

CS221: Project Proposal

Nimisha Tandon

nimisha@stanford.edu

Shaila Balaraddi

shailaab@stanford.edu

Naman Muley

ngmuley@stanford.edu

October 24, 2019

1 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 its policies.

For the final project we will, first, classify a set of articles into a news groups. 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. The system will take as input some domain data which represents the brand. It will also present to the user certain categories or news groups that the brand is interested in.

1.1 Classification into News Groups

We will start by building a text classifier. This classifier runs on a recent set of articles provided to it by the application, either siphoned from the internet or gathered from the client. It will then classify these articles into pre-defined categories.

To build this we will explore multiple techniques like RNN using LSTM, CNN and ML. Based on our findings we plan to either chose one classification model or create an ensemble of the 3 models and use that to classify our article.

1.2 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 its 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

2 Oracle and Baseline

We undertook basic exercise to identify upper and lower bounds for our system. These bounds are on the impact score presented to the articles.

2.1 Oracle

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 amendment The following list of articles are categorized and evaluated.

Article Name Category Impact score(scale of 1 to 10) 53297 Gun 10 53294 Gun 10 176846 Gun 0

As per the Oracle, the scores are dependent on the context of the domain in question. Hence even though the article "crime-report" has words like shooting and homicide, it doesn't have any impact on the domain of "guns".

2.2 Baseline

For the baseline we implemented the following:

1. *Classifier*: SGD ML classifier, where TF-IDF 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 analyzer 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.
2. *Impact Analysis*: Given the domain information in the form of some vocabulary, were the set of tokens against which the impact score was being calculated. However the numbers were not great. We talk about these challenges in the following section

2.3 Oracle and Baseline Stats

We ran a baseline program for a brand like *guns*.

We used the *talk.politics.guns* category already provided in the Newsgroup20 dataset.

We took as domain input the following vocabulary: "guns", "shooting", "victim", "mass", "kill", "murder", "weapon", "gun", "nra", "handgun", "assault". The baseline performance can be noted down in the following way:

- The classifier has 82% efficiency in classifying the category
- The Impact Analysis program resulted into the following Top 10 impact scoring articles. The article indices are indices into the test dataset of articles:

7274	0.05
3613	0.03
2668	0.03
548	0.03
1665	0.02
6763	0.02
6896	0.02
3781	0.01
6620	0.01
2135	0.01

3 Challenges

Following are some key challenges facing the project:

- *Impact Score*: The impact score definition is purposefully kept a bit fluid. The idea of the impact score is to use the category to determine a unique set of characteristics that determine *impactful* texts in that category. In order to improve the impact score we want to explore using the following transformers:

1. Word2Vec : Using this we would include in the word frequency not just exact words which match but also words that are close in meaning.
2. Wordvector averaging
3. 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.

- *Category Classifier*: Build a classifier system that can compliment the impact score better. 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.

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

- *Dataset*: Obtaining the right data set for the training the classifier and impact score system is crucial. Getting the right data to train the classifier on the categories we want to support is a challenge. We have identified the BBC dataset but would like to explore more.

Training data for the impact score system is also another challenge. The impact score system will use the category output of the classifier and must also consider the domain in its training.

4 Infrastructure

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

4.1 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.

4.2 Compute

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

4.3 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.

Social Impact

A brand is a financial entity that is interested in knowing potentially impacting content on the internet. Substitute a brand with a social cause like *election*, *poverty*, *amazon rainforest* etc and this system will be able to provide a stack rank of articles that are important to such a cause.

Also, in the context of fake news, one of the features that the impact analysis system could build is *credibility* where the credibility of a news source is a feature that the impact scoring system takes into account for building an impact score. In that context, this system can be integrated with an RSS feeder plugin which can then recommend potentially impacting articles in a stack rank to people who subscribe to causes like the ones mentioned above.