Insights from a Century of New York Times Articles: leveraging SpaCy to analyze women's representation in the news

Title: Letters, Photographer: Paul VanDerWerf, License: Attribution 2.0 Generic (CC BY 2.0), Link

With AI and tools such as chatGPT making headlines recently, many of us started wondering how computers can analyze, understand, and generate human language. This is where NLP (Natural Language Processing) comes in. Our team wanted to learn more about NLP and explore various NLP techniques and see if we can gain insight into excerpts of human text using techniques such as preprocessing, sentiment analysis, and topic analysis. Hence, for this project, we set out to work with a dataset of New York Times articles from 1920-2020 to explore the relationship between women and the news. We wanted to know how often and when women were mentioned in the news over time. How frequently were women mentioned compared to men? Were articles mentioning women more likely to have family-oriented topics? Were they polarized? How did this change over time?

To answer these questions, we extracted features such as common topics, people, and places discussed in the titles and excerpts of articles for each decade, and determined how often women are mentioned in the news over the decades. We also wanted to assess the common

sentiments associated with women in the news and examine whether the sentiment in articles involving women has shifted over time. Our analysis suggests that while progress has been made in terms of women's representation in the media, there is still a long way to go in terms of fully recognizing and representing the diversity of women's experiences and achievements.

Our Project

Our data uses a Kaggle dataset of New York Times excerpts from 1920-2020. Each row represents an individual article, including year, title, and excerpt. Our goals were to extract features such as common topics, people, and places discussed for each decade, and to determine how and how often women are mentioned in the news over the decades. We also aimed to find common sentiments associated with women in the news, and examine whether the sentiment in articles involving women has shifted, and analyze what topics are commonly associated with men versus women.

Different approaches to NLP

There are several approaches to do NLP which can be run on a variety of platforms or using a variety of python packages. Some examples of NLP approaches are rules-based methods, machine and deep learning, statistical methods, or a combination of methods referred to as hybrid methods. Rule-based methods use predefined rules and patterns to extract specific information, such as part-of-speech tagging and sentiment analysis. Machine learning uses algorithms to learn patterns and rules from labeled data to classify and extract information from new data. Named entity recognition and text classification are examples of machine learning. Deep learning uses artificial neural networks to learn patterns and representations of text data, such as sentiment analysis and language translation. Statistical methods analyze and extract information from text data, including frequency analysis and clustering. Finally, hybrid methods combine multiple methods which are selected depending on the goals of the project.

NLP with SpaCy

For our project, we are using SpaCy because it has a variety of built-in functions for text processing, including part-of-speech tagging, entity recognition, and removing common words. Different libraries and functions are available depending on your goals, which makes SpaCy a flexible yet powerful tool. SpaCy can also be used for sentiment analysis: in our case, we used Textblob, which reads in text blobs and scores the text in terms of polarity and subjectivity.

Dealing with Text Datasets

One of the biggest challenges with NLP is the inconsistency of human language. You have to ensure that you're feeding the computer fairly clean and consistent language so that it can pull the most relevant topics. For this project we also struggled with different file types and the size

of our file. Our dataset was very large and inconsistent, had file types that were not convenient, and many titles were short, lacked context, and often had missing excerpts. NLP also requires a lot of computing power, especially when using such a large dataset. To deal with the messiness and size of our dataset, we only used a small subset of our data, used preprocessing to make text simple and consistent, and combined title and excerpts and treated combined text as a single "text blob." To help with computing and file types, we converted larger subsets of data to CSVs, used batch processing, and ran some of the computation-heavy processes on google cloud.

