Machine Learning Engineer Nanodegree

Capstone Proposal

Simon Phua
August 24, 2017

Proposal

Domain Background

The advent of mobile technologies and social media has given rise to new opportunities for companies to engage consumers. One challenge faced by companies is effectively managing the swathes of feedback they receive. Of particular concern is negative feedback, that if not addressed quickly, could go viral and damage a company's reputation. My project proposes to investigate how sentiment analysis can be employed by companies to better manage its social media interactions with consumers.

The following research materials were consulted to better understand existing efforts in the field of sentiment analysis.

- *'SentiBench – a benchmark comparison of state-of-the-practice sentiment analysis methods'*, by Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves and Fabrício Benevenuto (https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0085-1)
- *'Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank'*, by Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts (https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf)
- *'NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets',* by Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu (http://aclweb.org/anthology//S/S13/S13-2053.pdf)
- *'Detection and Scoring of Internet Slands for Sentiment Analysis Using SentiWordNet'*, by Fazal Masud Kundi, Shakeel Ahmad, Aurangzeb Khan and Muhammad Zubair Asghar (https://www.researchgate.net/publication/283318703_Detection_and_Scoring_of_Internet_Slangs_for_Sentiment_Analysis_Using_SentiWordNet)
- *'Context-Aware Spelling Corrector for Sentiment Analysis'*, by Fazal Masud Kundi, Aurangzeb Khan, Muhammad Zubair Asghar, Shakeel Ahmah (https://www.researchgate.net/publication/284344945_Context-Aware_Spelling_Corrector_for_Sentiment_Analysis)
- *'Lexicon-Based Sentiment Analysis in the Social Web',* by Fazal Masud Kundi, Aurangzeb Khan, Muhammad Zubair Asghar, Shakeel Ahmah

(https://www.researchgate.net/publication/283318830_Lexicon-Based_Sentiment_Analysis_in_the_Social_Web)

Problem Statement

Social media managers sift through scores of tweets and posts one by one. For popular brands, comments could number in the hundreds and thousands on a daily basis. On some platforms, users leave a rating or a score with their feedback, e.g. Yelp, AirBnB, and Amazon, so social media managers can quickly identify the negative comments. On other platforms, e.g. Twitter, LinkedIn and Facebook, no such rating mechanism exists. For this latter group, in particular, social media managers need a way to prioritise the negative comments that deserve their attention first in order to quickly limit any reputational damage and commence service recovery.

I propose building a sentiment analysis model trained on the dataset's 'Score' and 'Text' inputs (other features will be examined, but these will be the minimum two key features used). The model will then be used to predict if any new given review is a 'negative' review.

Datasets and Inputs

The Amazon Fine Food Reviews dataset (obtained from Kaggle at https://www.kaggle.com/snap/amazon-fine-food-reviews) contains 568,454 reviews for food products on Amazon up to October 2012. The dataset consists of:

- Score – rating between 1 and 5 given to the product by the user;
- Text – text of the review by the user;
- Summary – title of the review by the user;
- UserId – unique identifier of the user;
- ProductId – unique identifier of the product reviewed;
- ProfileName – username of the user leaving the review;
- HelpfulnessNumerator – number of users who found the review helpful;
- Helpfulness Denominator – number of users who indicated if they found the review helpful or unhelpful; and
- Time – timestamp of the review

This dataset was selected for the following reasons:

- Large number of unique reviews (including a large variety of unique reviewers, words used in the reviews and products reviewed) makes this dataset suitable for a variety of machine learning techniques, and suggests it could generalise well;
- The 'Score' input provides an effective way to categorise the reviews by positive and negative reviews;
- The 'Text' input provides the basis for building a sentiment analysis model around the words used in the reviews; and
- Reviews were generated relatively recently which minimizes bias due to evolving language/lingo over time when applied to present day examples.

Given that the dataset is based on food product reviews, the immediate potential application is for sentiment analysis related to food products.

Solution Statement

In building a sentiment analysis model, I will utilise classification algorithms. Given the large feature set (i.e. number of different words in the dataset), it is likely that a Naïve Bayes classifier will be most appropriate. I will also attempt building a neural net to predict negative sentiments, and compare both approaches.

Benchmark Model

A social media manager combing through comments one at a time has a random chance of identifying a negative comment – the specific probability depending on the distribution of 'positive' and negative comments. I will compare the model's ability to correctly identify negative comments against random chance based on the dataset's distribution.

Check out other benchmark models

Evaluation Metrics

I will examine metrics including:

- Accuracy, given by (correct predictions) / (total predictions); and
- Recall, given by (correctly predicted negative reviews) / (total number of negative reviews).

In addition to the model's predictive power, I will also examine the additional time taken to employ a machine learning model for sentiment analysis, vis-à-vis a social media manager randomly reviewing comments. The improvement in predictive power must exceed the additional time required to implement the machine learning model in order to justify its use.

Project Design

Data exploration

- Examine the number of reviews, and those with missing data;
- Examine the distribution of scores;
- Examine the number of unique products reviewed and the range in number of reviews available for each unique product;
- Examine the number of unique users and the range in number of reviews left by each unique user; and
- Examine the HelfulnessNumerator and HelpfulnessDenominator distributions.

Data prepossessing

- Categorise reviews by ‘positive’ and ‘negative’ based on scores.

Implementation

- Implement a naïve bayes classifier with bag-of-words; and/or
- Implement a neural network classifier with one-hot-encoding.

Potential Refinement(s)

- Iterate with stop_words;
- Iterate with limited number of most-frequently-occuring words;
- Iterate with different model architectures;
- Iterate between including or excluding ‘Summary’ from dataset;
- Iterate between classifying a score of 3 as ‘positive’ or ‘negative’, or omitting it from the dataset; and
- Iterate between utilising only reviews with a minimum ‘HelpfulnessNumerator’ value.

Assessment

- Evaluate metrics;
- Assess and discuss usefulness of model; and
- Recommend further study and potential improvements.