A Comparative Study of English and Dutch Abortion-themed Press Release Texts

Danna Shao
d.shao@student.vu.nl

1 Introduction

The subject of abortion has continuously produced controversy. Its dispute revolves around the mother's right to personal freedom and the fetus's right to life is mingled with very complicated religious, cultural, political, and other societal backdrops. Shocking the world in 2022, the U.S. Supreme Court overturned Roe v. Wade three days before the Dutch parliament passed a law doing away with the mandatory five-day reflection period before an abortion (Eerste Kamer, 2022), as women are believed to be capable of making their own choices with the help of the doctors.

According to Amy T. Schalet, despite Dutch and the US teenagers initiate sex at comparable ages, Dutch adolescent fertility and abortion rates are eight and two times lower than American youths (Schalet, 2011). Amy argues in her book *Not Under My Roof* that the different cultural frameworks in the US and the Netherlands account for this discrepancy. Therefore, Dutch and English texts might both reflect this cultural difference. Through sentiment analysis on news texts reporting on the subject of abortion in both Dutch and English, this project aims to investigate this distinction.

2 Dataset Description

This section provides how the data is obtained and pre-processed, together with some general descriptive statistics of the datasets.

2.1 Crawling and Filtering Choices

The English data is scraped from the prestigious Guardian newspaper with open access to all its online content with its official api (Guardian, 2023), whereas the Dutch data is scraped from the Nederlandse Omroep Stichting (NOS) `nos.nl` by html scraping. As one of the largest news websites in the country, NOS is a government-funded broadcasting organization that is a part of the Dutch public broadcasting system.

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. Given the abundance of data in English, the three-month discussion is more likely to focus on this U.S. Supreme Court decision and its social background. The Dutch data is scraped using search keyword 'abortus' and page 1 to 35 (with invalid links removed) since the data is way less than the English ones, ranges from 2016 to present.

Each language's dataset contains 600 articles, randomly (seed=1) split into 80% train and 20% test sets. The metadata for these two dataset are stored separately in corresponding .csv files. Since data from both sources are already clean, no extra pre-process is involved except for the line wrapping between paragraphs is removed since it is not important for sentiment analysis.

2.2 General Descriptive Statistics

To provide a first impression of the dataset, this section lists general statistics of the training set. The text-token-ratio is calculated with the help of python package `stanza` using its tokenizer.

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.

The table shows that NOS articles are significantly shorter than the Guardian's. The potential influence of this difference will be examined in the following sections.

2.2.1 Section Distributions

The sections of the articles are classified by both news medias and can be obtained directly from the data source. Below are the distribution of the sections for both training sets.

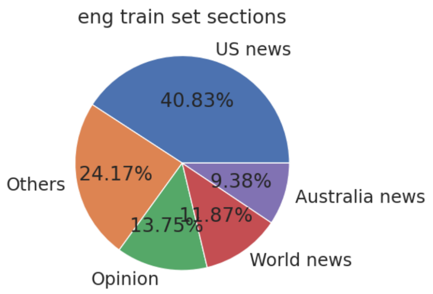


Figure 1: Section distribution for English training set. "Others" contains 15 articles from "Politics", 13 articles from "Global development" and other 21 categories having less than 10 articles each.

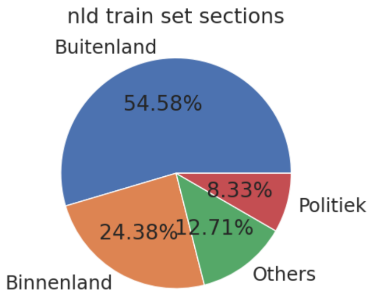


Figure 2: Section distribution for Dutch training set. "Buitenland" stands for foreign and "Binnenland" means domestic. ("Politiek" is indeed politics.) "Others" contains 18 articles without section and 23 other categories having less than 5 articles each.