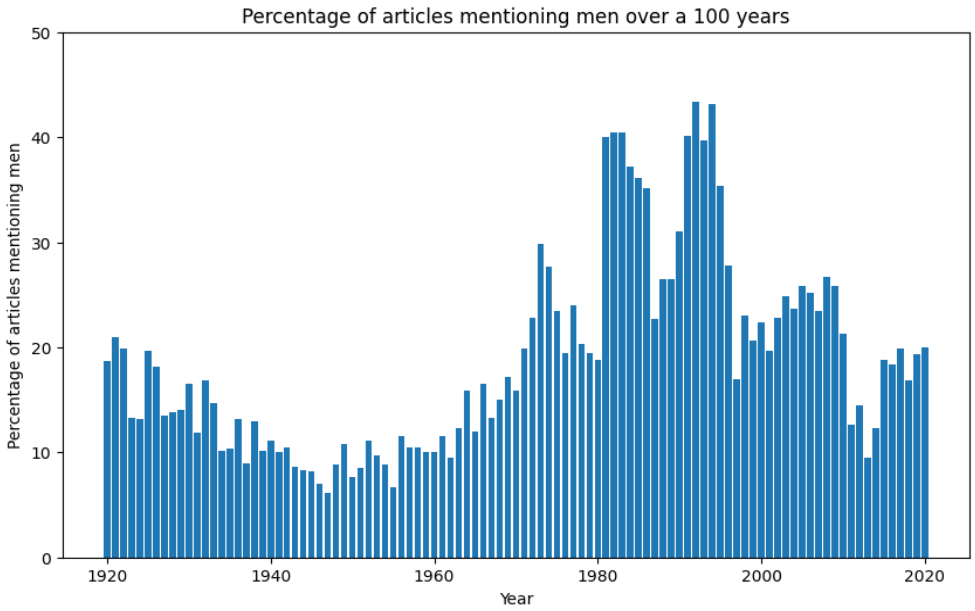
Preprocessing

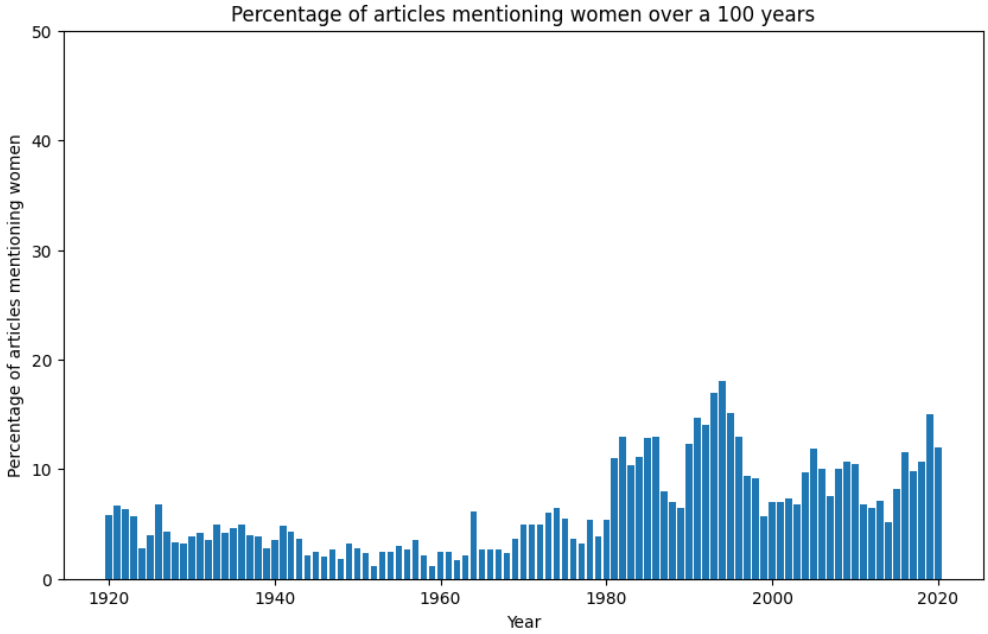
In order to effectively utilize NLP to analyze data, we first need to preprocess the dataset. To do this, we used SpaCy to perform name entity recognition, which identifies noun chunks as entities as labels as they fall under certain categories (organizations, companies, agencies, institutions, etc.). Preprocessing includes part-of-speech tagging that identifies the words in excerpts as nouns, prepositions, verbs, etc. In order to increase the power of NLP analysis, removing common words, such as "a," "is," "in," and "the", provides us with more important information from the excerpt. We then explored common entities from the excerpts and indicated their significance by identifying the important text entities. We then explored how topics shifted by decade by removing common words and focusing on nouns that were unique to the excerpts from each decade.

