Machine Learning Engineer Nanodegree

Capstone Proposal

Brian Long

September 24th, 2019

Proposal

Domain Background

Twitter is a micro-blogging platform where users can post short messages referred to as “tweets”. Tweets used to be limited to 140 characters but this limit was increased to 280 characters in 2017. This enforced conciseness made twitter a popular platform for people to quickly voice their opinion about a topic without needing to publish a multi-page article. Millions of people write multiple tweets every day to express their thoughts or opinions, both formal and informal. Twitter has become a place for friends to share pictures of their lunch, as well as a place for professionals to discuss emerging practices and technologies.

The thing that interests me the most about Twitter as a social media platform is the frequency and openness with which people post. In many ways, a person’s twitter feed is like a journal of their day to day life. There is a wealth of information that seems almost purposefully structured for data science. For example, the use of hashtags to label tweets allows them to be categorized by topic without breaking up the platform into multiple smaller communities that are defined by that topic. Using the public Twitter API, developers can access data about tweets that can then be aggregated and used to drive business decisions across a variety of fields.

Twitter stands out as a tool for sentiment analysis because the text differs heavily from the usual data that comes from sources such as movie reviews, product reviews, or news articles. These other sources are typically large bodies of text with proper grammar, whereas tweets are have a strict character limit which keeps them short while

also having a very informal tone that doesn't follow grammatical convention. Twitter has been studied as a source of sentiment analysis in the following papers:

Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1*(12)*.

https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., Sentiment Analysis of Twitter Data. Columbia.

http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf

Problem Statement

The information I would most like to extract from these tweets is the general sentiment in terms of whether it has a positive or negative tone. Sentiment analysis is a popular problem in the field of Natural Language Processing. The intricacies of language make it difficult to algorithmically extract meaning from a given body of text. While true sentiment contains much more complexity than can be expressed by the broad definitions of having a positive or negative point of view, this is still valuable information as it can be used to quantify the general public's feelings towards a company, product, or idea.

Datasets and Inputs

I plan to use the Sentiment140 dataset to build my Twitter sentiment analysis model. This dataset is a collection of 1.6 million tweets that were gathered from the Twitter API and are labeled as positive or negative. Since this dataset was intended for use in sentiment analysis, it was set up such that there is a perfect class balance with exactly 50% of the training data being positive and 50% being negative. The tweets were labeled based on their inclusion of emoticons, with variations of smiling emoticons being labeled positive, and variations of frowning emoticons being labeled negative. The emoticons were also removed from the dataset after the sentiment labels were added. This dataset was put together using distant supervised learning as a way to showcase

its viability in labeling data this way. The data was tested against Naive Bayes, Maximum Entropy, and Support Vector Machine models, all of which scored over 80% accuracy. The Sentiment140 dataset is available at the following link: http://help.sentiment140.com/for-students

Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1*(12)*.

https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Solution Statement

I will be attempting to create a model that can determine whether a tweet expresses a positive or negative sentiment. This model could eventually be used to build a tool that uses the Twitter API to have the sentiment evaluated for new tweets. This could further be used to show trends in how the general public views a given topic and leveraged to make business decisions.

Benchmark Model

I will be using a Naive Bayes model with a Bag of Words approach as my benchmark model. Naive Bayes is often used as a benchmark in Natural Language Processing problems since the implementation is simple but fairly effective. This model will then be tested for accuracy in terms of how many tweets were correctly identified as positive or negative.

Evaluation Metrics

The main metric I will be using to measure the benchmark and solution models will be accuracy. I choose to focus on accuracy over precision or recall because mislabelling a tweet as positive or negative has the same impact. In this instance we are not specifically trying to identify all positive tweets or avoid incorrectly labelling a negative tweet as positive, but to correctly identify as many tweets as possible, regardless of whether they are positive or negative. Since the target class distribution is perfectly

balanced between positive and negative tweets, using accuracy should not be affected by simply favoring one choice over the other. I will also be looking at F1-score as a secondary metric so that precision and recall are not entirely discounted, although my main focus will be on improving accuracy.

Project Design

My workflow will begin by downloading and importing the dataset into a dataframe using pandas. I will then split the data into training, validation, and testing sets. Next, I will perform some preprocessing on the text data. Tweets are considerably more informal than text that is usually used for sentiment analysis and very frequently has acronyms, shorthand versions of words to save space, and misspelled or exaggerated words that have many repeating letters for emphasis. The text will need to be tokenized to break up the sentence into individual words or n-grams. I will also attempt to transform this text into something more uniform to reduce the dimensionality of the data. I will try doing things such as removing hashtags and usernames referenced in tweets by a preceding “@” symbol and commonly used filler words such as “a” and “the”. Finally, word vectorization will transform the text into numerical values to get it ready to be used in training a model.

I will build a Bag of Words Naive Bayes model to use as a benchmark, fit it to the data, and test its performance against the validation set. I expect this to perform fairly well, since the Naive Bayes model used to measure the performance of the Sentiment140 distant supervised labeling method had a pretty high accuracy.

My approach creating a strong sentiment analysis model will be to use a deep neural network. I will try out a few architectures including Recurrent Neural Networks and Long Short Term Memory networks because I have read about their success in Natural Language Processing tasks. I will iteratively be training the model on the training set, measuring its performance against the validation set, and tweaking the architecture and hyperparameters. The performance of the model will be measured in terms of accuracy and F1-score.

When I feel that I have tuned the model to the best of my ability, I will measure the performance of both the benchmark model and the final model against the testing dataset to see how they match up. I will also be plotting the performance in multiple ways such as in an ROC curve so that it is easy to visualize the improvement of the model over the benchmark.