Measuring Impact and Consumption of Local News in a Changing Environment

Stanford CS224N Custom (Mentored) Project Milestone

Anish Saha, Austin Pennington, Yu-Chen Tuan
Department of Computer Science
Stanford University

asaha@stanford.edu, atpenn@stanford.edu, ytuan@stanford.edu

Abstract

In this project, we aim to identify "impactful" (award-winning) articles in a large dataset of millions of publications from newspaper and journals across the United States, collected over the past 2 decades (with publication dates spanning the years 2000 to 2019) by leveraging baseline metadata within the articles retrieved using a dedicated XML scraper, alongside a meticulous process of extracting various synthetic features using various natural language processing methodologies. The first step of this binary classification task will be to use NLP methodologies to retrieve features that quantify the prevalence of anecdotal data, quantitative analysis, logistical records, and the extent of further investigations and discourse on the topic, as well as to to detect major entities and topics, using a smaller subset of the dataset of articles. The next step will be to scale this model, perform feature engineering, and optimize feature/hyperparameter selection to develop a model that can effectively classify "impactful" (award-winning) articles on the entire dataset.

1 Key Information to include

- External Mentor: Shoshana Vasserman (svass@stanford.edu)
- External Collaborator: Gregory Martin (gjmartin@stanford.edu)

2 Approach

Theory: Our hypothesis is that earlier authors exert more influence in propagating events, and that this can be inferred to article-article many-to-many connections. Author mentions, article mentions, and shared events all constitute the kinds of "connections" we are interested in extracting from the article corpus. We will primarily be developing deep learning based NLP to extract named entities and/or events, and then construct graphs, where each vertex represents an article and an edge represents similar named entities and/or events between those two articles.

Fundamentally, we will rely on generated PageRank scores that depend on article exogenous features, namely article-article mentions and "endorsements" to contribute to the notion of "influential". Most deep learning tasks in this project will be done for the purposes of feature extraction [1].

There are many state of art papers talking about named entity recognition and events extraction. Gregoire et al applied dual-CNN to target the problem of event identification in crisis situations, and explore how semantic information can be used to enrich the deep-learning data representations. [2]

Yubo et al [3] adopt a convolutional neural network to capture sentence-level clues. Andrew's method [4] is unifying aspects of both information extraction and text summarization.

Sen et al address issues with insufficient training data using a method to automatically generate labeled data by editing prototypes & screening out generated samples by ranking the quality [5].

Baseline: We considered a naive model, using the preliminary features of the author, publication date, and word count. Given these initial features that have been extracted, we pass the data through a Logistic Regression classifier and obtain baseline results (see the "Experiments" section below).

Architecture: All of the code for this project has been written by the team (XML article data processing, content text data preprocessing, web scraping for NYT articles, word embeddings generation, and synthetic feature extraction).

3 Experiments

- **Data**: For this project, the dataset we will be using is the NewsBank corpus, containing millions of articles from local newspapers, appended to a dataset of manually scraped articles from The New York Times. The time period we will be considering is from 01/01/2000 to 12/31/2019. Meanwhile, the awards considered include all local-level journalism awards and Pulitzer Prizes from all categories. We developed an XML parsetr to extract all metadata and content text data for analysis as well as create a dataset in tabular form to track the metadata for each article. See the sample rows below:

Source	Author(s)	Publication Date	Word Count	Content	Award
New York Times	Jane Ada	01-06-2002	982	[01...1...0]	1
Oakland Tribune	John Doe	03-02-2008	189	[00...0...1]	0

- **Evaluation method**: For the purposes of the evaluation of our results, we will include the following performance metrics: Accuracy, True Positive Rate (Recall), F1 Score (evaluation metric combining both precision and recall scores), and the AUC value (Area under ROC Curve). Since the dataset is heavily imbalanced in that most articles do not win awards, the accuracy metric is not the most reliable metric to evaluate the performance of our models. These metrics will be measured on both the training and test sets to give us an effective means of measuring the performance of our classification models. Ideally, we want a high True Positive Rate (Recall), as the research group is primarily focused on identifying the distinguishing qualities of "impactful" (award-winning) journalism.

- **Experimental details**: So far, we have scraped all necessary articles in the dataset using an XML parser to extract all necessary metadata and content from the NewsBank corpus. Following this, we appended data from New York Times articles to this dataset, reformatted with the same metadata columns, and randomly sampled 10,000 articles from the merged dataset. We also developed scripts to numerically encode the content text data into word embedding vectors (for legal reasons, as per the data contract with NewsBank). After cleaning the data and performing some preliminary feature selection, the metadata columns utilized in our baseline models were as follows: `author, publication_date`, and `word_count`, with the binary target variable `award` (where `1` corresponds to award winners, `0` corresponds to all other articles). Using this smaller dataset and a typical `80-20` train-test set split, we utilized a Logistic Regression Classifier to attain results for our baseline model. We also explored a Random Forest classifier, which showed significant performance improvements.

- **Results**: Here are tables highlighting our results so far:

Table 1: Accuracy of Classifiers with small set of features

Classifier	Validation set	Test set
Logistic Regression	87.00%	87.95%
Random Forest	95.80%	95.35%

Table 2: Other performance metrics of Classifiers with small set of features

Classifier	Precision	Recall	F1	AUC
Logistic Regression	0.9664	0.3267	0.4883	0.6621
Random Forest	0.9059	0.8210	0.8614	0.9014

4 Future work

Now that we have scraped all the articles from both data sources (NewsBank and The New York Times) we need to analyze, cleaned and encoded (for legal reasons) the content text data as necessary, and established a clear baseline for our project, our next steps will be to extract more synthetic features from the content text data in order to build a more effective classifier. The features we will be exploring include: `top_keywords`, `anecdotal_data_metric`, `quantitative_analysis_metric`, `has_logistical_records`, and `investigation_metric` that quantify intuitively "impactful" aspects of journalism. To extract these features, we will analyze the article content data for certain keywords, numerical information, and specific phrases. We will also be using various unsupervised neural methods detect major entities and topics within the articles, as certain topics and trends are intuitively more "impactful" than others. For example, an article highlighting a major social injustice is more likely to win an award than an article discussing a new celebrity fashion trend. These synthetic features will allows us to train a more effective classifier that will achieve superior performance in terms of distinguishing "impactful" (award-winning) articles within the dataset.

References

[1] Rajeev Motwani Lawrence Page, Sergey Brin and Terry Winograd. The pagerank citation ranking: Bringing order to the web.

[2] Grégoire Burel, Hassan Saif, and Harith Alani. Semantic wide and deep learning for detecting crisis-information categories on social media. pages 138–155, 10 2017.

[3] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. Event extraction via dynamic multi-pooling convolutional neural networks. volume 1, 07 2015.

[4] Andrew Hsi. Event extraction for document-level structured summarization. In *Knowledge Media Institute, The Open University, United Kingdom*, 2016.

[5] Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng li. Exploring pre-trained language models for event extraction and generation. pages 5284–5294, 01 2019.