Next step: Topic Analysis
Are men mentioned more often than women?
The first step we took to answer this question was to look at both excerpts and titles to see if they contained pronouns such as she/her/hers or nouns such as woman/women/lady/dame. We then added a boolean column if the excerpt and/or titles showed gendered language. We repeated that step but this time for men, looking for masculine pronouns (he/his/him) and male-specific nouns (men/man/gentleman etc.), adding a boolean column if any of these words were found in the title and/or the excerpt.

The first question we asked ourselves was: is the number of articles mentioning men much higher than the number of articles mentioning women? How do these numbers change over time?

As shown below, although the frequency of women being mentioned increases over time, articles featuring men are more prevalent than those featuring women, regardless of the year.





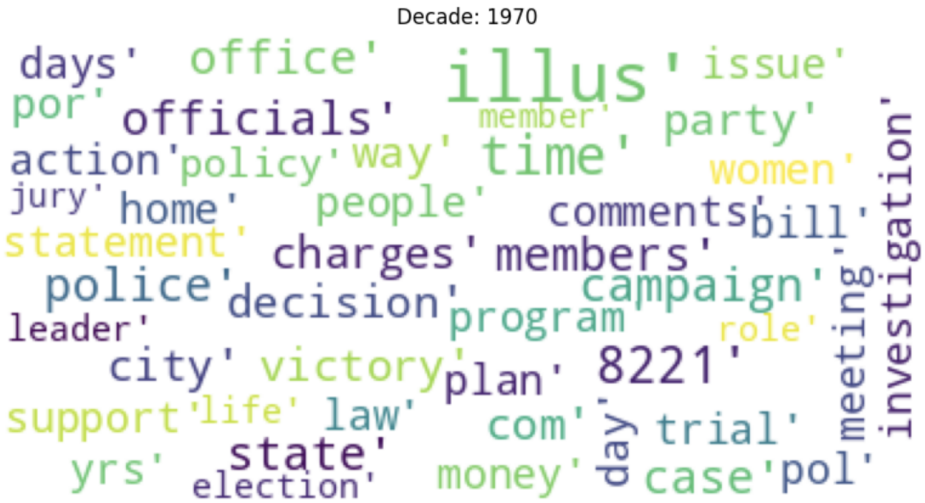
Interestingly enough, there is a clear uptick in mentions of women after 1980, suggesting that there was a cultural shift around that time. But if women get mentioned more often, does this mean that the culture and context around those mentions shifted as well? To answer this question, we set out to start a topic analysis.

Topic Analysis: In what context are women mentioned? Do these topics change over time?

Our next step involved extracting topics from these articles to better understand in what context were women mentioned. Were there recurring topics? And if so, do these topics change over time? What about articles mentioning men?

