Utilizing a Random Forest Classifier to Categorize Headlines

Irené Masabarakiza
Computer Science & Engineering
Texas A&M University
College Station, Texas
imasabo2k18@tamu.edu

Rohan Lingala
Computer Science & Engineering
Texas A&M University
College Station, Texas
rlingala@tamu.edu

I. Abstract

News and media are integral parts of the modern world because they are the lens to the global, foreign, and local issues that are in our society. They also show us other discussions of topics of interest, whether that be editorials or opinion pieces. Furthermore, the way we interpret and perceive media is just as important as we need to be able to identify what is subjective and objective, and furthermore what people are writing about in a quick and digestible way. Armed with this information, our challenge was to see if we could train a machine learning model to categorize an article into one of 42 different topics if only given its headline. The idea is that we can create a classifier that can identify the category of a headline, and if they cannot, we need to understand why. This project is a deeper exploration into natural language processing while trying to apply what we have learned in a new context that will challenge us and offers insights into the categorization of news headlines using machine learning and aims at enhancing news aggregation and recommendation systems.

II. Introduction

Using a data set comprised of news article information, we use Text-Hero to pre-process our headlines, which are then fed into a Word2Vec text vectorizer and then sent into a Random Forest Classifier with 100 estimators. We then use a confusion matrix and a classification report to showcase our results. Our training set consists of a JSON file of 210,000 data points, each point of which contains a category and a headline most importantly, but contains other auxiliary information as well. There are 42 different categories that the headlines can be assigned.

III. Related Works

The paper Application Research of Text classification based on random forest algorithm by Yanxiong et al, was one of the first pieces of literature we read in preparation for this assignment, it gave us the confidence and inspiration to attempt our own modeling. This paper reviews the random forest algorithm, and a modified version of it to better suit text based classification.

IV. Methodology

A. The Two Main Problems

When we first began to tackle this problem, we had many questions in regard to how we should try to create this model. Our two big issues were what information we give the model, and what kind of word-processing model we should we use. To answer the former, we decided to only keep the category and headline and to answer the latter, we were debating between two text vectorizers: TF-IDF, and Word2Vec.

B. The First Problem

Inside our data set, we had many different parameters we could have factored into our training, some of which were somewhat irrelevant such as date, website link, and author, however one seemed quite relevant as a potential extra parameter, that being the short description section of the data. We originally had high hopes for using this data until we realized that the descriptions given were basically the same information as the headline, meaning it would be redundant to have both. Furthermore, we only wanted to have the control variable be the headline because that is what we have decided to promise in terms of what we were going to offer, so we stuck with the headline analysis solely for this reason. This does not mean we didn't try and experiment with different ideas, instead of modifying the data set, we tinkered with various models.

C. The Second Problem

TF-IDF and Word2Vec were the two vectorizers we settled on because we were already familiar with the workings of TF-IDF because of prior assignments, and Word2Vec was very simple and intuitive in terms of understanding the details of how it processes words. We eventually decided on Word2Vec because of how well it handles semantics, we needed a tool that could analyze the sentiment of a short string of words, in this case, a headline.

Word2Vec is a method for creating high-dimensional word embeddings, which are numerical representations of words. The method makes use of a neural network to discover the connections among words in a corpus of text. A series of vectors, one for each distinct word in the corpus, are produced by the neural network after it processes a sizable amount

of text input. These vectors are produced by teaching the network to anticipate the words that will follow a particular input word in a sentence. The vectors are designed to capture the associations between words, so they will have similar vectors if they are connected or have similar contexts. The resulting vectors can be utilized for applications involving natural language processing.

V. Results and Conclusions

We were able to only achieve a weighted accuracy of 43 percent with Random Forest, however with utilizing different models, we were able to realize a ceiling of about 53 percent accuracy. There is nuance regarding the overall accuracy, most of which implicates our data set.