It is also shown that, despite the Guardian is an UK news media while NOS is a Dutch one, most of the news is still about US and foreign countries. An attitude comparison can therefore be conducted on those datasets.

2.2.2 Key Words

The key words are extracted with `KeyBERT` (Grootendorst, 2020), a minimal keyword extractor using BERT embeddings with multilingual support, and visualized with `WordCloud` (Mueller, 2020)

Except for the abortion related words, we can also see politician names and political entities from the keyword word cloud for both languages' dataset. Dutch keyword word cloud also shows that there are plenty of NOS articles are focusing on US politics.

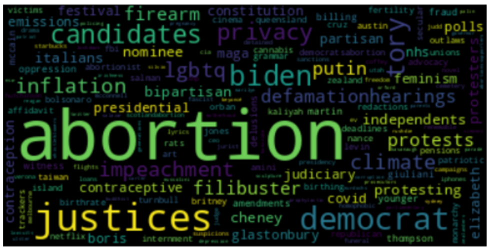


Figure 3: Keywords for the English training set.

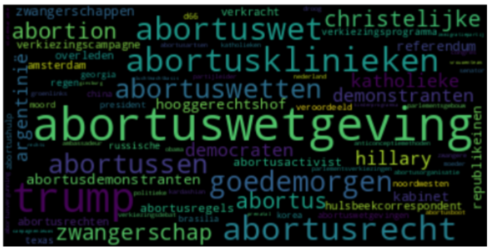


Figure 4: Keywords for the Dutch training set.

3 Explorative Linguistic Analysis

This section demonstrates the major and other findings of the difference in English and Dutch news texts. The analysis focus on the how Dutch and English article differs in sentiment when reporting domestic or foreign news in order to reveal the attitude differences. Readability of the articles is another factor taken into account. The correlations between readability and sentiment scores are also examined to gain insights.

3.1 Major Findings

In order to examine how American and Dutch cultures differ in their attitudes on abortion, the aim is to investigate the subjective and emotional differences between domestic or international news from two languages' datasets.

The sentiment analysis of this study is conducted using `Pattern` (Tom De Smedt, 2020) by API provided with `TextBlob` (Loria, 2020). `Pattern` is a multilingual web mining tool that integrates sentiment analysis based on pattern library. Two sentiment scores, polarity and subjectivity, are evaluated for each article. Polarity score indicates how positive or negative the text is; positive polarity scores indicates the article has overall positive emotion. Subjectivity score ranges from 0 (very objective) to 1 (very subjective). The scores are assessed based on a lexicon of adjectives (such as great, awful, etc.) with their corresponding polarity and subjectivity scores.

Based on the previously stated facts on the

Netherlands versus the US, it is assumed that Dutch news are more positive when reporting domestic ("Binnenland") events than those from other countries ("Buitenland"). This hypothesis is confirmed in the training set and will be examined on the test set in the following section.

NLD	Domestic	Foreign
polarity	0.031	0.011
subjectivity	0.493	0.499
ENG	**US**	**non_US**
polarity	0.088	0.086
subjectivity	0.431	0.430

Table 2: Sentiment difference between domestic and foreign section for the training sets.

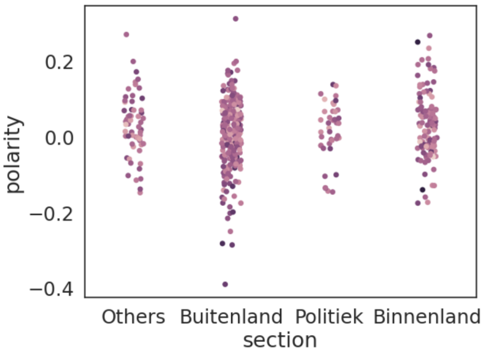


Figure 5: Polarity according to sections for nld train set, colored by its subjectivity. The darker the color the more subjective the article is. This figures shows that domestic articles generally have a more positive sentiment. The plot for all the datasets are listed in Appendix 8.