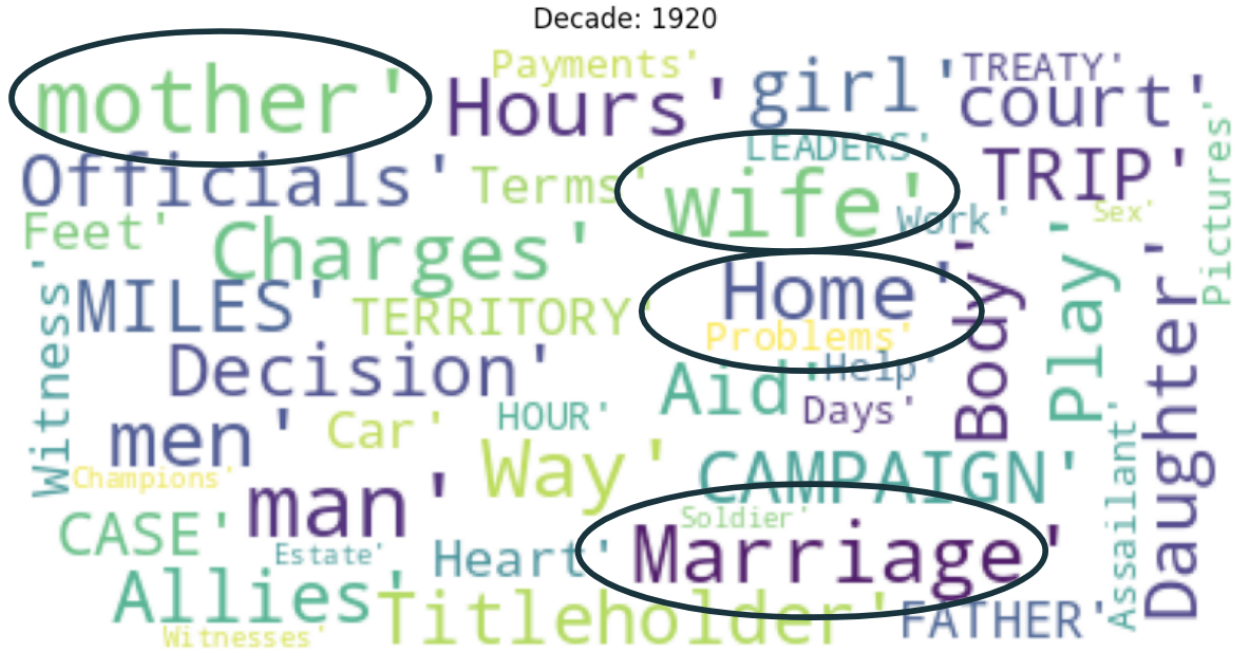
To easily visualize our data and answer these questions, we created word clouds for each decade showing what topics are mentioned often for women vs men. Men's topics were not very instructive: they were diverse and seemed aligned with historical events (‘campaign’ during presidential elections etc.).

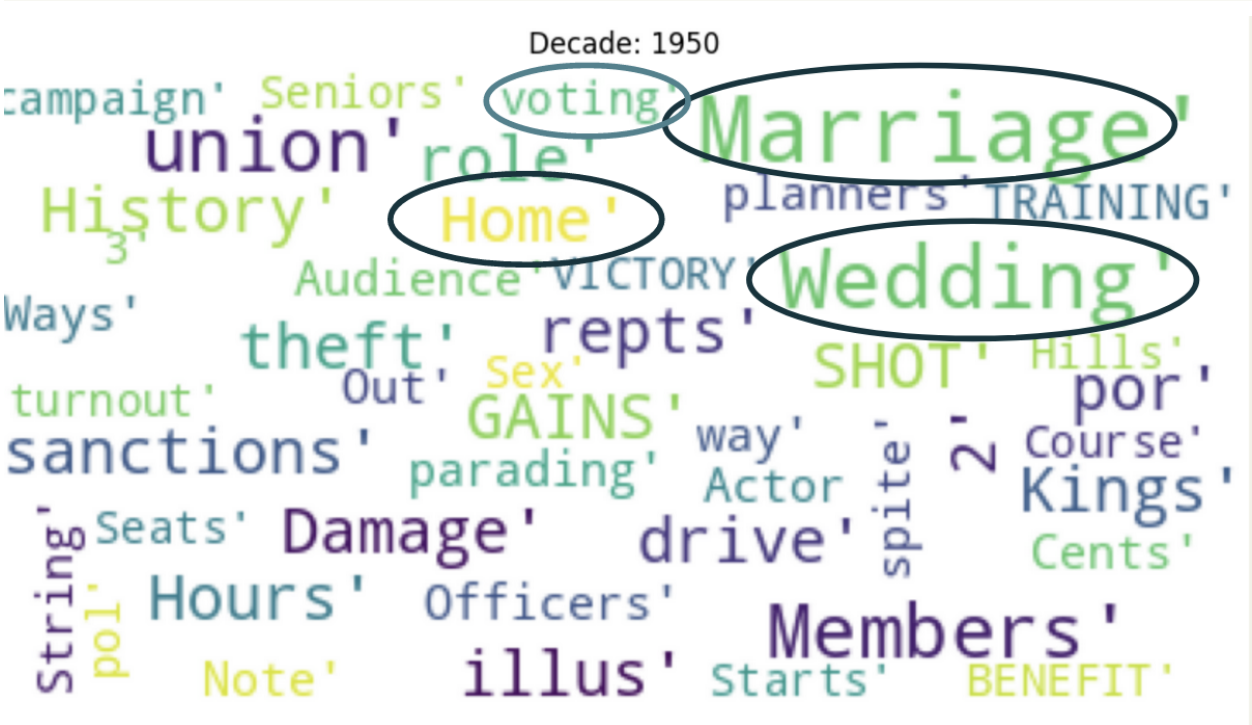
Examples of Men’s topics:



On the other hand, women's topics followed the same trend for most of the twentieth century: most topics were family-oriented. As we moved closer to the 2000s, we found that articles mentioning women became more diverse in their topics, and women were no longer solely portrayed as caregivers and homemakers. After 2000, women were mentioned more often in a variety of contexts, including in leadership roles, politics, and the workplace. Overall, our analysis suggests that while progress has been made in terms of women's representation in the media, there is still a long way to go in terms of fully recognizing and representing the diversity of women's experiences and achievements.

Women's topics in articles by decades (few examples):





Sentiment Analysis with SpaCy

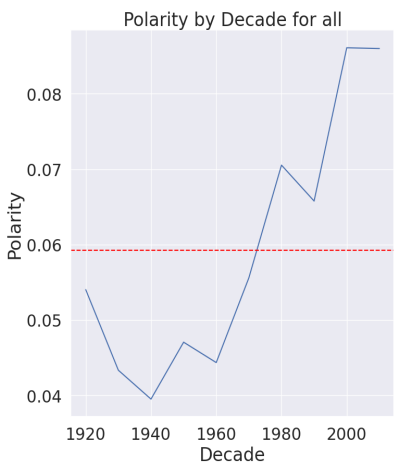
We used a “sentiment analysis pipeline” from spaCy named text-blob to do sentiment analysis on our text data. Text-blob can be used to evaluate both the polarity and subjectivity of the text. It uses a pre-trained classifier, which was trained on movie reviews, and a lexicon of words that have been assigned a polarity score and subjectivity score to calculate the overall sentiment of a text. Words are scored individually and combined to give a score to the text as a whole.

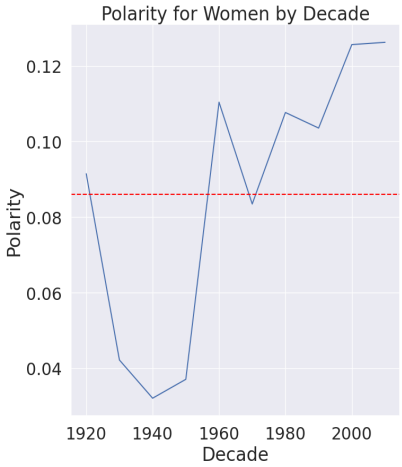
Here is an example of how text-blob assigns these values:

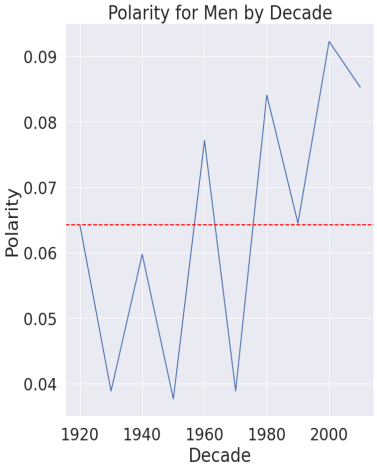
```
Row 5:
        Words: ['greater']
        Polarity: 0.5
        Subjectivity: 0.5

Row 5:
        Words: ['open']
        Polarity: 0.0
        Subjectivity: 0.5

Row 5:
        Words: ['classic']
        Polarity: 0.16666666666666666
        Subjectivity: 0.16666666666666666

Row 5:
        Words: ['third']
        Polarity: 0.0
        Subjectivity: 0.0

Row 5:
        Words: ['round']
        Polarity: -0.2
        Subjectivity: 0.4

Row 5:
        Words: ['back']
        Polarity: 0.0
        Subjectivity: 0.0

Row 5:
        Words: ['new']
        Polarity: 0.13636363636363635
        Subjectivity: 0.45454545454545453
```

