Impact Lens : Classification and Stack Ranking Information Relevant to a Domain

Naman Muley | Nimisha Tandon | Shaila Balaraddi

Stanford University

Overview

Motivation: Identifying news that could impact a brand is extremely useful for large brands or analytics divisions to get ahead of the PR cycle and formulate an early response. We attempt to build a Machine learning classifier for multi-class classification, which categorizes the articles using Stochastic Gradient Descent and then builds an impact score for these articles and presents a stack rank.

This system will also be useful to NGOs and Governments working to build a brand and fight misinformation on a national scale. Impact Score, as a concept can include credibility and factfullness.

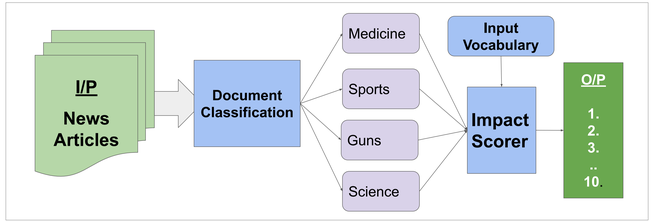
Goal: Build a system that

1. **Identifies articles most relevant to a brand and**
2. **Give them an impact score for how potentially impactful the article could be for a brand**

Solution Approach

We follow a two step approach:

1. **Take a set of articles and classify them into categories**. From these categories, we ask a brand to choose a subset of categories that are most relevant to them. We take as input a set of vocabulary that is important for the brand. Based on this vocabulary,
2. **Create an impact score for the brand in question**. The articles with top 10 impact scores can be provided to the client as having high potential for impact to the brand.



Methodology

Classification:

- Stochastic Gradient Descent with TF-IDF transformers

Classifier/ Hyper-params	Loss	Regularization	Max Iterations	Alpha	F1
SGD	Hinge	l2	20	1.00E-03	0.81088067
SGD	squared hinge	l2	20	1.00E-03	0.826460415
SGD	modified_huber	l2	50	1.00E-03	0.826561936

Key experiments were to try the SGD with and without tfidf and we observed there was a tremendous increase in f1 score on including Tf-Idf.

- Multinomial Naïve Bayes Classifier - We compared the accuracy score from SGD classifier (with Hinge Loss) to MultiNomial Bayes, and noticed that SGD got us slightly better classification accuracy.
- Pre-processing done before classification

Impact Score

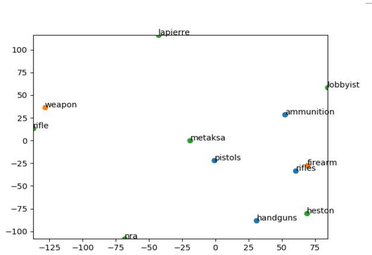
- Vocabulary frequency where we counted the terms in the articles.
- Enhanced Vocabulary using Glove word embeddings to find similar words and then count those in the articles.Using word embedding improved our results as embeddings were able to look beyond the defined vocab and look for words which were of similar significance.
- Impact score could also be calculated using the “term frequency” calculation of the TFIDF method on each document. We tested this and confirmed very similar results as the Vocab frequency. In addition the results brought previously unseen relevant documents to the fore-front.
- Opinion Mining

Dataset and Results

DataSets:

- Classification: Newsgroup 20
- GloVe: Wikipedia + Gigaword5

Extended vocabulary with **GloVe learning**



New York Times Articles

New York Times archive of December 2018 had 6200 articles. Score for #3967 (“Nigerian Military kills unarmed protesters”) go above #2888 (“Mass shootings in Parkland, FL”) with GloVe.

Article Index (out of 6200)	Score
2888	0.060
158	0.057
182	0.057
1492	0.043
1502	0.043
3967	0.040
3834	0.032
188	0.031
632	0.031
5224	0.025

Article Index (out of 6200)	Score (with GloVe)
3967	0.071
2888	0.044
158	0.044
182	0.044
5237	0.039
5255	0.039
1492	0.038
1502	0.038
4846	0.031
3834	0.028

Discussion

- We had success categorizing and generating impact score.
- Classification - F1 score was highest for modified_huber loss (among several variations tried)
- A modified_huber punishes more on outliers
- With one by one gradient descent, it learns faster to avoid miss-classification
- Ran grid search to optimize/improve F1 score for classification, however accuracy did not improve
- Use the intersection of the words from the Vocab and those found as opinion words to enhance the impact score. However, considering our vocab was user input and our opinion mining techniques were rudimentary we did not observe significant improvements in the impact score.
- It was also very difficult to pull sentiment from news articles as the words and their contexts were not assertive in opinions, or sometimes ambiguous and sarcastic.

Presentation Video & Reference

- Video Link:https://drive.google.com/open?id=18SmhS4tPAFJeX-hz8EvH_uF1_EKAcyLV
- Thanks to our mentor Jon Kotker for his guidance
- Helpful links used to build the feature extraction and classification algorithms:
- *https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html*
- *https://medium.com/syncedreview/applying-multinomial-naive-bayes-to-nlp-problems-a-practical-explanation-4f5271768ebf*