CS221: Project Proposal

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

October 23, 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 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 would try to classify an article into a news group and once classified we would try to extract Company/Organization names and analyse the article to produce an impact score for a set of pre-defined parameters, which would give an insigt into how the article may impact the Company/Organization or Brand in question.

Project Scope

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 ensamble 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:
1. Bag of words
2. Word2Vec using Glove vectors
3. TFIDF
4. Part of Speech Tagging
5. Averaging word vectors
6. etc

Once the article is categorized we will explore to identify the following impact parameters and try to provide a score for each of them: Features: are people talking about a particular aspect of your product or service? Wishes/Desires: is the article talking about expressing particular desires?
Price: is the article talking about how the price of a product is being perceived. Competitors: how does your brand compare to competitors? What are your strengths and weaknesses?

Dataset

We will be using the "20 Newsgoup" 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.

Oracle and Baseline

———

Next Steps

———

Identify the attributes which build up the impact score

———

Identify the attributes which build up the impact score

———

Improve Model for classification and Impact score

———

Challenges

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.