Twitter Sentiment Analysis Presentation Notes

# Introduction and Thanks

Thank you all for coming today. I’m Joanna Payoyo and for my 5590 project I did a Sentiment Analysis on tweets about Women in Technology. I’d like to thank Shoba and Yasaman for their help and guidance for this project. I couldn’t have done this project without you.

# Agenda (30 secs)

I’ll be going over why I chose this topic and my research questions. Then I’ll be going over the data analysis techniques I’ve used on this project. Then I’ll be doing a literature review. I’ll go over my code for extracting twitter data to ETL and analysis. I’ll explain my key findings and go into my conclusion. I will then go over what I learned and my struggles throughout this semester. Finally I will go over future work.

# Why I did this project and Research Questions (3 min = 375)

## Why

Since I decided to start a career in technology, I’ve become more aware with issues regarding a woman’s place in this field. There’s been some recent controversies such as the Google Memo, sexual harassment at Uber, and #ILookLikeAnEngineer that’s made me concerned about joining a field that aren’t welcome to women. Although my experience at MRU has, for the most part, been great, I’ve had some bad experiences which I attribute to me being a woman such as my group members not listening to me when I try to talk about the project but when I ask one of my male friends in the group to speak for me, everyone else immediately listens and responds. I’ve been talked down to multiple times by male classmates. And in general, being only one of two or three women in a class does not make me feel like this program is inclusive.

So I wanted to find out what the general public was saying about women in technology. I chose Twitter because people use it for two main things: 1) keep up with current events and 2) to tell people what they are thinking about and have discussions about the above topics.

## Research Questions

* In general, is the sentiment about women in technology on Twitter positive or negative?
* Are different keywords statistically significantly different from each other?
* What days of the week or month do people have more positive or negative sentiment regarding this topic?

# Background Study

Natural Language Processing (or NLP) is a multi-disciplinary field within data science in which we research how the human language can be understood by computers and then used for a variety of purposes. There are a number of ways to approach NLP. These include “machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems.”

The history of NLP begins after World War II due to tensions between the West and Russia. Machine translation of Russian to English was paramount to the US government as they needed to translate as many texts as fast as possible. In the 1950’s and early 1960’s, researchers found more ways for computers to understand language. In 1964, ELIZA, a precursor to Alexa and Siri, successfully impersonated a psychiatrist. This was the same year the Automatic Language Processing Advisory Committee (or ALPAC) was created by the US National Research Council to monitor NLP research. In 1966, ALPAC cut funding to machine translation research as it was found that machine translation was far more expensive than manual translation, at the time. It wasn’t until the 1980’s that NLP research started to pick up again due to increased computer power, and in the 90’s, the rapidly increasing use of the Internet, which in turn created more data with which to do research on. The focus of NLP switched from simply trying to understand the language to using computer speech recognition to do practical things. Today, NLP has gone so far as to give us digital assistants, such as Alexa, Siri, and Cortana that can automate our homes. IBM’s Watson is also a product of NLP. Famously, it competed and won against top Jeopardy contestants.

So we know that NLP can tell us individual bits of information but we are now moving towards a future in which NLP can help us understand mass amounts of data, such as determining the future price of oil by “reading” millions of articles regarding the oil industry.

Sentiment analysis is a branch of NLP which refers to the analysis of opinionated content. Sentiment analysis can be done in different ways such as feature-based sentiment analysis, which examines the different features a sentence has then determines if what was said about the feature was positive or negative. For example, in a product review regarding a vehicle, a feature found may be mileage and the sentiment found for a Hummer may be negative as opposed to a Prius which would be positive. Intent analysis determines the intention of a particular sentence, as in, is this sentence a question, suggestion, news, or an ad.

The most common method of sentiment analysis, as well as the one used for this project, is sentiment classification. This can be done in two ways: subjectivity and polarity. Subjectivity refers to the classification of a sentence as subjective or objective. Polarity determines how positive or negative a sentence is. For this project, I focused on the polarity of tweets.

This classification can be done through a rule-based approach that tokenizes words and assigns a value for each word. Classification can also be done through machine learning by using training the algorithm.