A. The Caveats

One of the caveats of our model was that the data set categories were split in somewhat of an arbitrary way, for example, there are categories such as ARTS, ARTS & CULTURE, and CULTURE & ARTS that are all considered different by the person who labeled the data, however one could argue that all three of these categories could be considered within the same bounds. This type of issue was something we did not predict, and to correct this one could re-label the data set to congregate more like-minded topics together. An observed issue is that when the category becomes too specific and not general enough, the model gets confused and picks the wrong category. Here are all observed examples of this potential confusion: HEALTHY LIVING and HOME & LIVING, PARENTING and PARENTS, STYLE and STYLE & BEAUTY, WORLD NEWS, and WORLDPOST and U.S NEWS and GOOD NEWS, MEDIA and ENTERTAINMENT, ENVIRONMENT and GREEN, and finally FOOD & DRINK and TASTE.

Another caveat with our data set was how vague certain topics meant, as well as how the author decided what headlines belong in what categories. For example, it is not particularly clear why something would be considered WEIRD NEWS, FIFTY, or WORLDPOST. There are also cases where something is categorized as STYLE however that same headline could potentially also be categorized into STYLE and BEAUTY, these distinctions are subjective in nature and are based on the whims of the author. The massive potential overlap between labels hurts our data analysis because we are not able to properly use natural language processing to determine what categories a headline would go in, as the categories are defined somewhat arbitrarily and vaguely.

B. Observations and Conclusions

As seen on the right, we have the statistics from the model given each category. It is notable that categories with pronounced enough distinction such as DIVORCE and ENVIRONMENT have at least a 0.50 precision value, with a recall of at least 0.40. If we were to only look at only the top 15 categories in regards to sample size, we observe a massive spike in the recall, showing that with more sample size and more generalized data, we could have gotten a model that was more permissive in categorizing something with the correct label.

Wanting to achieve a higher accuracy score, we decided to explore other models and see how they compare with our baseline of Word2Vec and Random Forest.

Our first attempt was creating a Naive Bayes Classifier with a Count Vectorizer to process our words, which gave us an accuracy of 53 percent, an improvement to the 43 percent value given by the Random Forest. The Bayes theorem states that the likelihood of an event occurring is equal to the probability of the occurrence given the previous knowledge about the event. A naive Bayes classifier is a straightforward probabilistic classifier based on this theory. A naive Bayes classifier is trained on a dataset of labeled instances in the classification context, where each example consists of a set of characteristics and a label. The classifier learns the probability of each label given various feature value combinations using the training data. The classifier employs the learned probabilities to forecast the most likely label given a fresh set of features. Assuming that the traits are independent of one another gives the word its ”naive” quality. Naive Bayes classifiers are commonly used in text classification tasks like spam detection. The features are the individual words in the text and the label is the classification category. This is why we choose this classifier as another tool we could use to categorize the headlines. This model showed improvement regarding accuracy in relation to the Random Forest Classifier but has little to no room for further fine-tuning to increase this accuracy further.

Our second attempt was utilizing a sequential model with word embedding and a Recursive Neural Network, we had looked into the word embedding earlier for processing Word2Vec outputs and believed that it would work well with a Recursive Neural Network, which is generally very good for handling text-based problems. With this model, we saw an upward trend in accuracy from one epoch to the next. However, due to the lack of sufficient computing power, running several epochs with this model took a while. With this attempt, we also hit a wall at 52 percent accuracy.

category	precision	recall	f1-score	support
BLACK VOICES	0.42	0.12	0.18	908
BUSINESS	0.32	0.17	0.22	1199
COMEDY	0.5	0.19	0.27	1077
ENTERTAINMENT	0.37	0.7	0.49	3466
HEALTHY LIVING	0.16	0.03	0.06	1322
FOOD & DRINK	0.46	0.52	0.49	1242
HOME & LIVING	0.54	0.42	0.47	843
PARENTING	0.35	0.44	0.39	1863
PARENTS	0.32	0.1	0.15	815
POLITICS	0.52	0.87	0.65	7050
QUEER VOICES	0.52	0.41	0.46	1297
SPORTS	0.46	0.28	0.35	1016
STYLE & BEAUTY	0.61	0.67	0.64	2053
TRAVEL	0.42	0.6	0.59	1935
WELLNESS	0.37	0.69	0.48	3484
MACRO AVG	0.4226666667	0.414	0.3926666667	1971.333333
WEIGHTED AVG	0.436211701	0.5672850862	0.4738549205	3249.418194

