Kevin Brockman

CS 723 Project Proposal

Problem / Goal

Spam emails cause significant waste of both computational resources to transfer and store them, as well as the waste of individual’s time to sort through them. Spam filters which filter out spam (undesirable) from ham (desirable) emails are a long-standing answer to this. I would like to contribute to the ongoing arms race between spammers and spam filters to keep inboxes clean and save people time and headaches.

Methods

1. Tools
   1. Work will be performed in python, primarily with the NLTK, Scikit, and Keras (using tensorflow) libraries, as well as word embeddings from GloVe (developed by Stanford NLP Group)
2. Training Data
   1. Training data will be sourced from multiple datasets of spam, ham, and mixed spam / ham emails
3. Normalizing Data
   1. As multiple datasets will be combined for training a common format will need to be applied, they will be narrowed down to three data fields: Subject Line, Body Text, and Classification (spam or ham)
4. Model Generation
   1. At least 3 models will be created, optimized, and compared
      1. Convolutional Neural Network[1]
         1. Data will be tokenized and tagged for input, GloVe word embedding data will be used t o form the input layer of the network
      2. Recurrent Neural Network[1]
         1. Similar preparation to Convolutional
      3. Support Vector Classification[2]
         1. Data will be vectorized based on individual words scored by their TF-IDF

Evaluation

Each model will be rated using standard statistical measures (Accuracy, Precision, Recall, and F1 Score) as well as training time and evaluation time.

References

1. Yoav Goldberg. 2016. A primer on neural network models for natural language processing. J. Artif. Int. Res. 57, 1 (September 2016), 345–420.
2. Huynh Tan Hoi and Le Vu Truong. 2020. Application of Support Vector Machine Methods in Classification of Customer Communications Data. In Proceedings of the 2020 International Conference on Computer Communication and Information Systems (CCCIS 2020). Association for Computing Machinery, New York, NY, USA, 15–18. <https://doi.org/10.1145/3418994.3419002>