Text Classification of News Headlines

Ruoting Liang, Jessica Moloney, Mehvish Saleem, Prabhjot Singh, Karanjit Singh Tiwana
{ruotingl, moloney, mehvishs, psa83, ktiwana} @sfu.ca

Motivation

Digitization has changed the way we process and analyze information. There is an exponential increase in information from different sources, especially information that comes in the form of text. The need to classify text has grown dramatically in many industries. As a result, text classification is a core problem in many applications.

The goal of this project is to provide Natural Language Processing and Machine Learning approaches to classify news headlines into a pre-defined category.

News sites may want a way to automatically tag and cater content as a way to improve the browsing experience of their viewers..

Data

The dataset contains around 125,000 news headlines spanning the years 2013- 2018 from HuffPost. The features included in the dataset are the following:

Variable	Description
Author	Journalist who wrote the published the article
Category	The category of the news article **(Target Variable)**
Date	Date the article was written
Headline	The headline text of the article
Article Link	Link to the original news article
Short Description	Short description of the news article.

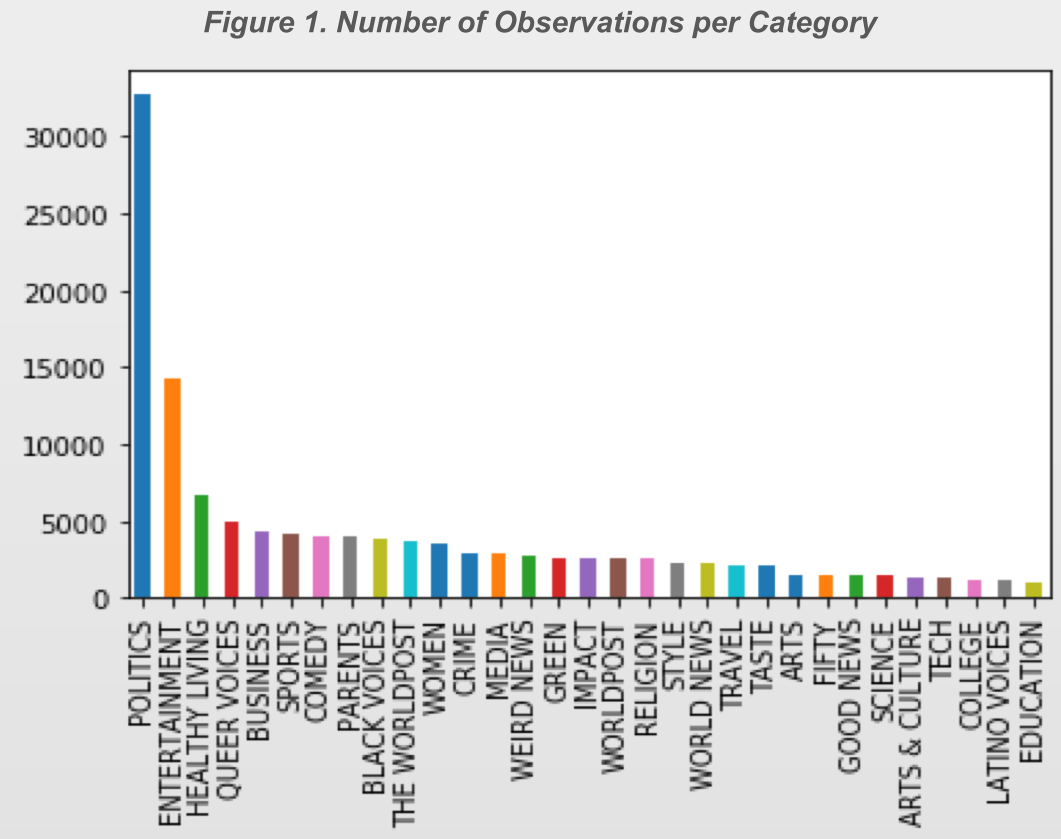
Table 1. Variables within the data set

The main features to be used here is the Headline variable as well as the Short Description to predict the news category.

Methodology

A variety of approaches were considered. In order to first do classification with text data, we need the words in some sort of numerical form in order to start to understand the relationships between the observations. For this, we try multiple word and sentence embedding techniques. We then apply a learning algorithm on the vectorized data focusing on methods that use neural networks.

Figure 1. Number of Observations per Category



The following methods have been considered for **text representation**:

- Bag of words:
  - TF-IDF
- Word embedding
  - Doc2Vec
  - Pre-Trained Glove Embeddings
- Neural Network Embedding Layer

The following methods have been used to **classify the data**:

- Linear SVC
- Long Short Term Memory Model (LSTM)
- LSTM + CNN Model

The above work representation methods and classification models have been combined Into the following **experimental model**s

1. TF-IDF with Linear SVC
2. Doc2Vec with Linear SVC
3. Neural Network Embedding Layer with LSTM Model
4. Neural Network Embedding Layer with LSTM + CNN Model
5. Pre-trained Glove Embeddings with LSTM + CNN Model

Exploratory

An investigation into the features within the dataset was carried out. One idea for potential analysis was to identify the entities most talked about per category. An initial look was done into the dataset to reveal the most mentioned entities in the entire data set. These are in displayed below in Figure 2.