3.2 Other Findings

Readability is the ease with which a reader can understand a written text. As one of the major factors in evaluating a press release is its professionalism and target audiences, readability is therefore an important measure. Flesch reading-ease test is a measure for article readability developed by Rudolf Flesch in *The art of readable writing* (Flesch, 1962). Modified by Douma (Douma, 1960), this measure is also suitable for Dutch texts. The Flesch-Douma formula is given by

$$206.835 - 0.93\left(\frac{total\ words}{total\ sentences}\right) - 77\left(\frac{total\ syllables}{total\ words}\right)$$

The scores can be interpreted as educational level. The lower readability score a text have, the higher educational level it corresponds to. Texts scored 0-30 are for academics, 30-50 are for college students, 50-70 for secondary education and 70-100 for primary school. In this project, this score is calculated with `TextSTAT` (Hüning, 2022)

Subjectivity seems to have a positive correlation between readability on both languages. One possible explanation is that the higher subjectivity the article have, the more likely it has a persuasive intention and thus needs to be easier to read.

Another cross lingual pattern is the "V" shape in polarity-subjectivity correlation, that is, subjectivity have a positive correlation with the absolute value of polarity (see Fig. 9). The stronger emotion an article has leads to a higher subjectivity score. This may be caused by subjective feelings are always emotional compared with objective facts derived from rational inferences.

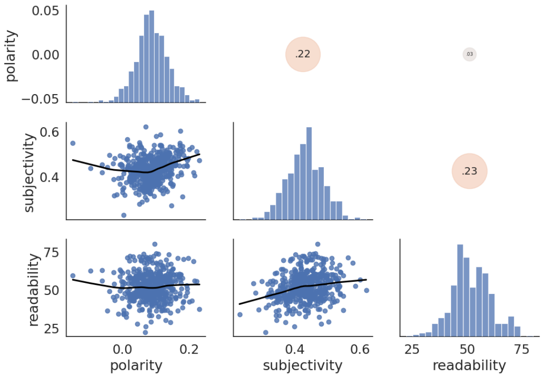


Figure 6: Correlation matrix for the English training set. Correlation values are calculated with Spearman rank correlation since we do not assume linear correlations.

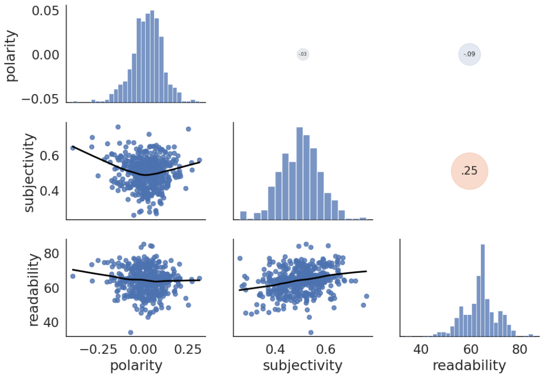


Figure 7: Correlation matrix for the Dutch training set. The correlation between subjectivity and polarity is negative here because of the polarity is of the original but not absolute value.

4 Discussion

In section 3, we stated our primary and secondary hypotheses with their corresponding explanations based on the findings on the training sets and known facts:

- Dutch news articles about abortion are more positive on domestic news than foreign news, since Dutch culture is more tolerant towards this topic.
- Subjectivity is positive correlated to readability and (absolute value of) polarity, since persuasive contents need to be easier to read and subjective contents are more emotional.

In this section, the hypotheses are tested on the test sets. Alternative explanations towards the results are also provided.

4.1 Hypotheses testing

Unfortunately, our first hypothesis is not true on the Dutch test set:

NLD	Domestic	Foreign
polarity	0.026	0.028
subjectivity	0.497	0.500
ENG	US	non_US
polarity	0.096	0.077
subjectivity	0.439	0.432

Table 3: Sentiment difference between domestic and foreign section for the test sets.

