Udacity Capstone Project

Machine Learning Nanodegree, 2020

1. Domain Background

Twitter is a micro-blogging platform that allows users to broadcast real-time messages called tweets. Twitter has 330 million monthly active users tweeting 500 million tweets daily.

This growth has caused many companies to create an official twitter account to engage with their customers and stakeholders. And some companies even automate this process using bots.

One significant academic study estimated that up to 15% of **Twitter** users were automated **bot** accounts. The prevalence of **Twitter bots**, coupled with the ability of some **bots** to give seemingly human responses, has enabled these non-human accounts to garner widespread influence.

Most bots produce their contents by listening to specific hashtags and re-broadcasting to a broader range of audiences that are interested in them. The problem with bots who source their content on the twitter platform is they can be tricked to broadcast tweets contrary to their subject matter.

The Internet has come a long way; nevertheless, unwanted content called spam is still a huge problem today. And Twitter is not exempted.

When people hear of spam, it is common to think of it as unsolicited emails one receives from an unknown person; interestingly, spam does not target inboxes only. Spams target everything these days, even social media feed.

Analyzing the content of tweets, retweets or mentions, is very important for companies who want to give control to bots to automate and engage customers on their behalf.

2. The Problem Statement

"Twitter is a wild place," said a friend.

After creating a @hubofml, I had no choice but to agree with him. Indeed twitter is wild. @hubofml is a Twitter page that broadcasts interesting articles and happenings on machine learning to people who are interested in machine learning. I built it by creating a bot that retweet tweets with machine learning and data science hashtags. The bot got a whooping 100 followers in 24hrs.

Things went South after the people of Twitter realized there is a new parrot in town that repeat tweets if you append **#machinelearning** to it.

People would do silly things like tweeting "*RT @hubofml Everything is F**cked! #machinelearning,*" and the bot would retweet it.

Can text categorization using machine learning help detect tweets related to the subject of interest and mark everything else as spam?

Text categorization is the process of automatically assigning one or more predefined categories to a text document.

The aim of this project is to categorize tweets into two categories, spam and non-spam using a supervised learning approach with an LSTM network. The spam determinant being whether a given tweet is related to machine learning or non-related. The project will analyse a dataset of CSV file containing over 10,000 tweets sourced directly from the Twitter platform through the process described in the section below.

3. Data Source

Twitter provides stream APIs one can use to retrieve tweets of certain categories (hashtags). I'll use this API to obtain datasets needed for this project. Two datasets will be used: 1) Machine learning tweets, and 2) Normal tweets. To collect machine learning tweets, the hashtag **#machinelearning**, **#ai**, **#datascience**, **#computervision** are used. That is, tweets with **#machinelearning**, **#ai**, **#datascience**, **#computervision** are retrieved from twitter API and tagged "not spam". Normal tweets are retrieved based on randomly selected keywords from the WordNet lexicon and tagged "spam".

Raw Input Data fields:

- ***Tweet****: The field contains the tweet in its raw form.*
- ***Spam****: An indication that a tweet is not related to machine learning. The indicator can take a yes or a no. "Yes" means spam and "no" means not a spam.*

	tweet	spam
0	RT @DrThomasPaul: Bill Gates and his foundation are bio-terrorists.\n\nBill is a computer science freak who is in the business of creating ch…	yes
1	RT @PoliticsPollss: Denmark has announced any companies registered in tax havens will not be eligible for state aid programmes during the p…	yes
2	RT @aajtak: उद्धव को MLC मनोनीत करने के लिए राज्यपाल खामोशी अख्तियार किए हुए हैं #Maharashtra #politics \nhttps://t.co/0rF9RLVPYE	yes
3	Nice one sis	yes
4	💞🍀👍♥ @missiontrustme Marianne Rauch – intelligent, geistreich, schön. ♥ #ff #7lines ..und wie geht es weiter? #Spannung und #Crime in "Träum süß stirb schnell", Langeweile war gestern! #ebook #Leseempfehlung #amazon #Bücher https://t.co/potjykXVYb	yes
5	@dwnews #Shame#Enough #Politics@critical situation.\nY don't you write a book and make it documented.\nLet's people's also feel your pain for present govt. Time has gone for Artificial Awards, It's #New#India.	yes
6	HK jails Indian businessman for breaking quarantine laws - https://t.co/WNsKV3KJ6V \nGet your news featured use #IndiaPostUSA \n#Coronavirus #COVID19 #Epidemic #HongKong #India #IndianAmerican #IndianOrigin #NRI #Politics #WHO	yes

