NLP PROJECT - TWITTER DATA CLASSIFICATION

Your progress

Announcements

NLP Module Twitter Data Classification Project

You are given a Twitter dataset in the form of a csv file that includes tweets classified by the human editors. In this project, please come up with a machine classifier that will replace the human editors. That means your classifier should correctly classify the individual twits as Business or not Business. Here is a guideline to help you to accomplish the project:

1. Decide on your programming language (for instance: Python vs R)

2. Analyze the Twitter data. Decide the feature and target fields. Explain why you include/not include any field.

3. Is your data set balanced/imbalanced? What are some of the predictable challenges ahead of you? Please list all of the alternatives/techniques that you can use in case of imbalanced data set.

4. What are some of the packages/libraries that you need to use? Are you planning to use NLTK, Spacey, re (regular expressions) … etc. Please make a note of all of them.

5. What kind of initial data cleaning steps are you proposing in terms of NLP text data? What is the size of your feature set in terms of vocabulary? How are you planning to reduce the dimension of your feature vector? Please explain in detail.

6. How are you going to tokenize, and vectorize your text data? What are some options in front of you? What are the advantages/disadvantages of BoW if that is what you are going to use?

7. Come up with a baseline classifier. What is your baseline classifier performance? Make a note of it. What performance metrics would you be using in this problem?

8. Revisit your NLP steps. How does it affect the performance to remove the stop words, or any tokens that do not carry any information such as numbers? How do stemming, lemmatizing, bi-grams, tri-grams affect the performance, or the size of your feature vector?

9. What classification algorithms are you planning to try in addition to your baseline classifier. Please come up with at least two candidates and explain why you choose them.

10. Apply cross validation and gridsearchCV techniques for hyper parameter tuning. Please explain how it helps you.

11. What is your best classifier? What is the performance of it? How do you compare it with your baseline classifier?

- Data Source:

https://drive.google.com/file/d/153Aw9XcU-2I7KL_UizEcbRsATnX698PJ/view?usp=share_link

https://drive.google.com/file/d/13Tx0YAS-H-lwzaAbqd0-sTw1ib68khe_/view?usp=share_link

NLP Project - Twitter Data Classification Project Submit