Comparing text processing approaches

Shiva Kumar Pendem

Abstract

This is the first task of the assignment provided. We need to take the given data and pre-process it in two different ways. Then, we need to tokenize the data and produce the outputs of the questions posed in the assignment.

1 Dataset

The dataset is a csv file named 'trainingdata-all-annotations.csv' which contains 2814 tweets of about various targets and the stance of that tweet, opinion, and sentiment of the tweet. Targets included in the data set are 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', 'Legalization of Abortion'. The Stances of the dataset are 'Against', 'Favor', 'None'. The possible sentiments of the tweet are 'Positive' and 'Negative'. The opinion towards could be 'Other' or 'Target'.

2 Pre-processing of data

There are two kinds of pre-processing for this assignment: maximal pre-processing and minimal pre-processing. In maximal pre-processing, we remove words that contain '@' tags and '#' tags, normalize the words to lowercase letters, define our own stop words, and then remove those stop words from the tweet.

In the case of minimal pre-processing, instead of deleting the entire word that contains '@', we anonymize the user who posted that tweet, normalize all the words to lowercase except the words that are fully uppercase.

3 Programming language and Libraries used

I used **Python** for this programming assignment. The libraries I used are **NumPy**, **Pandas**, **collections**, and **Scikit-learn**. All the programming has been done in **Jupyter Notebook**, which is a notebook-style environment to run code in blocks.

4 Stop words

Stop words are the words in a stop list which are to be filtered because they are insignificant. I have considered 40 most common words in all of the tweets as stop words.

5 Questions

5.1 What is the average length of each instance?

The average length of each instance for maximal pre-processing is *62.42*. And, the average length of each instance for minimal pre-processing is *103.32*.

5.2 What is the total number of words in the corpus?

The total number of words in the corpus is *305,052*. The total number of words in corpus for maximal pre-processing is *175,648*. And the total number of words in the corpus for minimal pre-processing is *290,752*.

5.3 What is the average length of tweets for each target?

Target	Maximal	Minimal
Atheism	63.98	106.46
Climate Change is a Real Concern	58.17	101.12
Feminist Movement	66.79	105.26
Hillary Clinton	58.68	99.05
Legalization of Abortion	63.01	104.47

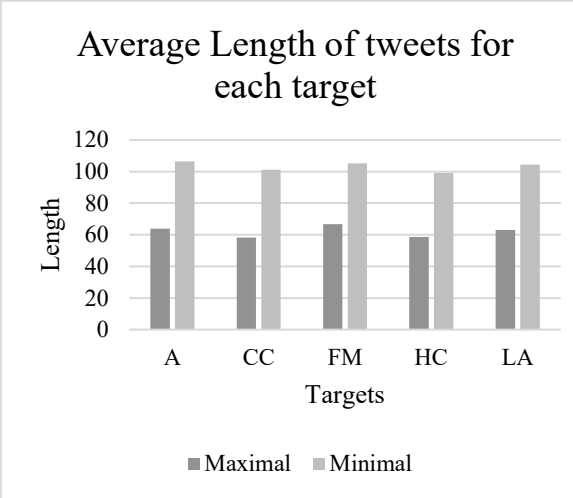

Fig. (a) Average length of tweets for each target.

In the figure above A, CC, FM, HC, LA refers to targets 'Atheism', 'Climate Change is a real concern', 'Feminist Movement', 'Hilary Clinton' and 'Legalization of Abortion' respectively.

5.4 What is the average length of tweet for each instance type?

Stance	Maximal	Minimal
Against	65.34	106.38
Favor	63.95	103.92
None	55.59	97.17

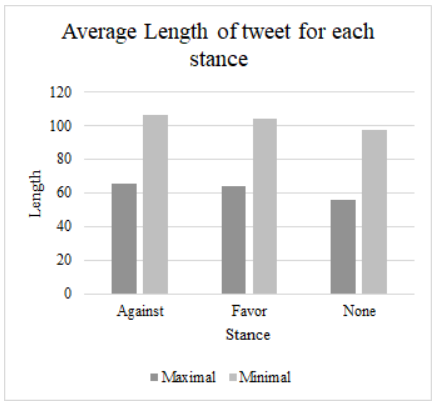

Fig. (b) Average length of tweets for each target.

