|  |
| --- |
| **Feature Selection: Comparing Features** |
|  |
| **Kaushik Parvathaneni** |
|  |
|  |
|  |
|  |

Abstract

This is the my first module assignment . We need to take the given data and do feature selection comparing the features.

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 "newinfo.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.

Polarity Calculation

For this assignment, first I have done POS tagging using NLTK and then calculated polarity with respective functions from part 1.

Programming language and Libraries used

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

Questions

What are the average polarity scores?

Average polarity of all the maximal filtered data 0.2566237694997919. Average polarity of all the minimal filtered data is 0.1827827710364166.

What are the average polarity scores across different kinds of stances?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Against** | **Favor** | **None** |
| **Stances** |  |  |  |
| **Positive** | 0.19 | 0.21 | 0.21 |
| **Negative** | 0.05 | 0.04 | 0.05 |
| **Average** | 0.25 | 0.25 | 0.26 |
|  |  |  |  |

*Max processing*.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Against** | **Favor** | **None** |
| **Stances** |  |  |  |
| **Positive** | 0.14 | 0.15 | 0.14 |
| **Negative** | 0.03 | 0.03 | 0.03 |
| **Average** | 0.18 | 0.18 | 0.17 |

*Min processing*.

What are the average polarity scores across different targets?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **positive** | **negative** | **average** |
| **Targets** |  |  |  |
| **A** | 0.20 | 0.04 | 0.25 |
| **CC** | 0.20 | 0.04 | 0.24 |
| **FM** | 0.19 | 0.06 | 0.26 |
| **HC** | 0.20 | 0.04 | 0.25 |
| **LA** | 0.20 | 0.05 | 0.26 |

*Max processing*.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **positive** | **Negative** | **average** |
| **Targets** |  |  |  |
| **A** | 0.15 | 0.03 | 0.18 |
| **CC** | 0.14 | 0.02 | 0.16 |
| **FM** | 0.14 | 0.04 | 0.19 |
| **HC** | 0.13 | 0.03 | 0.16 |
| **LA** | 0.14 | 0.03 | 0.18 |

*Min processing*.

What kind of challenges does NLTK face when POS tagging social media   
data?

Challenges nltk face when pos tagging social media data are

-Sparse data

-Noise in data

-Ambiguity

-Annotation challenges

-Difference in language

What kind of mistakes does NLTK make when trying to POS tag social media   
data?

The mistakes nltk make when trying to pos tag social media data are:

-Incorrect tagging

-Misinterpretation of Emoticons and Emojis

-Misinterpretation of Slang and Informal Language

1. What pre-processing steps seem to be “easier” for NLTK to accurately POS   
   tag?

Pre-processing steps which would make easier for NLTK to accurately POS tag are:

-Text Normalization

-Tokenization

-Removing Stop Words

-Stemming and Lemmatization

**8)** **Based on the descriptive statistics, do you think adding polarity scores as a feature would be helpful for classification?**

Polarity scores can be a useful as they provide information about the sentiment expressed in a piece of text.