Two kinds of preprocessing will be performed on tweet datasets:

1. tokenization
2. usernames, URLs, and emojis removal.

Usernames, URLs and hashtags are removed from tweets as they do not provide any information about the status of the words and might be noise for the classification process

4. Solution

The solution will be the prediction of whether a certain tweet is related to machine learning or non-related (that is, spam) in the test dataset. First, I'll preprocess the data. Then, perform vocabulary extraction and select features such as word length, word count distribution, character count. Each input will be encoded or converted to vectors.

The vectorised input will be trained using a Long Short Term Memory Network (LSTM). LSTM is a good option for this problem due to its ability to maintain an internal memory state and gates to control the flow of information inside each LSTM unit. The LSTM network will have 5000 embed size, 50 hidden layers and a fully connected layer with 2 outputs.

I'll do some tuning to both hyper parameters and the network architecture during implementation. Therefore, the architecture described in this proposal is not final.

5. Benchmark Model

For this problem, the benchmark model will be Support Vector Machine (SVM). The F1 score of the trained model will be compared with the F1 score of an SVM model on kaggle solving a similar problem. Since, I'm using tweets which are relatively shorter than emails, I do not expect my trained model and the SVM model to have a close F1 score.

6. Evaluation Metrics

Prediction results are evaluated on the cross-entropy loss between the predicted values and the ground truth. The ground truth is the set of tweets that have been classified as not spam.

Since I'll be dealing with thousands of tweets here, human errors can not be entirely overruled. As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate and may include incorrect labels.

7. Project Design

- The first step is data gathering; Since the dataset is not available, I'll create a python script to download datasets of tweets using the Twitter API. This could take up to 48hrs to gather the dataset required for training, and then saved to a CSV file.
- The CSV datasets will be converted to data frames using panda.
- Once the data is available in data frames, I'll do a quick calculation of the number of records in the datasets. That is the total number of tweets marked as spam, and the total number of non-spam tweets.
- Visualization provides better insight into data; I'll generate a graphical representation of the dataset, preferably a histogram representation of spam and non-spam tweets.
- Tweets usually contain irrelevant data like stopwords, punctuations, "RT," handles, and emojis. To make the datasets fit for training, I'll pre-process the data and encode categorical variable “spam."

```
def clean_tweet(tweet):
        # Remove usernames, "RT" and "#"
        # I might retain hashtags, they are very useful. Most Image tweets don't have descriptions, only hashtags.
        # Remove links in tweets
        # Remove punctuations and stopwords
        # Convert tweet to lower case
        # Tweets are usually full of emojis. We need to remove them.

        return tweet.strip()
```

- I'll generate a vocabulary list from the datasets and encode each tweet (sentence) to vectors.
- At this test, I'll split the prep-processed datasets into train and test datasets. Test datasets will be used to test how the model generalizes on unseen data.
- I'll train the model, tune hyperparameters, calculate scores until I achieve a satisfactory level of accuracy.
- Finally, I'll deploy the model to production and continue to monitor it.

8. References

Transfer Learning in NLP for Tweet Stance Classification

A Deep Learning System for Twitter Sentiment Classification

Detection and Classification of Social Media-Based Extremist Affiliations Using Sentiment Analysis Techniques

Cyberthreat Detection from Twitter Using Deep Neura Networks

Twitter by the Numbers: Stats, Demographics & Fun Facts