Machine Learning Engineer Nanodegree Capstone Proposal

Yosef hesham nwear
october 4, 2019

Classifying SMS or emails Messages as either Spam or not Spam (Spam Filtering - Text Classification)

Domain Background

- Natural Language Processing (NLP) is a sub-field of Artificial Intelligence that is focused on enabling computers to understand and process human languages, to get computers closer to a human-level understanding of language. Computers don't yet have the same intuitive understanding of natural language that humans do. They can't really understand what the language is really trying to say. In a nutshell, a computer can't read between the lines.

- The goal of this field is to get computers to perform useful tasks involving human language, tasks like enabling (human-machine communication , improving human-human communication, or simply doing useful processing of text or speech)

That being said, recent advances in Machine Learning (ML) have enabled computers to do quite a lot of useful things with natural language! Deep Learning has enabled us to write programs to perform things like language translation, semantic understanding, and text summarization. All of these things add real-world value, making it easy for you to understand and perform computations on large blocks of text without the manual effort.

The paper I upload it with submit file it called : (Text_Classification_Using_Machine_Learning) . Problem Statement The problem is that many people receive spam (emails,sms message , …) These type of messages aims to collect users personal information ,so in This project I have data set contains 5574 classified message and aim To use NLP and ML to predict that new message is spam or not spam . Datasets and Inputs SMS Spam Dataset, The size of the dataset is 5574 sample row data and contains two columns. First column contains classification of message (ham(legitimate) or spam) and Second one is message content . SMS Spam Collection has a total of 4,827 SMS legitimate messages (86.6%) and a total of 747 (13.4%) spam messages. I will take 25% of data as Testing set and 75% as Training set .

Solution Statement This is a classification problem, to solve it, we build a model to predict if message is (ham or spam) , K Nearest Neighbors , Decision Tree , Logistic Regression , Naive Bayes , SVM Linear . Benchmark Model We will use benchmark model to compare to our final result to make sure it works; we will use Naive bias classifier , as it is easy and has an initial result without tuning or using complex methods. Evaluation Metrics First thing to do is deal with imbalanced classes “or data in the dataset” by using Confusion matrix or F1 score or ROC and AUROC

Project Design

I will start with the classical workflow

1- Exploring and Load data set
2- Preprocessing : Tokenization: It is the process of converting each word in the document into different token. It splits a sentence into number of pieces called tokens. Stop-words Removal: Stop words like connecters or prepositions are removed. For example:" is, am, are, the, and, but, for, if etc." Stemming: It is task of converting different words into the root Word.
3- Regular Expressions : use regular expressions to replace email addresses, URLs, phone numbers, other numbers .

4- Feature engineering: the words in each text message will be our features. For this purpose, it will be necessary to tokenize each word We will use the 1500 most common words as features .
5- model: split data and use algorithms to make prediction, I will use : "sklearn.model_selection.train_test_split()" To split the data. And about prediction i'll use classification models. 6- Evaluation: nltk.classify.accuracy(model, testing) or F-Score Measure

References

1- https://en.wikipedia.org/wiki/Natural_language_processing
2- http://www.nlp.com/what-is-nlp/
3-https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63
4-https://www.nltk.org/
5- https://www.nltk.org/book/ch06.html
6- https://en.wikipedia.org/wiki/Mean_squared_error
7- https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

8-https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28