Trivial NLP Solutions for Classifying Social and Transaction Interactions

The Problem Statement - Separate Social Interactions from Transaction Interactions.

Given Resources - Raw Unlabelled Uncleaned Dataset.

Approach -

- The first step in solving any problem with machine learning is to clean and format the data in a useable format. In fact, data cleaning and preprocessing is one of , if not the most important facets contributing towards a projects success.
- The dataset given was cleaned properly using first generic data cleaning procedures followed by text cleaning to reduce the features, redundant information.
- After extensive dataset cleaning and analysis by the "Text Processing.ipynb" , the cleaned dataset is trialled by the Spam Filtering - unlabelled trial.ipynb file.
- The final tf-idf based classical binary classification and the unsupervised clustering are both done finally in the Spam Filtering - unlabelled trial.ipynb.
- Please Bare in mind, these solutions are the most trivial and fastest solutions possible. Any solution beyond these will require more time but can be done.
- The output ipython notebook can be viewed as a html -
    - Processing Social Messages.html

Future Approaches -

- The current solution vectorizes the input using trivial tf-idf vectorizing.
- This can be easily improved by using word embeddings such as word2vec and doc2vec, spacy and so on but that would require additional computational power. (more capable machines)
- The data can be cleaned even more by knowledgeable data scientists into a processable format.

- Furthermore sampling a portion of the data to feed into a ML pipeline by labelling the sampled dataset thus changing our problem to a supervised problem would make the task much simpler and easier.

Effort -

- The project so far has taken me roughly 9 hours total from start to finish.