|  |
| --- |
| **Evaluating different Text processing approaches** |
|  |
| **Kaushik Parvathaneni** |
|  |
|  |
|  |
|  |

Abstract

This is the my first module assignment . We need to take the given data and pre-process it in two different ways which are minimal pre-processing and maximal pre-processing. Then, we need to tokenize the data and produce the outputs of the questions posed in the assignment.

Dataset

The dataset consists of 2814 tweets on various targets, together with their stance, opinion, and sentiment, and is stored as a csv file with the name "trainingdata-all-annotations.csv." Targets including "Atheism," "Legalization of Abortion," "Climate Change is a Real Concern," "Feminist Movement," and "Hillary Clinton" are included in the data set. "Against," "Favor," and "None" are the dataset's positions. The sentiment of the tweet could be either "Positive" or "Negative." "Other" or "Target" are two possible attitudes.

Pre-processing of data

For this assignment, there are two types of pre-processing: maximal pre-processing and minimal pre-processing. The words that contain the "@" and "#" tags are removed, the words are normalized to lowercase letters, we construct our own stop words, and finally we delete those stop words from the tweet. This is known as maximal pre-processing.

For minimal pre-processing, we anonymize the user who sent the tweet, normalize all the words to lowercase, save for the words that are entirely uppercase, and do not remove the entire word that contains the @ symbol.

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.

Stop words

Stop words are words that should be filtered out of a stop list because they are unimportant. All of the tweets' top 50 most frequent words were taken into consideration as stop words.

Questions

What is the average length of each instance?

The average length of each instance for maximal pre-processing is *60.48*. And, the average length of each instance for minimal pre-processing is *86.66*.

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 *170,199*. And the total number of words in the corpus for minimal pre-processing is *243,879*

*.*

What is the average length of tweets for each target?

|  |  |  |
| --- | --- | --- |
| **Target** | **Maximal** | **Minimal** |
| **Atheism** | 62.21 | 87.20 |
| **Climate Change is a Real Concern** | 56.81 | 86.50 |
| **Feminist Movement** | 63.95 | 87.96 |
| **Hillary Clinton** | 57.29 | 84.97 |
| **Legalization of Abortion** | 60.96 | 86.67 |

*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.

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

|  |  |  |
| --- | --- | --- |
| **Stance** | **Maximal** | **Minimal** |
| **Against** | 63.41 | 88.85 |
| **Favor** | 61.66 | 87.20 |
| **None** | 54.01 | 82.17 |
|  |  |  |

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

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Against** | **Favor** | **None** |
| **Targets** |  |  |  |
| **A** | 85.74 | 90.53 | 88.39 |
| **CC** | 95.2 | 89.83 | 81.54 |
| **FM** | 91.14 | 87.19 | 80.95 |
| **HC** | 89.76 | 80.18 | 77.79 |
| **LA** | 88.15 | 86.52 | 83.74 |

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | Against | Favor | None |
| **Targets** |  |  |  |
| **A** | 61.49 | 67.53 | 59.88 |
| **CC** | 74.46 | 61.41 | 49.44 |
| **FM** | 68.45 | 62.35 | 54.90 |
| **HC** | 60.70 | 54.04 | 52.07 |
| **LA** | 62.63 | 63.76 | 55.78 |

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

Pre-processing across different platforms

I discovered that the terminology utilized in various media varied greatly. Newspapers, for instance, typically use a greater variety of terms that are easier to understand, however this can vary depending on the topic matter (e.g. the Financial Times vs. The Sun). There is a lot of information available on Twitter, which is primarily text-based, in the form of tweets. However, solely using pre-processing techniques is insufficient when politicians are attempting to communicate via this platform. On the Reddit site, pre-processing methods are not always successful. Because of the complexivity over there.

Pre-processing for my own stance detection

I would utilize pre-processing procedures that are equivalent to those used in a comparable assignment in my research on stance detection. I would, however, incorporate a neural network into my pre-processing model that gives the use of words in context priority. This inclusion would help to capture the intent of tweets and enhance the precision of the classification process because context is so important, especially on social media platforms like Twitter.

For instance, my machine learning model would evaluate the word in the tweet's context and determine whether it was intended to be offensive if a comedian tweeted a joke about a politician that contained that word. The derogatory word would be present if the tweet was categorized as a joke, but it would be blurred out if it was categorized as having "intended harm." This method takes into account both the individual words and the larger context, which is important for deciphering meaning and correctly categorizing tweets.