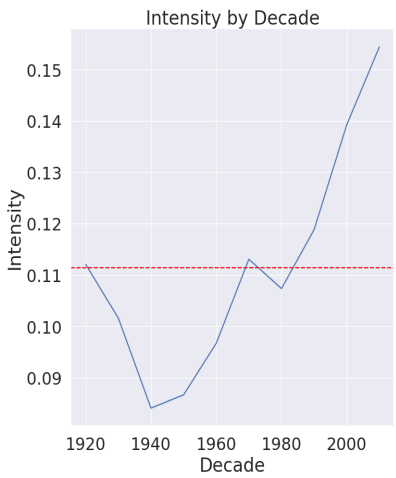
In addition to the polarity and subjectivity scores, we created an additional column called 'intensity,' which provided the absolute value of the polarity score. This allowed us to understand the degree of emotion expressed in the text, irrespective of whether it was positive or negative. We also created a new column in our dataset called 'sentiment,' which categorized the text as 'Very Positive,' 'Positive,' 'Neutral,' 'Negative,' or 'Very Negative,' based on the sentiment score generated by text-blob.

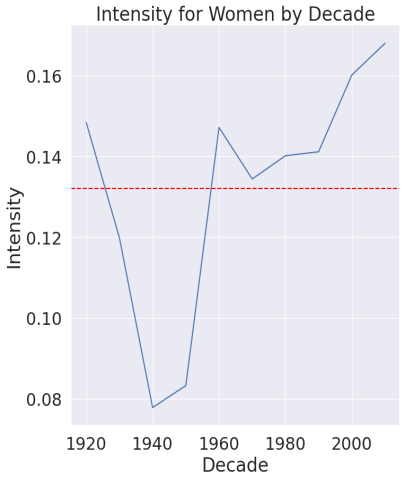
First, we graphed the average polarity, intensity, and subjectivity for each decade for all the articles, those mentioning women, and those mentioning men to see if there was a difference between groups.