Fig. 1. Our classification report output of the top 15 categories. (Random Forest)

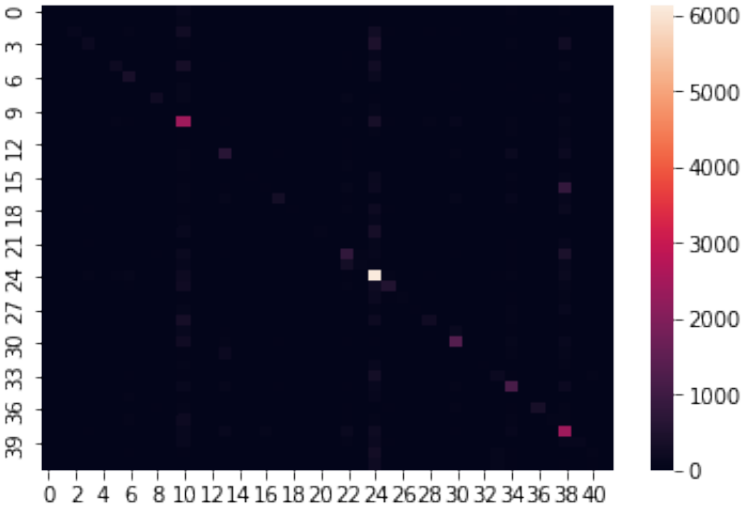

Fig. 2. Our confusion matrix output. (Random Forest)

	precision	recall	f1-score	support
ARTS	0.38	0.03	0.05	335
ARTS & CULTURE	0.33	0.00	0.01	248
BLACK VOICES	0.42	0.12	0.18	908
BUSINESS	0.32	0.17	0.22	1199
COLLEGE	0.44	0.03	0.06	237
COMEDY	0.50	0.19	0.27	1077
CRIME	0.39	0.54	0.45	706
CULTURE & ARTS	0.84	0.09	0.17	225
DIVORCE	0.59	0.41	0.48	665
EDUCATION	0.33	0.02	0.05	204
ENTERTAINMENT	0.37	0.70	0.49	3466
ENVIRONMENT	0.83	0.06	0.12	313
FIFTY	0.33	0.00	0.01	279
FOOD & DRINK	0.46	0.52	0.49	1242
GOOD NEWS	0.27	0.06	0.10	282
GREEN	0.29	0.10	0.15	537
HEALTHY LIVING	0.16	0.03	0.06	1322
HOME & LIVING	0.54	0.42	0.47	843
IMPACT	0.21	0.05	0.08	687
LATINO VOICES	0.00	0.00	0.00	210
MEDIA	0.51	0.11	0.17	599
MONEY	0.57	0.06	0.11	350
PARENTING	0.35	0.44	0.39	1863
PARENTS	0.32	0.10	0.15	815
POLITICS	0.52	0.87	0.65	7050
QUEER VOICES	0.52	0.41	0.46	1297
RELIGION	0.61	0.12	0.20	500
SCIENCE	0.50	0.12	0.19	448
SPORTS	0.46	0.28	0.35	1016
STYLE	0.39	0.03	0.06	445
STYLE & BEAUTY	0.61	0.67	0.64	2053
TASTE	0.35	0.02	0.03	435
TECH	0.59	0.11	0.19	410
THE WORLDPOST	0.36	0.26	0.30	729
TRAVEL	0.42	0.60	0.49	1935
U.S. NEWS	0.25	0.01	0.01	270
WEDDINGS	0.69	0.54	0.61	720
WEIRD NEWS	0.27	0.10	0.15	581
WELLNESS	0.37	0.69	0.48	3484
WOMEN	0.44	0.13	0.20	758
WORLD NEWS	0.27	0.10	0.15	624
WORLDPOST	0.31	0.04	0.07	539
accuracy			0.45	41906
macro avg	0.42	0.22	0.24	41906
weighted avg	0.43	0.45	0.39	41906

