Feature selection: Comparing Features

Shiva Kumar Pendem

Abstract

This is the second task of the assignment provided. In this task, we have to do POS [1]tag on the dataset and measure polarity [2]of a tweet.

1 Dataset

The dataset used in this task is same dataset as the task 1 but has some additional columns such as the both pre-processed tweets and length of the tweet. For finding the polarity of the tweet I have downloaded a excel file from Kaggle[3], which contains Positive and Negative words.

2 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.

3 Measuring Polarity

3.1 What are the average polarity scores?

Polarity	Maximal	Minimal
Avg. Positive	0.203	0.214
Avg. Negative	0.05	0.037
Average Total	0.253	0.251

3.2 What are the average polarity scores across different kind of stances?

Stances	Polarity	Maximal	Minimal
AGAINST	**AP**	0.19	0.21
	AN	0.05	0.04
	AT	0.24	0.25
FAVOR	**AP**	0.2	0.21
	AN	0.04	0.03
	AT	0.25	0.24
NONE	**AP**	0.21	0.2
	AN	0.05	0.03
	AT	0.26	0.24

3.3 What are the average polarity scores across different targets?

Targets	Polarity	Maximal	Minimal
A	**AP**	0.2	0.23
	AN	0.04	0.03
	AT	0.25	0.03
CC	**AP**	0.2	0.2
	AN	0.03	0.02
	AT	0.24	0.23
FM	**AP**	0.19	0.2
	AN	0.06	0.04
	AT	0.25	0.25
HC	**AP**	0.2	0.2
	AN	0.04	0.03
	AT	0.25	0.23
LA	**AP**	0.2	0.21
	AN	0.05	0.03
	AT	0.25	0.25

1 https://en.wikipedia.org/wiki/Part-of-speech_tagging

2 Polarity refers to the overall sentiment conveyed by a particular text, phrase or word.

3 Postive and Negative words

In the table above, A, CC, FM, HC, LA refers to Atheism, Climate Change is a Real Concern, Feminist Movement, Hilary Clinton, Legalization of Abortion respectively.

Note: In the tables of this report, AP, AN, AT refers to scores of Average Positive Polarity, Average Negative Polarity, Average Total Polarity respectively.

3.4 What are the average polarity scores across targets and stances?

Stances	Targets	Polarity	Maximal	Minimal
AG	A	AP	0.21	0.2500
		AN	0.04	0.03
		AT	0.25	0.28
	CC	AP	0.08	0.17
		AN	0.03	0.03
		AT	0.25	0.24
	FM	AP	0.19	0.21
		AN	0.07	0.05
		AT	0.26	0.26
	HC	AP	0.19	0.2
		AN	0.05	0.04
		AT	0.25	0.24
	LA	AP	0.2	0.22
		AN	0.05	0.04
		AT	0.26	0.25
F	A	AP	0.2	0.21
		AN	0.05	0.05
		AT	0.25	0.28
	CC	AP	0.2	0.21
		AN	0.04	0.03
		AT	0.25	0.24
	FM	AP	0.2	0.2
		AN	0.06	0.05
		AT	0.26	0.26
	HC	AP	0.28	0.24
		AN	0.02	0.02
		AT	0.25	0.24
	LA	AP	0.2	0.2
		AN	0.04	0.04
		AT	0.26	0.25
N	A	AP	0.21	0.22
		AN	0.06	0.04
		AT	0.25	0.28
	CC	AP	0.23	0.22
		AN	0.04	0.02
		AT	0.25	0.24
	FM	AP	0.2	0.2
		AN	0.06	0.04
		AT	0.26	0.26
	HC	AP	0.21	0.2
		AN	0.04	0.03
		AT	0.25	0.24
	LA	AP	0.21	0.21
		AN	0.06	0.04
		AT	0.26	0.25

In the table above AG, F, N refers to Against, Favor and None stances respectively.

4 Challenged faced by NLTK

NLTK (Natural Language Toolkit) encounters several challenges when attempting to perform POS (Part-of-Speech) tagging on social media data. Firstly, social media data often contains informal language, such as slang and emoticons, that is difficult to parse using traditional POS tagging methods. Additionally, spelling errors and typos can impact the accuracy of the system by introducing non-standard spellings. The non-standard syntax used in social media, including incomplete sentences and unconventional punctuation, further complicates the task of POS tagging. Contextual ambiguity also poses a challenge, as meaning and tone can be heavily influenced by emojis, hashtags, and the specific platform being used. Lastly, multilingualism is another challenge for NLTK, as social media data often contains multiple languages, which can be difficult to handle using traditional POS tagging methods trained on a single language.

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

When trying to POS tag social media data, NLTK may make mistakes due to several factors. Firstly, social media language is often informal and may contain non-standard spellings, abbreviations, and emoticons that can be difficult to interpret. For example, "lol" could be interpreted as an abbreviation for "laughing out loud" or as a noun meaning

"league of legends". Secondly, social media data often includes a lot of slang, regional or cultural expressions, and new words that may not be in the NLTK's dictionary. This can lead to errors when the NLTK tries to tag a word that it does not recognize or misinterprets its meaning.

Another challenge with social media data is that it often lacks context, and the text can be very short, making it difficult to accurately identify the part of speech of a word. For example, in the sentence "That's sick!", "sick" could be an adjective meaning "cool" or a noun referring to an illness. Without context, the NLTK may not be able to accurately determine the correct part of speech.

6 What pre-processing steps seem to be "easier" for NLTK to accurately POS tag?

Pre-processing steps that seem to be "easier" for NLTK to accurately POS tag include standardizing spelling, correcting punctuation, and tokenizing text into individual words or sentences. These steps can help ensure that the text is in a format that is more easily recognized by the NLTK, improving the accuracy of the POS tagging.

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

Yes, adding polarity scores as feature would be helpful for classification when want to say whether it's positive or negative sentiment, we need additional data in order for it be classified specifically into other intentions such as happy, sad, angry, and etc.

8 Conclusion

I think polarity score is a valuable to measure sentiment although the calculation of that score is bound by the words, I believe my implementation isn't an exhaustive positive and negative words and can't produce accurate results but it does the job.