Linnea Frangén
Computational Literacy
Final project

1. Research Question and Dataset

In this project, I experiment with different computational methods to analyse the linguistic features of fake news. My research question is how language complexity differs between fact-checked fake and real news. Previous research has connected fake news with shorter words, higher lexical redundancy and thus lower lexical diversity as well as shorter article length in comparison to real news (Horne and Adali 2017). Other features that relate to language complexity are the number of prepositions, exclusion words and conjunctions, sentence length and the number of words with six or more letters (Tausczik and Pennebaker 2010). I have selected a number of these features which I analyse using Voyant (Sinclair and Rockwell 2016), Excel and Python. Deceptive speech has been connected with reduced complexity and also the level of complexity can indicate whether the text is showing multiple perspectives or not (Tausczik and Pennebaker 2010, 34–35). Based on these findings I hypothesize that the fake news will be less complex linguistically than real news, although the result may be less distinct since both the real and fake news come from fact-checking sites.

The fact-checked news data come from a MisInfoText GitHub repository that contains articles that have been manually verified and labelled by fact-checking websites (Asr and Taboada 2019b, link in the references). Asr and Taboada scraped the websites of two fact-checking sites, Buzzfeed and Snopes, and then manually cleaned and assessed a randomly selected portion of them (2019a, 7-8). I chose to use only the Snopes dataset in this study since the topics of fake and real news in this dataset are more varied and the labels were the most suitable for comparing real and fake news (Asr and Taboada 2019a, 9). The Snopes dataset was also thoroughly cleaned, whereas the larger Buzzfeed dataset consistently lacked whitespaces between words which would have distorted the results.

2. Data Processing

The data was downloaded in CSV format, which I saved as an excel file for pre-processing the data. This included filtering the articles based on their label: I deleted articles labelled as "mixture", "mostly false" and "mostly true" so that only articles labelled as "true" and "false" remained. I divided the news to separate files according to their label, one containing all fake news and one all real news. Additionally, I deleted all unnecessary information such as the URL, titles and additional notes, so that the files contained only the full text of the original articles. I also manually deleted the text "[Your user agent does not support frames or is currently configured not to display frames. However, you may visit the page menu.]" from the real news dataset as it is not meant as a part of the original body text of the article. I ran the data through OpenRefine as I initially assumed that the dataset contained duplicates. However, it turned out that there were no exact duplicates in the dataset, but instead separate articles written on the same topic which initially seemed as duplicates. Since the data had already been cleaned thoroughly no changes were done to the data in the OpenRefine. I then converted the files back into CSV for it to work more fluently with the Python code. Next, I will describe the workflow for each of the different methods used in this study.

Voyant:

- I input the data to Voyant and it automatically computed the following information:
    - Fake news dataset contains 33,712 total words and 6,526 unique word forms.
    - Real news dataset contains 54,997 total words and 8,987 unique word forms.
    - Vocabulary Density (type-token ratio): fake news 19%, real news 16%.
    - Average Words Per Sentence: fake news 21.8, real news 19.5.
- I also searched the occurrence of a selection of conjunctions using the Voyant interface, which were chosen based on a reading of Longman Student Grammar of Spoken and Written English (Biber, Conrad and Leech 2002, 30–31). These were processed further in Excel.

Excel

- I entered the raw counts of the 14 conjunctions in Excel (*and*, *but*, *or*, *nor*, *if*, *as*, *after*, *because*, *since*, *although*, *while*, *than*, *that*, and *whether*), separately for both fake and real news datasets.

- I calculated the total number of conjunctions.
- Entered the total number of words to the excel file which I previously got from Voyant.
- I calculated the normed rates using the formula: Normed rate = (raw count / total word count) ∗ the fixed amount of text. In this study, the fixed amount of text was 100 words.
- Rounded the normed rates of conjunctions off to two decimal places.
    - The normed rate of the 14 conjunctions in total is 5.76 per 100 words for real news and 5.74 for fake news. The exact rates for all the conjunctions are visible in the table in the next section.