# Literature Review

During my research, I found reports that women are not only being abused online, they were also being actively shut out of the conversation through constant, and sometimes violent, harassment. A particularly famous example of misogyny in the sphere of women in technology on twitter, was when Felicia Day, a female video game celebrity, was called a “glorified booth babe” by a male video game journalist. While these do worry me a lot, these papers focused on known instances of sexist tweets so I wanted to understand the bigger picture regarding the general sentiment regarding women.

Different words have different connotations. Researchers from the University of Copenhagen determined that the ten most frequently occurring terms in sexist tweets included “women”, “girls”, and “female.” All three terms had different distributions within a collection of pre-determined sexist tweets: women had a distribution of 0.83%, female had 0.34% and girl had 0.48%. For this project, I wanted to see if using these different keywords would also yield different results in a sentiment analysis of a more general data set.

Finally, I was interested in seeing if general twitter sentiment held true for the subject of women in technology. The University of Vermont has been doing extensive research regarding happiness on twitter over time. Some of the trends they’ve found are that Saturday’s are typically the “happiest” day of the week on Twitter, with a steady decline to it lowest point on Tuesday. In this project, I wanted to see if general trends held true for women in technology tweets.

# Methodology/Process/Code (5 min)

* Get tweets
  + The first step was to get the tweets from Twitter. Twitter has a public API in which you can get tweets up to a week before today. Another important thing to note is that “the standard search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results.”
  + There are a many things you can get from a Twitter post beyond just the tweet and the user. This includes the date/time it was posted, where it was posted, the language it’s written in, how many retweets, favorites, if it’s a reply to something, the background color of the user.
  + After looking at the different types of data I could extract, I decided to extract the date/time, full twitter text, number of retweets, number of favorites, and the location information.
  + The date/time could indicate chronological trends, the retweets and favorites can tell us about the popularity of the tweets and which tweets people like to engage with, and the location could indicate geographic trends.
  + After looking at the location data, it was apparent that not very many tweets contain location data so I took it out of the get code.
  + Overall, I collected 27,872 tweets.
* Cleaning the Data (part 1)
  + The tweets contained special characters so I cleaned them up by replacing them with the appropriate symbols.
* Sentiment Analysis
  + There are many different Python libraries for natural language processing and sentiment analysis. I chose TextBlob as it is one of, if not the easiest and most intuitive natural language processing libraries to use.
  + So the next step in my process was to run the sentiment analysis through my data. I used both the sentiment polarity function, which is in the range of -1 (negative) and 1 (positive), and the subjectivity function, which tells me how subjective or objective a sentence is.
* Removing Outliers
  + After some data exploration, I removed some outliers to better understand what the tweets were saying.
  + I found the outliers by first plotting the sentiment polarity by the favorites as well as the retweets. Visualizing this allowed me to see exceptionally popular tweets.
* Data Exploration
  + Wordcloud
    - I generated a Wordcloud to see if there was anything interesting I could find. A Wordcloud is a visualized collection of the most frequently used words within a dataset. The larger the word, the higher its frequency. At its first iteration, the largest words were the keywords. After removing those, there were some words that said something more about the content of the tweets, however I found that it was not useful for me to understand what people are saying about women in technology on Twitter. Perhaps with more data cleaning and removing “neutral” words, a Wordcloud could be useful to look for reoccurring topics.
  + Descriptive statistics
    - After data cleaning, I have a total of 27,868 tweets to work with.
    - 15,516 or 56% of the tweets are positive (sentiment polarity over 0)
    - 6,912 or 25% of the tweets were neutral (sentiment polarity is 0)
    - 5,440 or 19% of the tweets were negative (sentiment polarity is under 0)
* Keywords
  + Are different keywords statistically significantly different from each other?
  + I used an Analysis of Variance or ANOVA test to see if the keywords Women, Female, and Girl made a difference in sentiment.
* Polarity and Subjectivity
  + In order to find the sentiment polarity and subjectivity of the tweets, I used TextBlob’s sentiment analyzer that uses the Pattern library’s sentiment analysis implementation. Later on, I also used TextBlob’s Naive Bayes implementation to find the sentiment polarity to compare the two.

