CS221: Project Proposal

Nimisha Tandon
nimisha@stanford.edu
Shaila Balaraddi
shailaab@stanford.edu
Naman Muley
ngmuley@stanford.edu

October 24, 2019

Introduction

Identifying information on the internet that could be potentially damaging to a brand is an important function of marketing and PR for a brand. News articles about the entire industry, a product ecosystem or the brand itself, could be relevant to a brand and impact it positively or negatively. A system that provides an impact score by analyzing news and other textual documents can provide useful leads for a brand to get ahead of a PR cycle. For example, articles in Mexican national newspapers about use of guns can be relevant to NRA as a brand in North America and will be helpful for it to shape it's policies.

For the final project we will, first, classify an article into a news group. Once classified we will analyze the article to produce an impact score for a set of pre-defined parameters. These parameters give insight into how these articles will affect a brand. The impact score can be used to stack rank articles, when presented to the brand.

Project Scope

This project can be broken down into two stages: a) classification and b) impact analysis. Both these operations will deploy ML techniques. Both these stages also require training effort. Each of the following stages will take a set of training data to train the algorithm. They will then take as input

Classification into News Groups

We will start by exploring to build a text classifier. To build this we will explore multiple techniques like RNN using LSTM, CNN and ML. Based on our findings we plan to either chose 1 classification model or create an ensemble of the 3 models and use that to classify our article. To name a few we would be exploring the below(and more) features to study each ones impact on the classification model. We will compare the output of the model looking at the Precision , Recall and F1 scores on the validation set:

- Bag of words
- Word2Vec using Glove vectors
- TF-IDF
- Part of Speech Tagging
- Averaging word vectors

Impact Analysis

Once the article is categorized, providing an impact score based on the classification is the next step. Creating a meaningful impact score is key and part of the challenge. Impact to a brand can be classified based on certain parameters that are important to a brand e.g. market share, sentiment of it's products, sales, research and development etc. Some of these parameters are specific to a category and some others are generally relevant to the brand.

Following are some parameters for a category like *geography*:

- *country*: Is the geographical location talked about in this article one that is highly relevant to the brand?
- *population*: Is this article relevant to a high population number or low?

Following are some general parameters that could be relevant to the brand:

- *Features*: are people talking about a particular aspect of your product or service?
- *Sentiment*: is the article talking about the brand in positive or negative light
- *Sales*: is the article talking price, sales or any other monetary factors relevant to the brand

Similar to classification, use of RNN or CNN or a combination of ML techniques can be used to build the impact score.

Oracle and Baseline

The Oracle will be a manual classification of articles, and impact score based on the content and its relevance to the subject. Following is a table of the results of the articles that have any impact to the subject of "guns". We are interested in articles that reflect a true impact to guns, in the context of gun control, gun legislation, NRA and 2nd amaendment The following list of articles are categorized and evaluated.

Article Name Category Impact score(scale of 1 to 10) 53297 Politics 10 53294 Politics 10 176846 Politics 0 crime$_{r}eportPolice0$

As per the Oracle, the scores are dependent on the context of the domain in question. Hence even though the article "crime$_{r}eport" haswordslikeshootingandhomicide, itdoesnthaveanyimpactonthedomainof" guns"$.

For the baseline we implemented a SGD ML classifier, where tfidf and vector count was used as features. With this we achieved an accuracy score of 82.4 percent.

The output of the classifier is what we fed into the impact score analyser where we tried to rank the documents in terms of their impact score.

For calculating the impact score we gave the function a set of words for a given category of the articles. Using these tokens we identified the frequency of these tokens in each of the articles.

For our baseline we only tried to count the frequency of terms in the article from the set of tokens against which the impact score was being calculated. However the numbers were not great.

In order to improve the same we want to explore using the following features :

a. Word2Vec : Using this we would include in the word frequency not just exact words which match but also words that are close in meaning.
b. Wordvector averaging
c. BERT and ELMO - Since these embedding take into account not just the word but also the context of the input this would help us improve how we calculate the impact score.

We would also like to improve our classification model for which we want to look at using Neural Network (RNN using LSTM CNN) using Keras and Tensorflow.

Challenges

Following are some key challenges facing the project:

- Build a system that understands impact score
- Build a classifier system that can compliment the impact score better
- Obtaining relevant data sets to train the system

Infrastructure

Infrastructure for the project consists of building instrumentation for the model to run

Application

The model will accept input from user on the category and will display the outputs as a stack rank of articles on the internet along with their potential impact scores. This application will interact with the model. This application will also be responsible for pulling articles from the internet in obtaining test data if required.

Compute

Google cloud credits will be used to run this model and perform training of the model.

Dataset

We will be using the "20 Newsgroup" data set. The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of my knowledge, it was originally collected by Ken Lang, probably for his Newsweeder: Learning to filter netnews paper, though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. This comes builtin with scikit learn. And can be used directly from there.
We are also exploring to use the bbc data set which consists of 2225 documents from the BBC news website corresponding to stories in five topical areas business, entertainment, politics, sport, tech.

Infrastructure

Project Prompt

Define the input-output behavior of the system and the scope of the project. What is your evaluation metric for success? Collect some preliminary data, and give concrete examples of inputs and outputs. Implement a baseline and an oracle and discuss the gap. What are the challenges? Which topics (e.g., search, MDPs, etc.) might be able to address those challenges (at a high-level, since we haven't covered any techniques in detail at this point)? Search the Internet for similar projects and mention the related work. You should basically have all the infrastructure (e.g., building a simulator, cleaning data) completed to do something interesting by now.