5.5 What is the average length of tweet for each stance type across targets?

	Against	Favor	None
Targets			
A	107.23	106.91	104.11
CC	109.73	104.91	95.58
FM	108.27	106.14	95.97
HC	103.92	93.91	91.91
LA	106.28	105.55	100.11

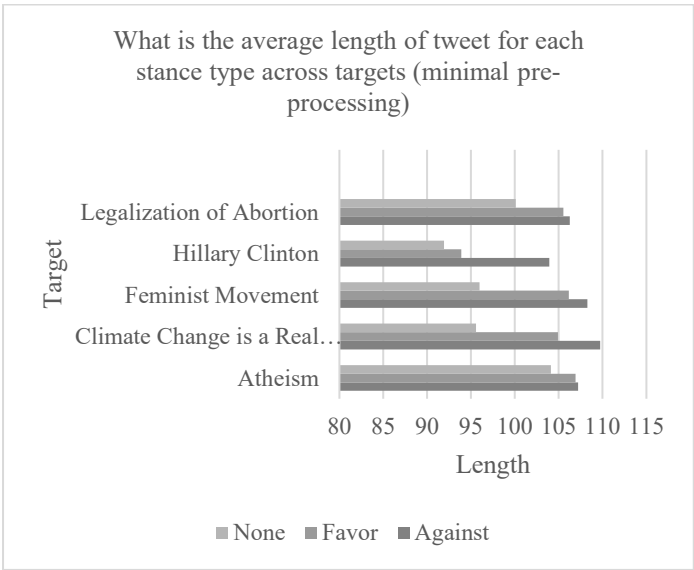

Fig. (c) Average length of tweets for each stance type across targets for minimal pre-processing.

	Against	Favor	None
Targets			
A	63.16	69.63	61.67
CC	75.46	62.68	50.92
FM	71.3	65.66	56.79
HC	62.06	55.53	53.46
LA	64.54	67.11	57.27

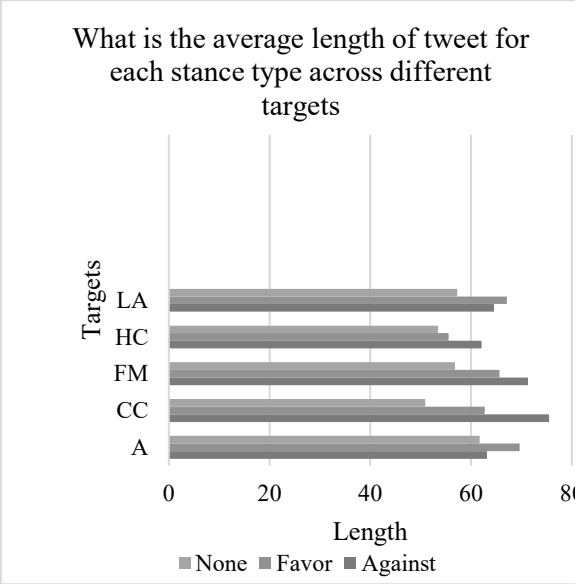

Fig. (d) Average length of tweets for each stance type across targets for maximal pre-processing.

6 Pre-processing across different platforms

I noticed that the vocabulary used across different types of media varies considerably. For example, newspapers generally employ a wider range of words that are more easily comprehended, though this can depend on the subject matter (e.g. the Financial Times vs. The Sun). When it comes to Twitter, which is mostly text-based, there is a great deal of data present in the form of tweets.

However, it is not enough to simply employ pre-processing techniques when politicians are attempting to communicate via this platform.
Reddit is a platform where pre-processing techniques do not always prove to be successful. For instance, the r/wallstreetbets [1] community utilises words such as "stonks," "tendies," "HODL," "YOLO," and other unusual terms that the average person might not be familiar with. These particular words, which are exclusive to this particular community, demonstrate the complexity of the language found on Reddit and the need to employ more specific strategies when attempting to decode its content.

7 Pre-processing for my own stance detection

In my research on stance detection, I would adopt similar pre-processing conventions to those used in a comparable assignment. However, I would add a neural network to my pre-processing model that prioritizes the context in which words are used. Given the importance of context, especially on social media platforms like Twitter, this addition would help to capture the intent of tweets and improve the accuracy of the classification process.

For instance, if a comedian tweets a joke about a politician that includes an offensive word, my machine learning model would analyze the word in the context of the tweet and assess its intention. If the model classified the tweet as a joke, the offensive word would remain; if the tweet was classified as "intended harm," the word would be blurred out. This approach considers not only individual words, but also the broader context, which is vital for interpreting meaning and classifying tweets more accurately.

[1] https://www.moneyunder30.com/what-is-r-wallstreetbets