

Machine Learning Engineer Nanodegree

Capstone Proposal

Patrick O'Sullivan
June 20th, 2019

Proposal

Business Public Sentiment

Domain Background

Businesses are in every aspect of our lives from the moment we are born (and earlier) to the moment we die (and after). They have huge control over us and can manipulate us in many ways (and often do). A number of business have built platforms that allow their customers rate their one-off experiences with a business but what about the overall sentiment of a business.

Understanding the overall sentiment of a business may help us make a more informed decision about which business we want to use for a given service and hence encourage businesses to be more conscience and pro-active about their public sentiment. It would also allow other businesses to decide which businesses they wish to partner with or provide/receive services from.

Data Source:

- <http://help.sentiment140.com/home>

There are several research papers in this area:

- [https://www.researchgate.net/publication/301649777_Analyzing Scientific Papers Based on Sentiment Analysis First Draft](https://www.researchgate.net/publication/301649777_Analyzing_Scientific_Papers_Based_on_Sentiment_Analysis_First_Draft)
- [https://www.academia.edu/39483256/Business Sentiment Analysis. Concept and Method for Perceived Anticipated Effort Identification](https://www.academia.edu/39483256/Business_Sentiment_Analysis_Concept_and_Method_for_Perceived_Anticipated_Effort_Identification)

There are several news articles on this topic highlighting its importance:

- <https://www.forbes.com/sites/jiawertz/2018/11/30/why-sentiment-analysis-could-be-your-best-kept-marketing-secret/#91f358e2bbec>
- <https://www.businessinsider.com/negative-social-media-sentiment-hurts-sales-2013-6?r=US&IR=T>
- <https://www.theguardian.com/news/datablog/2013/jul/15/reputation-management-business-swallow-bitter-pill>
- <https://www.business2community.com/branding/measuring-corporate-sentiment-02091306>

My personal motivation for working on sentiment is to understand the importance of how what we say and do effects how people perceive us. I'm starting with businesses but "us" could be a team or a person also. Once we understand how we (person, team or business) are perceived we could provide feedback on how to improve that perception.

Problem Statement

The main objective of the project will be to use Machine Learning to decide the sentiment of text. When give a string of text we want to be able to say whether the sentiment of the text is considered positive or negative. If we can build a model that can accurately say whether a string of text is positive or negative, we can then take live data feeds for various companies (from twitter or other sources) and track the public sentiment over time.

Datasets and Inputs

For this project we will use a data-set called [Sentiment140](#). The data-set is split into both a training set and testing set. The training set contains 1600000 tweets. The test set contains 498 tweets. I may split the training set further as the test set is very small.

The tweets are is a csv file with the following fields:

- id
- date
- query
- user
- text

The tweets are classified as

- 0 = negative
- 2 = neutral
- 4 = positive

Half (800000) of the tweets are classified as negative and half (800000) of the tweets are classified as positive. There are no neutral tweets in this data-set. This means the data-set is balanced so neither class should dominate during training.

I can use the twitter [Standard search API](#) to get real data as input.

Solution Statement

The proposed solution to this problem is to use the Natural Language Toolkit (NLTK) and Machine Learning technique that have proved to be successful in the classification of sentiment.

First, we will read the data-set (see Data-set section above) and do any pre-processing that is needed to make sure the data is as clean as possible. Then we will split the training set and test set and build and

compile our model, then evaluate and validate the accuracy of our model and finally get a prediction and accuracy score.

Benchmark Model

There are several projects on Kaggle in this area which I can use to compare my approach.

- [twitter-sentiment-analysis](#)
- [python-nltk-sentiment-analysis](#)

As Naïve Bayes is a quick and easy way to predict classes I will use it as my benchmark model. At the very least my final solution should be as good if not better than this benchmark model.

Evaluation Metrics

The evaluation metric for this project is an accuracy score.

Project Design

There are several steps needed to complete this project:

- Exploration: Understand the data been used in this project. In this phase I'll download the data-set, have a look at it and see what I need to do to prepare it for preprocessing. For example, I've already hit the excel limit on opening the training file as it contains more than a million rows. I'll use PowerShell to split the file. "\$i=0; Get-Content <PATH>\training.1600000.processed.noemoticon.csv -ReadCount 250000 | %{\$i++; \$_ | Out-File <PATH>\splitfile_$\\$i$.csv}"
- Preprocessing: In the phase I will look to clean the data to make it easier to work with. Will need to remove some features often seen in text like tweets, e.g. URLs, usernames, etc.
- Split: The data set comes with a training set and a test set, but the test set seems very small. May need to split the training set further.
- Models: I will start with a Naïve Bayes model as it is a good model for text-based classification and I will use it as a benchmark and see if I can create an even better model. The data is labelled and Support Vector Machines (SVMs) are another good model for data classification so I will use SVMs for my second model.
- Model Training: Start training the model. Try different setting to improve the model. Make sure the model isn't under-fitting or over-fitting.
- Evaluation: Look at the results the model is producing, accuracy score, confusion matrix and use that evaluation to try to improve the model.

References

1. Sentiment140: <http://help.sentiment140.com/for-students/>

2. Twitter Search API:

<https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets>

3. Kaggle: <https://www.kaggle.com>