On the contrary and surprisingly, this is true for the English articles. Same as those in the training set, articles from English test set are more positive on US news than non-US news. The Spearman rank correlation on country (domestic/foreign) and polarity is -0.21 ($p = 0.02$) for English and -0.03 ($p = 0.77$) for Dutch, implying there is no significant correlation on Dutch test set. However, for the training sets, the statistics are -0.01 ($p = 0.82$) for English and -0.09 ($p = 0.05$) for Dutch. These values shows that more data is needed for a solid conclusion.

The secondary hypothesis has been verified on the test sets for both languages. The correlations are significant despite not very strong.

NLD	score	p-value
sub-ABSpol	0.232	0.018
sub-read	0.210	0.021
ENG	US	non_US
sub-ABSpol	0.338	0.0002
sub-read	0.215	0.018

Table 4: Spearman rank correlation between subjectivity (sub), absolute value of polarity (ABSpol) and readability (read) on the test sets.

4.2 Alternative explanations

The alternative explanation of the secondary hypothesis is that, since sentiment scores are calculated based on pattern dictionaries, the adjectives with higher polarity scores are also the ones with higher subjectivity scores. Despite this may be true under certain situations, for example, "disgusting" can be a word that has high score for both measurements, we are unsure about the accuracy since the context is lost.

5 Future work

Firstly, more data is needed for the sentiment analysis on domestic versus foreign comparison. Since the Dutch articles are significantly shorter than those of English, it is important to obtain data from more different media sources.

Secondly, more advanced sentiment analysis tools that takes contexts into account (e.g. large language models) are required for a more reliable conclusion on subjectivity and polarity.

Data availability statement

The data and the analyzing programs that support the findings of this study are available in dannashao/LD at https://github.com/dannashao/LD/tree/main/finalproject_DannaShao. The repository will be open access after the grading of this project within three months.

The data were derived from the following resources available in the public domain: the Guardian (https://www.theguardian.com/), NOS (https://nos.nl/).

References

W.H. Douma. 1960. *De leesbaarheid van landbouwbladen : een onderzoek naar en een toepassing van leesbaarheidsformules*. Number no. 17 in Bulletin / Afdeling sociologie en sociografie van de Landbouwhogeschool Wageningen.

Eerste Kamer, 2022. 2022. Stemming afschaffing minimale beraadtermijn voor afbreking van zwangerschappen. `https://www.eerstekamer.nl/verslagdeel/20220621/afschaffing_minimale_beraadtermijn`. Online; accessed: 2023-12-06.

R. Flesch. 1962. *The Art of Readable Writing*. Wiley.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Guardian, 2023. 2023. The guardian open platform documentation.

Matthias Hüning. 2022. Textstat - simple text analysis tool.

Steven Loria. 2020. Textblob documentation.

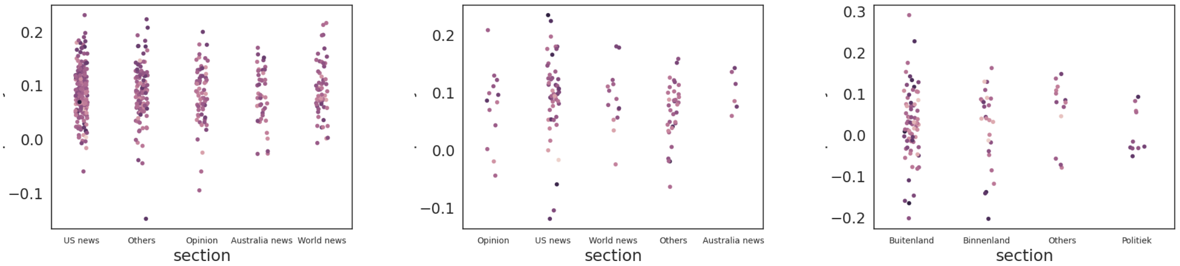
Andreas Mueller. 2020. Wordcloud for python documentation.

A.T. Schalet. 2011. *Not Under My Roof: Parents, Teens, and the Culture of Sex*. University of Chicago Press.

