Final Project Template

Danna Shao
d.shao@student.vu.nl

1 Introduction

1-2 paragraphs on the relevance of the topic and the choice of target languages

Background: Abortion related news

Language choice: According to Amy T. Schalet, despite Dutch teenagers initiate sex at comparable ages, their birth and abortion rates are eight and two times lower than American teenagers, and this difference is caused by different culture frames.

2 Dataset Description

2.1 Crawling and Filtering Choices

The English data is scraped from the guardians with its official api (https://open-platform.theguardian.com/documentation/). The Dutch data is scraped from nos.nl by html scraping.

The English data is scraped using search keyword 'abortion' and date range 2022-06-24 to 2022-09-24', three months after the Roe v. Wade overturned. The Dutch data is scraped using search keyword 'abortus' and page range 1 to 35 (with invalid links removed) since the data is way less than the English ones.

Each language's dataset contains 600 articles, randomly(seed=1) splitted into 80 percent train and 20 percent test. The metadata for these two dataset are stored separately in corresponding .csv files.

2.2 General Descriptive Statistics (train sets)

Stats	eng	nld
article length	7483.06	3609.53
title length	79.90	63.55
sentence length	27.08	17.28
type-token-ratio	0.43	0.52

Table 1: Genral descriptive statistics of the train sets.

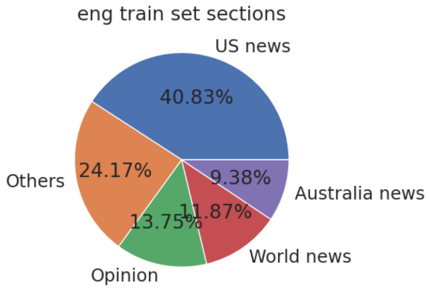

Figure 1: Section distribution for eng train set

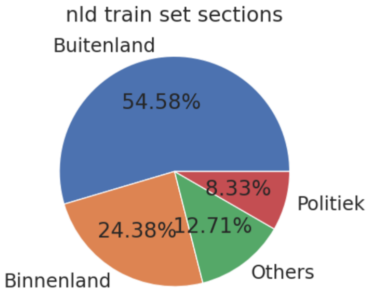

Figure 2: Section distribution for nld train set

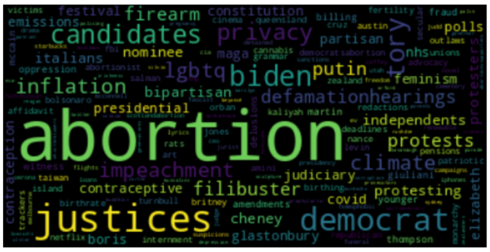

Figure 3: Keywords for eng train set

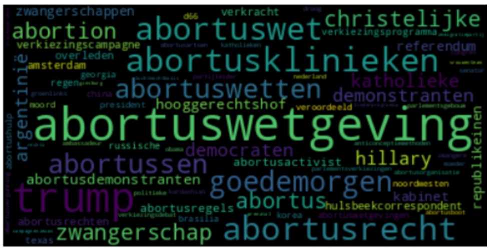

Figure 4: Keywords for nld train set

3 Explorative Linguistic Analysis

3.1 Major Findings

- Analysis goal: Find emotional and subjectivity difference in two languages corresponding to domestic or foreign news.
- Motivation: Explore the attitude difference between American and Dutch culture towards abortion.
- Methodological choices: NLP pipeline TextBlob as it contains nld and eng sentiment analysis.
- Findings (Hypotheses): Dutch news will have more positive attitude if it belongs to the domestic ('Binnenland') section than foreign ('Buitenland') news.

Stats	Domestic	Foreign
polarity	0.03	0.01
subjectivity	0.49	0.50
readability	63.89	63.54
type-token-ratio	0.52	0.52
sentence length	17.14	17.40

Table 2: Difference between domestic and foreign section for the nld news train set

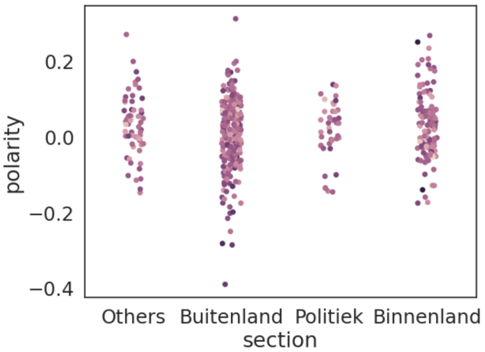

Figure 5: Polarity according to sections for nld train set, colored by its subjectivity. The darker the color the more subjective the article is.

3.2 Other Findings

Subjectivity may have a positive correlation between readability. This may be caused by the higher subjectivity the article have, the more likely it has a persuasive intention and thus needs to be easier to read.

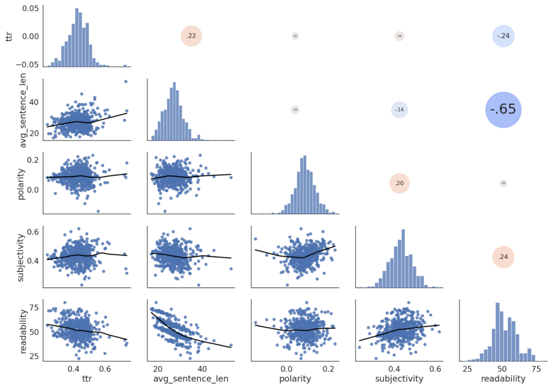

Figure 6: Correlation matrix for eng train set.

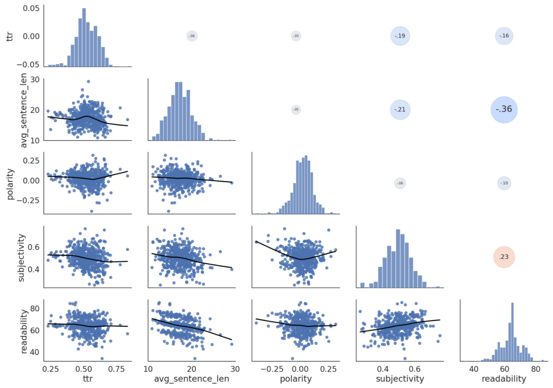

Figure 7: Correlation matrix for nld train set.

4 Discussion

Reflection on the findings, explanation of **testable hypotheses** that will be evaluated on the test set, discussion of alternative explanations

References

Author Name. Year. *Example Title*. Publisher.