ISyE 6740 – Summer 2021
Project Proposal

Team Member Names: Yuheng Tou (903651231), Bingyu Fan (903650948)

Project Title: News Classification

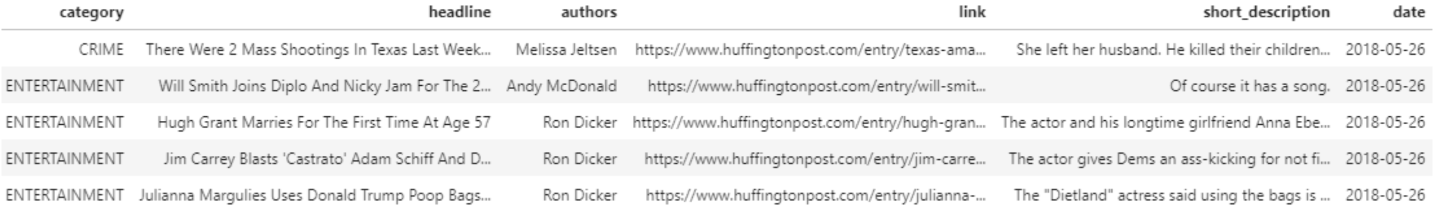
Problem statement

With the rapid development of digital platforms, the amount of online news published every day has grown drastically over the past few years. How must they be distributed to the appropriate audience in a timely and accurate fashion has been a trending research topic. News classification using machine learning methodologies has proven to be quite successful in providing information of interest in real-time.

Our group will utilize several algorithms to perform such news classification using data collected from Kaggle. We first perform vigorous exploration data analysis and choose several appropriate algorithms and finally We will compare the performance of these models.

Data Source

The input dataset consists of 200853 news from the year 2012 to 2018 obtained from HuffPost. The labels provided for this classification problem includes a total of 41 categories, such as Crimes, Politics, Wellness, Entertainment, etc. The text features are news title and descriptions. we will be applying NLP techniques to these columns and convert them into bag-of-words. Another feature that we might use is Author, which we assume most authors will only focus on a smaller number of topics. Suggested by Oliver et al.(2014), adding aggregation features to non-aggregated features could improve the accuracy of classification. Thus, we will create aggregated features such as number of Crime news by author, topics that has not been published by author, etc. However further data exploration is needed to be certain if they will be helpful to our model.



category	headline	authors	link	short_description	date
CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...	2018-05-26
ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	2018-05-26
ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	2018-05-26
ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fi...	2018-05-26
ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...	2018-05-26

After the initial data exploration, we have a few concerns regarding the labels provided by the data owner. For example, Good News and WEIRD NEWS are extremely broad category that may lower our model accuracy. Also, WORLDPOST and THE WORLDPOST, CULTURE & ARTS and ARTS & CULTURE mean the same thing and should be group together. Therefore, further in-depth data exploration and testing with models should be conducted along with proper visualizations to uncover the true categories and help us decide whether if we should merge some of the provided labels.

Methodology

Preprocessing and feature engineering

Standard practices for text preprocessing such as the removal of stop words, adjusts for punctuation, and stemming would be the logical first step, after which feature vector extraction can be performed. Both weighted feature methods such as TF-IDF or vanilla sparse vectors (word counts or binary-included-excluded) seem to have their respective merits for this data set with no obviously superior choice, so all three can be input into our models for empirical evaluation. Theoretically, TF-IDF's weaknesses, name its over-sensitivity to the 'extensive margin' and under-sensitivity to the 'intensive margin' (sensitivity to frequency of occurrence of words in the corpus, irrespective of their concentrations in particular documents) are not as applicable here as we are dealing with titles, in which repeated use of a word is unlikely. Thus, we theorize that TF-IDF would work well. TF-IDF also does not capture elements such as *position*, so more involved methods such as topic models and embeddings may be something we experiment with.

Train-Test Split

We propose an 80-20 split, with samples drawn in an even percentage from each class due to the imbalance of classes.

Models: Classical machine learning and recent advancements

Naïve Bayes is known to have an advantage in handling class imbalances, an issue prevalent in our dataset as the 'politics' class has ~30 times as many data points as many of the less well represented classes. It also is known to work well with small snippets of text. Since we are only dealing with headlines rather than complete passages, it could prove to be a reasonable choice. How much the assumption of conditional independence is violated is unknown, and different NB models such as multinomial etc. will have to be put to the test. To improve the model, metadata can be included via features for 'author' and 'period'. Articles published in the same period would have a higher probability of being associated with the same trending topic.

Besides Naïve Bayes, we will choose a set of appropriate models that are available in the SciKit-learn, such as SVM, logistic regression, decision trees, and boosting tree models. We will test all possible models and compare the accuracy of each.

For more modern models, we put forth the case that BERT may be an appropriate model for this data set given the importance of *context* in news articles, in contrast to more 'standardly worded' pieces such as academic texts, for instance. BERT's ability to learn bi-directionally allows it to 'view each piece as a whole' rather than as disjoint vectors. This, we propose, possibly makes it the preferred choice over LSTM etc. (and even LSTM variants with bi-directional capabilities).

Evaluation and Final Results

Our evaluation metrics are classification accuracy and F1 scores. We also intend to present visually the tradeoffs between precision and recall for each category (in one-vs-all format) via precision-recall curves, which are preferable to receiver operating characteristic curves (ROC) in this instance due to class imbalances—data points are heavily skewed towards the largest class, ‘politics’, at the expense of many relatively underrepresented classes that have up to 30 times fewer data points. Along with these graphs, we present the ‘area under curve’ (AUC) analogue to precision recall, the AP score. These visualizations offer insight as to how our models may be *sensitive* with respect to one class, but less *selective* with respect to that same class, or vice versa. This in turn will allow us to fine tune our model accordingly. In terms of practical value, it may be more important to have high sensitivity to one class, for instance, even if it meant less selectivity, so these visualizations would offer guidance on making a well-informed trade-off.

Reference

O. Schulte and K. Routley, "Aggregating predictions vs. aggregating features for relational classification," *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2014, pp. 121-128, doi: 10.1109/CIDM.2014.7008657.

Suleymanov U, Rustamov S (2018) “Automated news categorization using machine learning methods,” *IOP Conf Ser Mater Sci Eng* 459 12006

Akanksha Patro, Mahima Patel, Richa Shukla and Dr. Jagurti Save. (2020). Real Time News Classification Using Machine Learning. *International Journal of Advanced Science and Technology, 29(9s), 620 - 630*. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/13254