorrectly classified as being about men. This ocurred despite the fact that Lori Loughlin and Felicity Huffman's names were not listed as stop words, and thus were not removed from the headlines. We predict the model behaves this way because words related to crime, such as "scam," "charged," and "bribery" are more frequently linked to men than to women. This may have caused the neural network to classify the headlines as being about men even though they explicitly listed the actresses' names. A similar trend is likely to occur in the first misclassified headline in the table above, because the words "libel," and "Duerte" are more commonly associated with male subjects, even though the subject of this headline is female. The second headline in the table was classified as being about a man, when the child in question is female. This is likely evidence of bias regarding the words "genius" and "mathematics" being correlated with male subjects.

Although it occurs far less frequently, several headlines describing male subjects were incorrectly classified as female. In the third headline, it is clear that even though Christian Dior is male, the model is being influenced by the word "sisterhood." Some causes for error can be traced to ambiguity about how to define the "subject" of a headline that mentions people of multiple genders. For instance, the fourth headline was labeled as having a male subject, but classified as female by our neural net. While we labeled the true subject of the article as being the ex-boyfriend, headlines describing women as victims of crime are common. Thus, the words "death" and "body" likely influenced the model to classify this headline as having a female subject.

5.2 Publication Comparison

Since our results were based on averaging across headlines from eleven different publications, we also chose to examine each publication individually to discern how they may have influenced our overall results. Although we were limited by the significantly smaller dataset when cutting down to one publication, this practice allowed us to compare the publications relative to one another. We found that BBC News, New York Magazine, and CNN had the highest accuracy rates, with greater than 85% accuracy for classifying headlines. The next highest accuracy rates were for Daily Mail and MTV News, which each obtained 80% accuracy. We were intrigued by this finding because it contradicted our expectations. We had thought that the more informal publications that may tend to lean more towards celebrity gossip (such as Daily Mail and MTV News) would have the highest accuracy. These publications have a stereotype of being more superficial and sexist, theoretically making the language they use to describe men and women more distinguishable. However, it appears that overall topic and context were more indicative for classification than subtle differences in tone. Some of the more serious, general publications (BBC News, New York Magazine, and CNN) tended to discuss men and women in very different contexts – for instance, far more frequently discussing men in relation to politics or law. These disparities in topics associated with men vs. women make the classification task clearer and contribute to a higher accuracy rate.

5.3 Most Indicative Words

Although it does not pertain as directly to our more advanced models, with Naive Bayes, we were able to calculate the most indicative words for each class on each run. Some of the words that came up as notable across our trial runs are available in the table below:

Class	**Indicative Words**
Man	‘guilty,’ ‘summit,’ ‘emergency,’ ‘congress,’ ‘report,’ ‘do,’ ‘wall’
Woman	‘dress,’ ‘israel,’ ‘mom,’ ‘look,’ ‘daughter,’ ‘ex,’ ‘paris,’ ‘breaks,’ ‘husband’

Figure 5: Most indicative words

Although these words are highly representative of much of the gender bias that currently exists in the media, it is also important to note that many of these words are related to current events – such as Donald Trump’s ‘wall’ or ‘Paris’ Fashion Week, an event that many female celebrities attended. In order to lessen the effect of this, we could elect to filter words that are indicative of current events or expand our dataset to contain news from the past. However, if we want to focus on the biases in current news, it is difficult to remove more stop words, especially since headlines are already very short in length. Still, in addition to demonstrating the way in which men are more likely to be written about with regard to government whereas women are more likely to be written about in terms of their relationships with other people, these words also showcase the larger effects that these biases have in society. For example, if the United States had a female president, the political bias that is present in the ‘Man’ category might be less evident. But to this day, men are more likely to be presidents or congresspeople, whereas women are more likely to stay at home and care for their families. And although journalism exacerbates this issue by focusing more on men in political spaces and women in family spaces, these stereotypes are perpetuated further such that they can have concrete effects on the lives people lead.

6 Conclusion

We set out to build a model that could successfully classify a headline as being written about a male or female subject, using a dataset of approximately 1800 headlines from eleven different news sources. The basic structure of our final model is a convolutional neural network (CNN) with max-pooling, a fully-connected layer with dropout, and the softmax function. This task was an intriguing opportunity to experiment with different Natural Language Processing techniques and compare various combinations of features as input to this model. Our most successful version of the model is a CNN with GloVe pre-trained word embeddings [5] and word class features obtained from factor analysis [7].

When thinking about potential future work to improve this model and perform more accurately on this task, we think that ambiguity in headlines that mention both a man and a woman would be an important issue to examine. Human labelers made the distinction by labeling according to the more agentic person who is more clearly the subject of the headline. It's possible that implementing NLP techniques that provide the ability to examine sentence structure, such as dependency parsing, could help the model determine the subject of the headline when two people are mentioned.

In terms of the implications of this project, the fact that we were able to obtain an accuracy as high as 86.7% is indicative of both the gender bias present in journalism and the discrepancies in gender representation present in society as a whole. Ideally, there wouldn't be enough distinguishing factors present in headlines such that the gender of the person being discussed can be determined even without knowing their name or the pronouns used. There is potential, however, for projects such as this one to not only reveal issues present in today's society, but to help solve these problems. Perhaps if the author of an article were notified by an AI software that a word they used is a significant indicator of gender and is frequently used in a biased way, they could have a chance to re-examine the language they've chosen before publishing.

Machine learning is often criticized for learning and perpetuating the biases present in human society. We aim to leverage this disadvantage in a meaningful way, intentionally shedding light on societal biases and concretely representing them in order to work towards eventually removing them. This project is one specific application of the much broader idea that machine learning can be used to learn about unconscious bias in order to inform people about how to consciously work against it.

7 Additional Information

7.1 Mentor

Our mentor is Xiaoxue Zang.

References

[1] Arjun Mukherjee and Bing Liu. Improving Gender Classification of Blog Authors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[2] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

[3] Thorsten Joachims, et al. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. 1999.

[4] Denny Britz. Convolutional Neural Network for Text Classification in Tensorflow, 2018, GitHub repository, https://github.com/dennybritz/cnn-text-classification-tf.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. 2014.

[6] Tomas Mikolov, et al. Distributed Representations of Words and Phrases and their Compositionality. *In Proceedings of NIPS*, 2013.

[7] Shlomo Argamon, et al. Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 2007.

[8] Wenpeng Yin, et al. Comparative Study of CNN and RNN for Natural Language Processing. 2017.