Fig. 3. Our classification report output of all categories. (Random Forest)

	precision	recall	f1-score	support
ARTS	0.35	0.02	0.04	320
ARTS & CULTURE	0.00	0.00	0.00	301
BLACK VOICES	0.40	0.11	0.17	908
BUSINESS	0.29	0.17	0.21	1174
COLLEGE	0.43	0.02	0.05	243
COMEDY	0.55	0.17	0.25	1122
CRIME	0.37	0.55	0.45	673
CULTURE & ARTS	0.72	0.06	0.10	236
DIVORCE	0.62	0.39	0.48	671
EDUCATION	0.14	0.01	0.02	202
ENTERTAINMENT	0.37	0.70	0.49	3469
ENVIRONMENT	0.78	0.05	0.09	291
FIFTY	0.12	0.00	0.01	284
FOOD & DRINK	0.44	0.52	0.48	1180
GOOD NEWS	0.29	0.05	0.09	294
GREEN	0.33	0.10	0.15	505
HEALTHY LIVING	0.18	0.03	0.06	1373
HOME & LIVING	0.56	0.39	0.46	831
IMPACT	0.25	0.05	0.08	714
LATINO VOICES	0.00	0.00	0.00	221
MEDIA	0.47	0.09	0.15	627
MONEY	0.28	0.03	0.05	339
PARENTING	0.31	0.44	0.37	1717
PARENTS	0.27	0.07	0.12	818
POLITICS	0.52	0.87	0.65	7107
QUEER VOICES	0.53	0.40	0.46	1264
RELIGION	0.53	0.09	0.16	520
SCIENCE	0.38	0.10	0.16	441
SPORTS	0.44	0.26	0.33	1050
STYLE	0.39	0.02	0.04	427
STYLE & BEAUTY	0.59	0.65	0.62	2001
TASTE	0.15	0.01	0.02	407
TECH	0.67	0.09	0.16	421
THE WORLDPOST	0.36	0.24	0.29	760
TRAVEL	0.42	0.59	0.49	1978
U.S. NEWS	0.17	0.00	0.01	282
WEDDINGS	0.70	0.48	0.57	725
WEIRD NEWS	0.20	0.08	0.12	529
WELLNESS	0.37	0.70	0.48	3591
WOMEN	0.38	0.13	0.20	700
WORLD NEWS	0.27	0.09	0.14	653
WORLDPOST	0.19	0.02	0.04	537
accuracy			0.44	41906
macro avg	0.38	0.21	0.22	41906
weighted avg	0.41	0.44	0.38	41906

Fig. 4. Our classification report output. (Random Forest, with short description)

epoch	loss	accuracy	val_loss	val_accuracy
1	2.878	0.2874	2.357	0.4078
2	2.2021	0.4379	2.0505	0.4735
3	1.9375	0.4965	1.9331	0.4967
4	0.5311	0.5311	1.8797	0.5119
5	0.5567	0.5567	1.8664	0.5168
test loss	1.8507			
test accuracy	0.5204			

Fig. 5. Epochs for our Word Embedding into Recursive Neural Network Model.

REFERENCES

[1] S. Lei, "Research on the Improved Word2Vec Optimization Strategy Based on Statistical Language Model," 2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS), 2020, pp. 356-359, doi: 10.1109/ISPDS51347.2020.00082.

[2] Sahisnu Mazumder, Bazir Bishnoi, and Dhaval Patel. 2014. News Headlines: What They Can Tell Us? In Proceedings of the 6th IBM Collaborative Academia Research Exchange Conference (I-CARE) on I-CARE 2014 (I-CARE 2014). Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/2662117.2662121

[3] Y. Sun, Y. Li, Q. Zeng and Y. Bian, "Application Research of Text Clas-

sification Based on Random Forest Algorithm," 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2020, pp. 370-374, doi: 10.1109/AEMCSE50948.2020.00086.