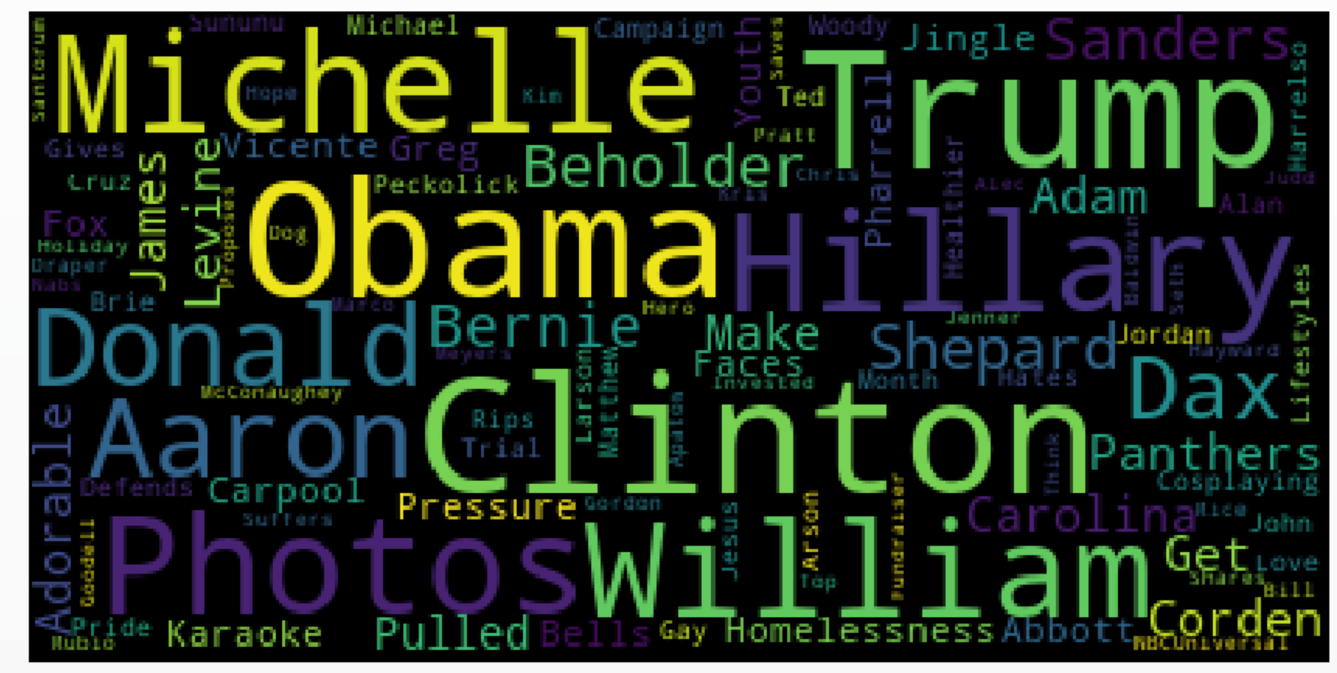


Figure 2. Most Mentioned Entities in the Headlines

We also found that there were over 20 news categories within the data set and that the classes were fairly imbalanced. This can be seen in Figure 1.

Results

The main optimization metric looked at is the F1-Score, though both precision and recall are considered as well. Below, we present the F1-Score of the models considered on the hold-out set in our experiments. Results of the experiments are as follows:

Model	F1-Score (Test Set)
TF-IDF + Linear SVC	0.60
Doc2Vec + LinearSVC	0.68
LSTM + Neural Embeddings	0.58
LSTM-CNN + Neural Embeddings	0.58
LSTM-CNN + Pre-Trained Embeddings	0.58

Table 2. Average F1-Score Results on Test Set

The F1-scores are averaged across all categories and their results are above. We've seen that we do relatively well in categories for which we have more data and not as well in categories with less observations.

Conclusion

This is a multi-class classification problem with over 30 classes. Some classes are clearly similar to others (e.g. Arts and Arts & Culture) which may make it hard to properly classify. Additionally, there exists less of these types of articles within the data set. Due to this, we do see that we perform well on categories for which we have a lot of data and not as well on categories that don't.

Moving forward, we could reduce the number of classes and put more sections together. This would give more examples of each type and create more distinct classes with less overlap.

HuffPost articles are by no-means neutral. There are many opinion articles and headlines which aim to promote sensationalism. This is however a dive into the classification of text into certain categories based on content. Moving forward, we expect that the classifier may not perform well on new headlines from other sources with a different writing style.

References

Le, Quoc, and Tomas Mikolov. *Distributed Representations of Sentences and Documents.* Mountain View, CA, 2014 cs.stanford.edu/~quocle/paragraph_vector.pdf