# Findings: Polarity Range in data, positive vs negative (2 min = 250)

The sentiment of the tweets are largely positive. The average sentiment was 0.114. Additionally, the range of polarity was most concentrated between 0 and 0.25. While this seems to be a good thing, this range also seems to be quite narrow. This tells me that perhaps people don’t feel very strongly about this subject. I would also like to note that most of the outliers in the favourites and retweets lie within this range. This implies that it may not be the sentiment of the tweet that drives the popularity of a tweet, but rather the content of the tweet itself.

The average subjectivity was 0.396. The range of subjectivity was most concentrated between 0.186 and 0.573. Again, this is an indicator that people don’t feel very strongly about this subject.

Upon closer examination, the “judgement” of the classifier comes into question. The most retweeted tweet was “Amazon sells facial recognition software to ICE who keep murdering trans women in their custody.” This was rated as 0 sentiment and 0 subjectivity. To examine this, I split each word in the sentence and ran each of them through the classifier. I found that all words had 0 sentiment and 0 subjectivity, including “murdering.”

Using a different classifier with the Naive Bayes classifier, I received results that made more sense. The Naïve Bayes method outputs the polarity of a sentence in both positive and negative percentages. For the aforementioned Amazon post, it was found to be 0.19 positive, and 0.81 negative. The caveat to using this method is that it is much, much slower.

# Findings: Keywords (Statistical Difference)

In order to figure out if using different keywords made a difference on sentiment, I used an ANOVA test. An ANOVA test should tell me if the distributions of the different keywords are equal.

After splitting the data with the keywords Women, Female, and Girl, I ran it through an ANOVA test. According to the test, the three keywords are statistically different.

Women has the highest mean at 0.145, followed by girl at 0.102 and finally female at 0.086.

However, looking at this composite graph, there is only a small 0.02 to 0.04 difference between the means of the three keywords. So although they are statistically different, I would have do more analysis in order to determine what a 0.02 difference means.

# Findings: Days of the week, Days of the month (2 min = 250)

* Lastly, I wanted to determine if the day of the week or day of the month had any effect on how women in technology were talked about on Twitter.
* By looking at this graph and checking with a calendar, the days with the largest spikes for negative sentiment were Sundays and Wednesdays.
* According to a study that measured happiness on Twitter, most people tweet negatively on Monday so these findings of Wednesdays and the weekend being particularly negative for women in technology may be a trend specific to this topic.
* A cursory look through the tweets during these days doesn’t really help describe why these days are particularly bad. Perhaps a more detailed topic analysis would help find the reason for the negativity.

# Conclusion/So What (125)

* According to my sentiment analysis, the sentiment is largely positive towards women in technology on Twitter. By looking at different keywords, we see that wording does in fact make a difference in how people feel about this issue. We can also see that the days of the week also influence how people discuss this subject.
* However, it is also apparent that I am just at the tip of the iceberg in truly understanding how people feel about women in technology on twitter. My findings have created more questions than answers.
* For example, what would be the difference in my results if I used different classifiers? Would the time of day tell me something more about the way women in tech are discussed?
* Although I did find answers to my initial questions, in order to paint a full picture of what people are saying about women in technology, I must continue exploring the data and use a variety of methods.
* NLP using social media data gives us the opportunity to better understand each other. I hope to continue to harness this tool to understand issues such as gender discrimination in hopes of finding a solution to social issues.

# Limitations, Struggles, and Future Work

* ETL never ends. About 2/3 of my project was spent on ETL and there are still a lot I could do to continue cleaning it up, such as deleting all “neutral” words such as prepositions (after, in, to, on, with), conjunctions (and, because, but, for), and determiners (a, an, the, every, this, those).
* Analysis never ends. There is always something to be found. Some of the things I would do include:
  + Using a different library, like NLTK or SpaCy and compare the results.
  + Train my own classifier.
  + Topic analysis – tell me more about the qualitative contents of the tweets
    - Contextual Semantic Search
  + How does this compare to how women are spoken about in other media?
  + Geographic study
* Keep collecting data – the larger the data set and the longer the time period it is taken in, the more accurately representative your findings will be.