Walter Daelemans Tom De Smedt. 2020. Pattern documentation.

Appendices

A Polarity strip plots



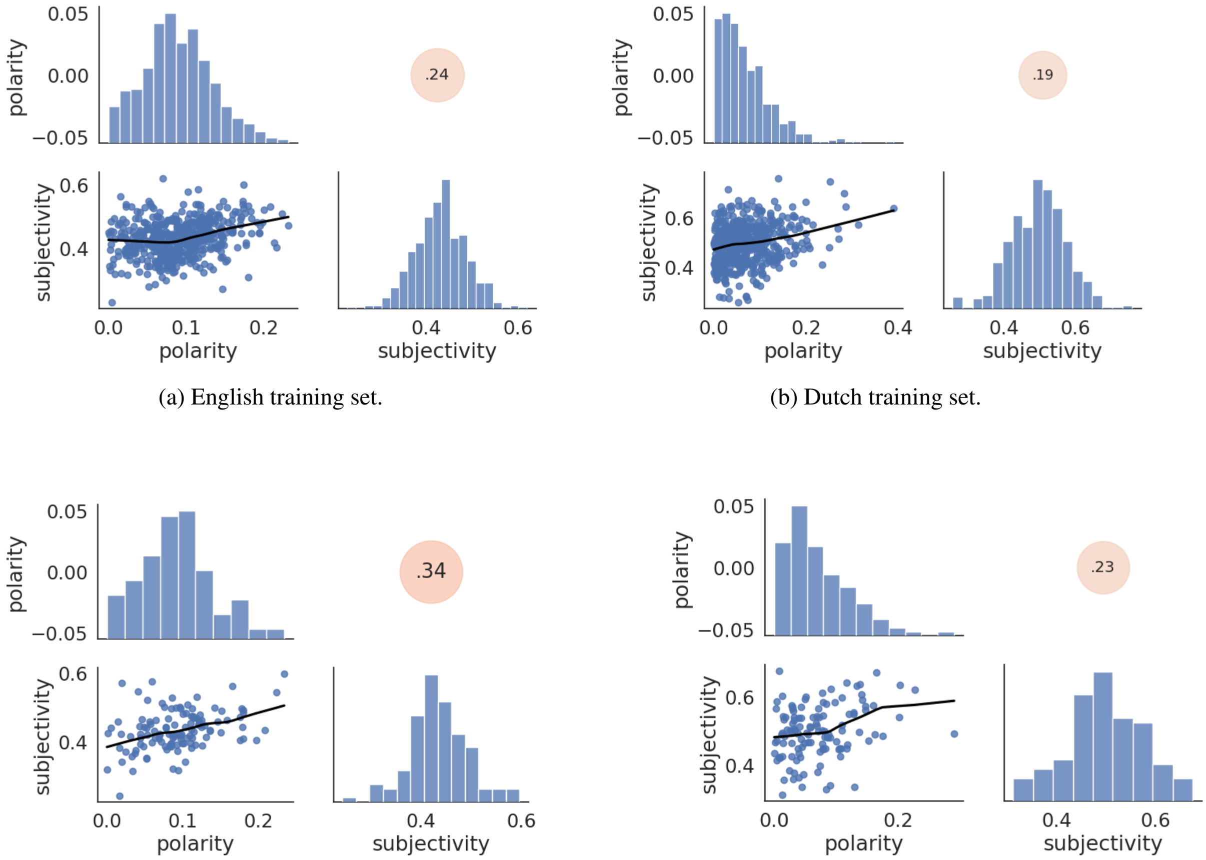
(a) English training set section polarity.

(b) English test set section polarity.

(c) Dutch test set section polarity.

Figure 8: Polarity according to sections colored by subjectivity.

B Correlation plots with absolute value polarity



(a) English training set.

(b) Dutch training set.

(c) English test set.

(d) Dutch test set.

Figure 9: Correlation plots with absolute value polarity