CS221: Project Progress Report

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

November 15, 2019

1 Introduction

We are tackling the problem of identifying articles that could be potentially dangerous to a brand. Our approach has two steps. First, we take in a set of articles and classify them into categories. In the final application, this could be done by pulling articles from the internet daily. Once these categories are identified, in the second step, we create an impact score for the brand in question. The articles with top impact scores can be provided to the client as having high potential for impact to the brand.

This problem has a rich application of various algorithms for natural language processing. Since proposing this project, we have explored a few algorithms to create a better sense of what an impact score can be. We looked into utilizing unsupervised learning algorithms like GloVe and supervised algorithm implementations like Naive Bayes. These have given us better impact scores. Further more, it feels like we could perform some more fine tuning to come up with better models to bring a richer understanding of the *impact score*

2 Methodology

We extend our two step approach which we presented in the proposal. We have spent time exploring different algorithms for both these stages. We seem to have settled on to a methodology, which we explain in this section.

2.1 Classification

In making progress for the project, we decided to give more weight to the impact score analysis section, since that's the more novel component of the solution approach. Our classification currently still uses a simple Stochastic Gradient Descent classifier. We found it's accuracy to be nearly 82 percent and thought we can improve upon that in the later parts of the project.

We still did look at using the Naive Bayes algorithm to perform classification, since that's what literature reported is it's primary use. It performed slightly better at classification than SGD but not by much.

2.2 Impact Score

Impact score needs to be a number that represents how relevant or impactful will a particular article be to the brand. Hence, we extended our previous approach of having a vocabulary of words that are relevant to the brand and applying a combination GloVe analysis, TF-IDF and Naive Bayes to come up with a score of how closely does the article's content relate to the dictionary relevant to the brand.

2.2.1 Input Brand Vocabulary

It was possible to decipher a set of words that are deemed contextually more dominant and important for a category like *guns* by various methodologies including TF-IDF itself. But we believe finding relevancy to a brand should allow some input bias from the brand. For example, a brand may way to find articles relevant to a subset of their very specific products and those may not be popularly present in the testing dataset.

We allow a list of words to act as our vocabulary for finding articles with highest potential impact. For example, for a brand like NRA, we take the following set of words as a vocabulary:

["guns", "shooting", "weapon", "nra", "handgun", "assault", "rifle", "america", "firearm"]

2.2.2 Similar Words using GloVe learning

A measure of how relevant an article is to a brand is to know how many of semantically or linguistically similar words occur in that article. For each of the words in the vocabulary, we get a list of *similar words* using the GloVe learning algorithm [1]. We then use a simple TF algorithm to understand how commonly do these similar words occur in an article.

Algorithm 1: Extend Vocabulary with Similar Words using Glove Model

$Vocabulary \leftarrow [\text{"}guns\text{"}, \text{"}rifle\text{"}, \text{"}weapon\text{"}, \text{"}nra\text{"}, \text{"}handgun\text{"}, \text{"}firearm\text{"}]$;
$model \leftarrow train\_glove($*glove.6B.300d.w2vformat.tx*$)$;
GetGloveWords $(Vocabulary, model)$
  $extended\_vocab \leftarrow Vocabulary$;
  foreach *word* $w \in Vocabulary$ **do**
    $similar_words \leftarrow model.get_similar_words($word=w, $topN = 10)$;
    $extended_vocab.extend($similar_words$)$;
  end
  return $extended\_vocab$;

Once the vocabulary is extended, we use this extended vocabulary to identify articles that have the most occurrences of these words. In the Proposal, we had used a similar algorithm to calculate our baseline but with the vocabulary as the raw input set of words. Following is a figure that shows how the expanded vocabulary behaves for a brand like NRA and input vocabulary = *["guns", "weapons", "nra"]*.

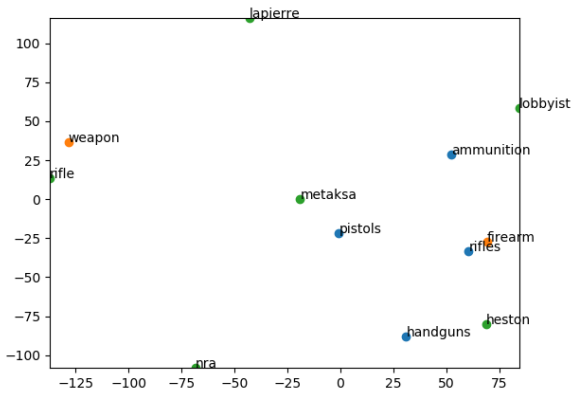

Figure 1: Similar Words obtained from GloVe Learning for *NRA*

We use this extended vocabulary to come up with an impact score by simply counting frequencies of these words. Our algorithm that performs this to come up with an impact score for each article in the

category is given below. The score dictionary contains the impact score calculated for each article in the set of articles.