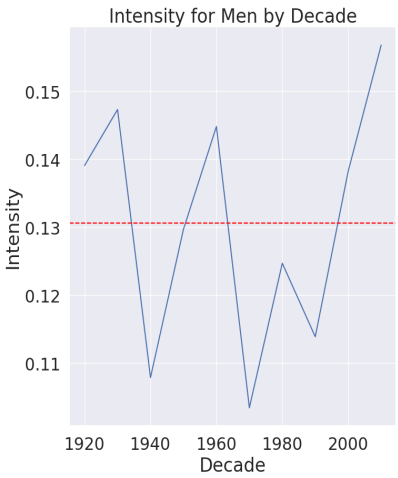


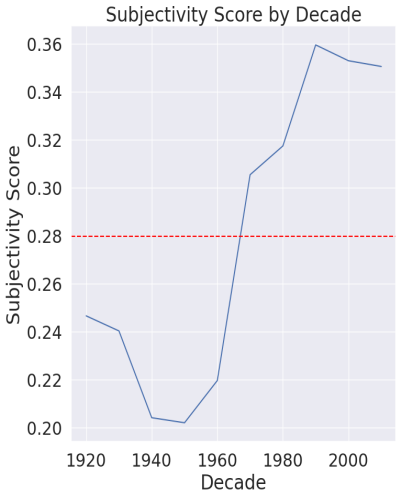


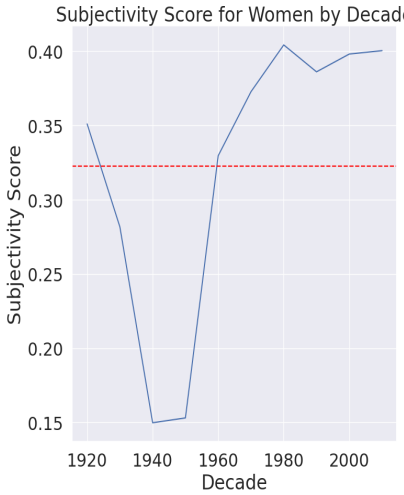


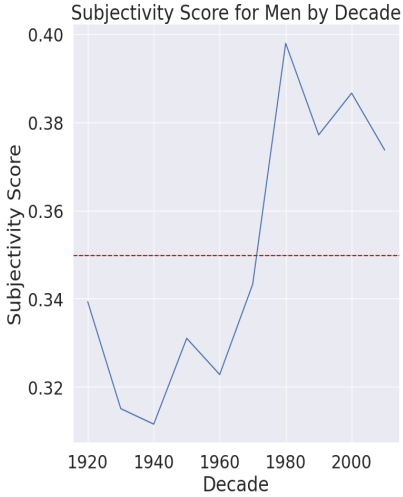








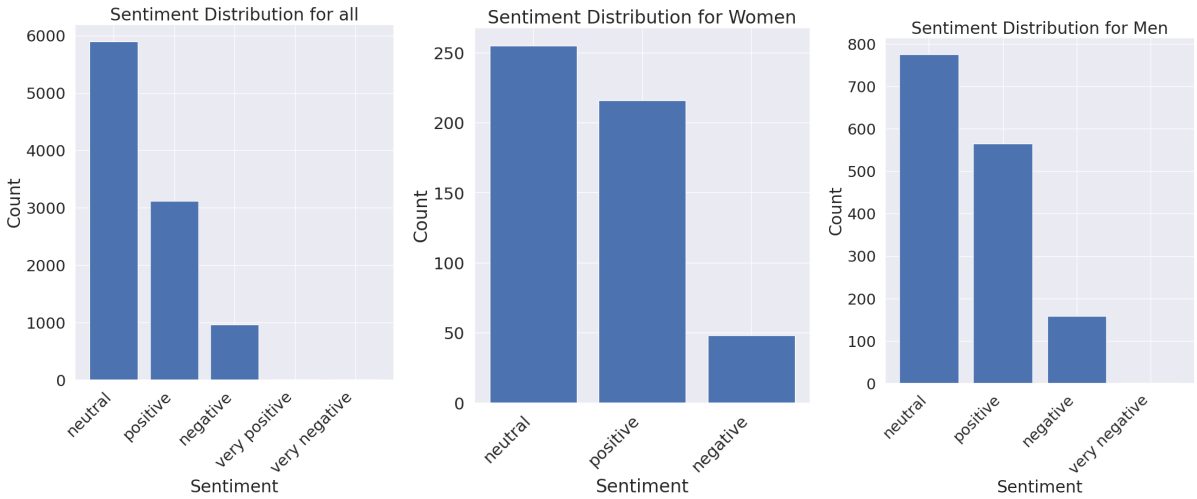




As we can see, all three scores increased throughout time and there was much more inconsistency in articles about men. Women on average had a higher polarity score and lower subjectivity score. Both men and women scored similarly in intensity, but both groups were higher than the average for all articles. Additionally, we can see that all three metrics had a dip

between the 40s-60s, especially for women. These trends, especially the large variation in scores for men, might be a result of the wide variety of topics associated with men.

Next, we compared the frequency for each sentiment for all the articles in the subset of our dataset with those mentioning men and women.



The lack of very positive and very negative articles is likely due to the type of text we're analyzing. Since text-blob was trained on movie reviews, it is better suited for analyzing reviews which are often more polarizing with very clear language and sentiments. The news, by nature, is a little less intense than reviews and scored more mildly across all groups.

Conclusion:

Overall, from this analysis we can see that women are mentioned far less frequently in the news than men and, historically, have been primarily associated with 'familial' topics like marriage, family, kids, and motherhood. Although we did observe an increase in the frequency of articles mentioning women after the 1980s, it still pales in comparison to the frequency of men in the news. Additionally, as time progresses we can see that women are associated with a wider variety of topics rather than just mothers and caretakers, whereas men are consistently associated with a huge variety of topics regardless of the decade. When looking at the sentiments associated with women and men over time, there is not a clear pattern or difference between the groups, though women had higher average polarity scores than men indicating that women are likely more often mentioned in a more positive context, which makes sense considering they are frequently associated with familial topics. Additionally, the variety of scores associated with men might be a result of the wide variety of topics associated with men.

Though we are thrilled to see that there has been some progress in women's representation in the media, there is still work to be done!