Python

I used Python to compute first the average word lengths in the datasets and then the standard variation of the word lengths, which enabled me to better evaluate its statistical significance. A code from Stack Overflow was adjusted to suit the needs of this study, and below is a description of what the code does. The exact code can be found in the GitHub repository for this project.

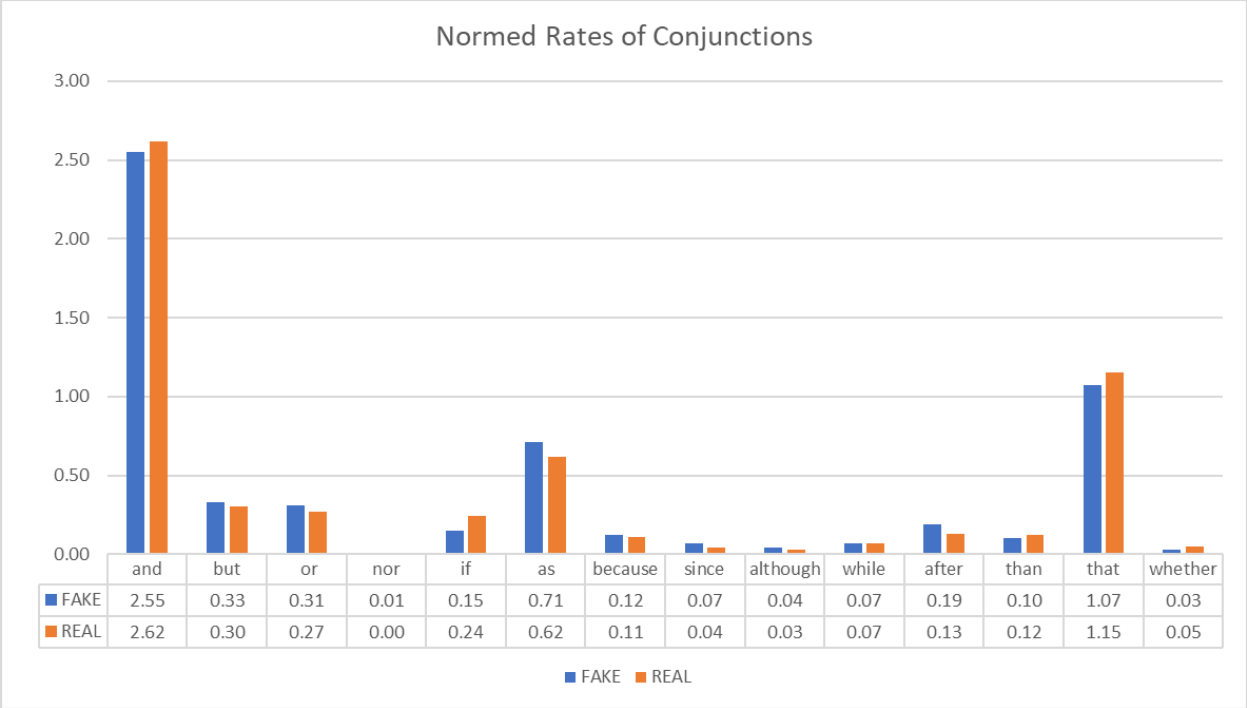
- The code filters out punctuation and numbers because otherwise each instance would have been counted as an individual word. The exact items that were excluded are ".,;!?-0123456789()[]{}.
- I initially intended to simply copy-paste the news data into the code, but because the data had empty lines between each sentence or each paragraph, the code was not able to read it beyond the first empty line. Removing the whitespaces would have been very time-consuming, so I created a step that fetched the data from the directory it is stored in.
- Prints out the average word length with all the decimals. When rounded off to two decimals the average lengths are 5.08 letters for fake news and 5.04 for real news.
- To test whether the difference is statistically significant, I included a step that counts the variance of the word lengths.
    - Variance for real news: 8.44
    - Variance for fake news: 8.56
- The code also imports a library called “math” in order to calculate the square root of a number, which is needed for the standard deviation. The code calculates and prints the square root of the variance:

- Standard deviation for real news: 2.91
- Standard deviation for fake news: 2.93

- The same code is repeated twice on different datasets (first for the real news data and then the fake news data).

