Developing Spam Detection and Prevention Schemes using Natural Language Processing

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Parv Ahuja
Spring, 2020.

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Yuan Tian, Department of Computer Science

Developing Spam Detection and Prevention Schemes using Natural Language Processing

Every day, hundreds of thousands of individuals sit in front of their computers and write fake reviews for products based upon the tasks they receive from the underground market for the purposes of manipulating product ranking. Spam callers, review writers, and other miscreants communicate using underground crowdsourcing, where they utilize instant messaging (IM) to coordinate their attacks (Undisclosed, 2019). The threat intelligence provided by this IM communication is invaluable to understanding and mitigating fraud, but hard to systematically gather and analyse. In comes Aubrey, the first autonomous chatbox that actively collects intelligence through communication with real-world miscreants. The chatbox, Aubrey, poses as a miscreant looking for tasks, and utilizes the underground communication pipelines to engage in conversation and extract relevant intelligence. To achieve autonomous conversation, Aubrey is modeled as a finite state machine where states represent different possible stages in the conversation, and state transitions are performed based on the responses of the miscreant. In order to facilitate this automata, Aubrey must be able to deduce meaning and intent from miscreant messages and then utilize that information to generate a valid response (Undisclosed, 2019).

Natural Language Processing is the field of study regarding natural human language and computer science. More specifically, NLP involves how to program computers and software to analyse and process linguistic data. In the context of Aubrey, NLP is used to ensure smooth sentence transitions and a seamless conversation. First, NLP is used to translate a miscreant message into some recognisable state for Aubrey. To do this, the message is broken up into different words, each of which becomes mapped to a higher-dimensional vector that represents the word's relationship to the entire sentence. Next, the vectors are processed and translated to specific intents based on a previously defined database of relationships. Finally, a classifier uses the derived intent of the message and produces the state probabilities, essentially mapping the message to a state within Aubrey's finite state machine (Undisclosed, 2019). Recall, that states within the FSM were designed to represent different possible stages within a conversation with a miscreant, and this process essentially maps a miscreant message to a particular stage in the conversation that Aubrey can recognise. After this, Aubrey can follow the state transitions and produce a response that reflects understanding of the miscreant message.

From this prior research, it becomes evident that Natural Language Processing has a significant role in autonomous conversation . Even outside of the scope of Aubrey and e-commerce fraud, NLP is a key factor in the future of spam detection and prevention. Autonomous conversation with spam callers, chatting with e-commerce miscreants, and detecting spam emails are all potential schemes that would require NLP. The remainder of the research assumes the use of NLP in spam detection, and refocuses on the specific NLP techniques that could enable these detection and prevention schemes.

In recent years, there has been tremendous progress made in the field of natural language processing, including more powerful models and increased relational data. One of the most basic frameworks within NLP is a transformer, which is a mechanism to gather relevant context of a given word, and then encode the context into a multi-dimensional rich vector (Rizvi, 2019).

Before BERT, a majority of transformers in application were long-short-term-memory based models, like the one applied for Aubrey. LSTM transformers operate much in the way that humans do when processing sentences, in that encoding and decoding for context happens left-to-right in a linear fashion (Rizvi, 2019). Although this model was powerful in its own manner, it ran into issues when context was hidden within a sentence. For example, in the two sentences below, a LSTM transformer would have trouble deducing the different meanings of the word "bank" looking only left to right (bolded words provide context).

1. I need to go to the bank **to deposit some money**
2. **We went to the river** bank

BERT, an acronym for Bidirectional Encoder Representations from Transformers, tackles this problem by jointly conditioning on both the left and right context. The tremendous framework was developed by Google AI just this year, in 2019, and has already made groundbreaking advancements in NLP research. BERT was able to use a single model to achieve state of the art results on 11 individual NLP tasks, using only an unlabelled dataset. Through personal investigation, in a simple classification problem, I was able to utilise a pre-trained BERT model to achieve higher accuracy than my previous fully trained LSTM model. In other words, without any finetuning or training, BERT was able to achieve better results than the LSTM model that had been fully trained on the dataset.

For the remainder of my research, I plan to delve into BERT and develop a model that could surpass results achieved within the spam detection field. In Aubrey, for example, a BERT model could replace the current LSTM model to reduce error in deducing meaning from miscreant sentences, overall enabling Aubrey to collect more threat intelligence. Specifically, I propose implementing BERT within Aubrey and examining the results and more research into spam-specific NLP tasks that could be better solved using BERT.