Algorithm 2: Calculate Impact Score for each article from a list of articles, given extended vocabulary

CalculateVocabImpactScore (V, D)
inputs : A list of words that form ExtendedVocabulary V; a list of articles D
output: A dictionary representing impact scores for all articles in D
$score \leftarrow \emptyset$;
$RangeD \leftarrow range(0, len(D))$;
foreach *index* $i \in RangeD$ **do**
score[i] = 0 ;
freq = [] ;
/* iterate over all words in i_{th} article to count frequency */
foreach *word* $w \in D[i]$ **do**
freq[word] += 1 ;
end
/* iterate over all words in V to build a score for article i */
foreach *word* $w \in V$ **do**
score[i] += freq[w] ;
end
end
score.Normalize() ;
return $score$;

We have detailed our results of the above score dictionary in the Results section.

2.2.3 Relevant words using TF-IDF

Term Frequency-Inverse Document Frequency commonly known as TF-IDF can be used to find the frequency of vocabulary words in a document to perform sentiment analysis or extract relevancy from a document.

This algorihm has 2 algorithms working together.

1. Term Frequency

$$tf(t,d) == \frac{noofoccurrencesoftermindoc}{totalnumberofallwordsindocument}$$

Before using this algorithm, we first do some preprocessing of the doc as follows:

- Remove all stop words.(ex: in, the,are it)
- Convert all words to lowercase.

Now using the term frequency alone will end up giving us words that are not unique (for ex: repeated words, should not add up to the relevance of the document)

2. Inverse document frequency

This gives us the uniqueness of a word:

$$idf(t,d) == log(\frac{nooftimesthetermappears}{noofdocumentscontainingtheword})$$

The final step is to multiply the two together to get the TFIDF score which would give us the impact score.

3 Implementation

Talk about how the models are trained, what data sets did we use for these models. how are the data sets relevant. Future work may involve using and creating better data sets

Currently, we have a small python application that performs both sets of actions.

4 Preliminary Results

Show similar words, maybe a fancy representation of similar words and important words as found by GloVe and TF-IDF respectively We ran the above algorithms and ran comparisons on various impact score results on the documents.

5 Future Work

This impact score, we feel, still does not capture all the richness we can provide using natural language processing algorithms. We intend to improve in the following areas:

1. *Better Trained Models*: Currently the GloVe model we use is trained from the wikipedia dataset. We intend to see if training it on a more relevant dataset creates a better set of similar words that we can use to calculate impact scores.

2. *More Factors in Calculating Impact Score*: Currently, we're using fairly simple ways of calculating impact based on contextual similarity between words. Adding more sophisticated ways to identify important articles is desirable.

Table 1: Impact Score chart

Top 5 relevant docs	ImpactScore
Subject: Re: Gun Talk – State legislative update. One might well ask if CA gun owners have given up on the NRA/CRPA. The national NRA doesn't march in and get things passed. They provide a convenient label for local activities/activists.	0.20
Subject: Re: The Dayton Gun "Buy Back" (Re: Boston Gun Buy Back) Is there something similar pro-gun people can do ? For example, pay 100 dollars to anyone who lawfully protects their life with a firearm ? Sounds a bit tacky, but hey, whatever works. How about a gun buy-back/charity? Get some sponsors to fund the purchase of used firearms, have a gunsmith check them over, and give or sell them at a low price to poor persons wishing to own firearms. ;-)	0.19
Our Gun Club has organized several of these (we just finished teaching another one last night, in fact) and they have been very well received. We get a lot of people who are novices interested in guns. We even get a few who are anti-gun, but feel they should know something about "gun safety" since members of their family keep guns at home.	0.18
Well Joe, I suggest that you talk to the Center to Prevent Handgun Violence or the Centers for Disease Control. If YOU look carefully you will see that YOU greatly underestimate the presence of guns in the lives of youths. The CPHV reports that 135,000 youth bring GUNS to school DAILY and that 400,000 bring GUNS to school at least once a year. The CDC estimates that 1 out 0f 25 high school students carried a gun to school at least once in 1990. The CDC also says that 1.2 million elementary-aged, latch-key children (kids who come home from school to an empty house), have access to guns in their home. California schools reported a 200student gun confiscations between 1986 and 1990, and a 401988 and 1990. Florida reported a 61schools between 1986/87 amd 1987/88. These are the "statistics".	0.17
Jim, I'm just saying how it is. I'm not saying if that is a good thing or not. From the police who I have talked with who run some of these gun buyback programs, I get the impression that they really think they are having an impact on the community. When I ask them if they have an evaluatory component to the program, they say "well no..." So, in answer to your question, no, false hope is not the intent. I think the intent is to show folks that police are attempting to do something to curb interpersonal gun violence whether its effective or not. Look, if you can't measure the impact of these programs using some sort of pre-test and post-test evaluation, what is the point? It must be symbolic in nature. The police are essentially saying "look, if you have a gun lying around and you don't want it, we'll give you 50 dollars for it...because we care about the community". If you, I and Joe could think of a way to measure the effectiveness or ineffectiveness of these programs we could become rich and famous.	0.15