3. Analysis and Discussion

The analysis shows that the differences in language complexity between the real news dataset and the fake news dataset are minimal. The average word lengths between the datasets differ by 0.04, and when considering that the standard deviation is between 2.91–2.93, the result is most likely not statistically significant. Similarly, the sentence length and conjunction use differ only marginally from each other, as shown in the table "Normed Rates of Conjunctions". The type-token ratio which describes the vocabulary complexity would indicate that the fake news is even slightly more complex than the real news, as it contains less repetition. Thus, the hypothesis that real news would be more complex than fake news was incorrect. However, the contradictory result may be caused by defects in the dataset as well as a methodology that was not refined enough.



	and	but	or	nor	if	as	because	since	although	while	after	than	that	whether
FAKE	2.55	0.33	0.31	0.01	0.15	0.71	0.12	0.07	0.04	0.07	0.19	0.10	1.07	0.03
REAL	2.62	0.30	0.27	0.00	0.24	0.62	0.11	0.04	0.03	0.07	0.13	0.12	1.15	0.05

One aspect that should be improved in this study is the size of the dataset. It was insufficient especially for some of the less frequent conjunctions, such as *nor*. The normed difference is 0.01,

and the actual raw occurrences were 2 in fake news and 0 in real news, which are too low to provide reliable information. Additionally, not all the articles in the MisInfoText dataset fit the criteria generally given to fake news, which defines fake news as non-factual content that tries to appear as if it were legitimate news to gain the credibility generally given to news media (Tandoc, Lim and Ling 2018, 143 and Gelfert 2018, 108). It seems that there is a lot more variation in the types of articles included in the data, the fake news dataset included, for example, an article from RationalWiki, which is a completely different genre. It is not news and it does not attempt to appear as such, which might have distorted the results.

However, what caused the lack of difference was most likely the fact that both the fake and real news stories were collected from fact-checking sites, instead of comparing, for example, the most trusted news outlets with counter media. The choice to compare fact-checked real and fake news was made deliberately to avoid possible bias caused by the decision making of fact-checking sites. They have been criticised for biased decision making since the articles are picked by individual people whose beliefs may influence the process (Asr and Taboada 2019a, 4) and it has been shown that fact-checking sites are more likely to pick up negative ads rather than neutral or positive ones (Amazeen 2016, 442). Furthermore, fact-checking websites aim to find and reveal misinformation, not to confirm accurate news, which affects the articles they choose to check (Amazeen 2016, 451). Therefore the articles included in the current dataset are likely to have all been written in a similar style. This indicates that the difference in complexity is not related only to the veracity of the articles, but rather their news genre.

4. References

Amazeen, Michelle A. 2016. “Checking the Fact-Checkers in 2008: Predicting Political Ad Scrutiny and Assessing Consistency.” *Journal of Political Marketing* 15 no. 4: 433–464.

Asr, Fatemeh Torabi, and Maite Taboada. 2019a. “Big Data and Quality Data for Fake News and Misinformation Detection.” *Big data & society* 6, no. 1: 1–14.

———. 2019b. MisInfoText. A collection of news articles, with false and true labels. Dataset. Accessed 3 November 2020. https://github.com/sfu-discourse-lab/Misinformation_detection.

Biber, Douglas, Susan Conrad, and Geoffrey Leech. 2002. *Longman Student Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Gelfert, Axel. 2018. "Fake News: A Definition." *Informal Logic* 38, no.1: 84–117.

Horne, Benjamin D., and Sibel Adali. 2017. "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News." Accessed 15 October 2020. https://arxiv.org/abs/1703.09398.

Sinclair, Stéfan and Geoffrey Rockwell. 2016. *Voyant Tools*. Accessed 3 December 2020. http://voyant-tools.org/.

Tandoc, Edson C. Jr., Zheng Wei Lim and Richard Ling. 2018. "Defining "Fake News": A typology of scholarly definitions." *Digital Journalism* 6 no. 2: 137–153.

Tausczik, Yla R., and James W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29, no. 1: 24–54.