Measuring Impact and Consumption of Local News in a Changing Environment

Stanford CS224N Custom (Mentored) Project Report
https://github.com/anish-saha/Measuring-Impact-and-Consumption-of-Local-News

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 from the NewsBank corpus, a dataset of over two million publications from local newspapers and journals across the United States, collected over the past two decades. We leveraged baseline metadata within the articles retrieved using a dedicated XML scraper, alongside a meticulous process of text data analysis for synthetic feature extraction. The first step of this binary classification task was 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 subset of the corpus of articles. After this, the next step was to extract further synthetic features to utilize a Word2Vec clustering methodology obtain categorical groups as features for each article, followed by the use of a pre-trained BERT model to extract major named entities within each article. As expected, the addition of these synthetic features worked successfully for the purposes of building a more effective classification model for the purposes of identifying "impactful" (award-winning) articles.

1 Key Information to include

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

2 Introduction

The value of impactful, award-winning journalism cannot be understated in terms of how it greatly shapes the public's perception of various ongoing trends and events around the world. On a more local level, impactful journalism can highlight ongoing major injustices by various corporate or federal entities, bring major world events, crimes, or court cases to the public eye, and even influence economics and perception surrounding affluent people, major political figures, large organizations, and flawed systems. However, while such journalism is recognized with awards such as the Pulitzer Prize, there exists no way to linguistically distinguish what exactly characterizes these impactful works. Such capabilities would have major positive implications, such as creating more economically viable models for publishers, fostering a more productive culture in the field of journalism, and gaining a better understanding of the aspects of journalism that influence people the most.

As such, in collaboration with the Stanford Graduate School of Business, our team's goal was to develop effective models to tackle the classification task of identifying "impactful" (award-winning) articles in a large dataset of millions of publications from newspaper and journals across the country collected over the past two decades. Leveraging the baseline metadata, various metrics derived using text analysis, and features derived using Word2Vec clusters and BERT entity extraction, our goal is to develop a predictive model that is able to analyze the content of an article, common contemporary trends or entities, and other article metadata to output whether a given article is "impactful."

3 Related Work

There has been much work done in the space of using unsupervised learning techniques, combining these methodologies with models to produce word embeddings, such as Word2Vec. [1] Cha et al take the approach of applying unsupervised clustering methods, on multidimensional word embeddings, as a feature extraction and dimensionality reduction technique, using the extracted categories as higher level features that can then be passed into text regression. The paper elects to use Brown clustering to extract their clusters/features and pass it into a linear SVM. We adopt a similar approach, electing instead to use k-Means clustering based Word2Vec embeddings in our approach, in order to extract categorical features to enhance the performance of our classifier.

Research in the field of events extraction and named entity recognition (NER) have been used to retrieve major events and entities from sentences. Gregoire et al applied a dual-CNN methodology to target the problem of entity identification in crisis situations, exploring how semantic information can be used to enrich the deep-learning data representations. [2] Andrew Hsi's method unifies aspects of both information extraction and text summarization to improve language understanding models. [3] Jason et al suggest an architecture that automatically detects word- and character-level features using a hybrid bidirectional LSTM-CNN architecture, eliminating the need for feature engineering. [4] These works influenced our decision to extract entities from the article content text data.

Thus, decided to explore a similar, less computationally intensive alternative. Google has continually introduced various pre-trained BERT models since 2018, where they have been applied to Named Entity Recognition (NER) tasks. [5] Our project aims to bypass our computational limitations by making use of Google's pre-trained BERT models, allowing us to use a similar deep learning approach and perform entity recognition to detect major events, people, locations, and organization within the article content text data. This methodology is then used to extract a categorical synthetic feature relating to common named entities from articles for the purposes of detecting named entities that "impactful" articles frequently discuss.

4 Approach

The first step of our approach was to build an XML parser to extract all relevant metadata from the NewsBank corpus. Due to computational limitations, we randomly sampled only 8000 typical articles and 2000 award-winning articles for the purposes of our analysis. The metadata features provided in corpus for each article include: the publication date, the headline, the content text data of the article, the author, the section, and whether or not the article won an award. We then manually extracted the word count from each article, and used all of the features available to build our baseline classification models, using a Logistic Regression Classifier and a Random Forest Classifier. Although the baseline models were relatively effective in terms of accuracy, there was still much room for improvement, especially in terms of detecting the true positives (award-winning articles) within the dataset.

We then utilized rudimentary text analysis and keyword-matching methodologies to extract intuitively useful synthetic features within each article; various metrics were explored to improve the viability of our classification models. One such metric analyzed the presence of quantitative data and analysis within the article, extracted by searching for numerical data in the content text data and by utilizing phrase-matching for common numerical analysis phrases. Another metric measured the expedience of an article, as it seemed intuitive that articles that discuss a recently trending news topic or breaking news event are more likely to win an award. Other metrics analyzed the text data to extract synthetic features that measure the amount of personal anecdotes, accountable sources, logistic records, and the amount of investigation done within each article.

Our next step was to leverage Word2Vec clustering methodologies to group articles together by topic, which would intuitively inform our models to more effectively classify "impactful" (award-winning) articles. Our topic feature extraction relies on two assumptions, that we leverage to deal with computational concerns while calculating our models. Firstly, Word2Vec relies on consecutive word orderings to extract embeddings, and so a dropout methodology or random sampling technique seemed to be inappropriate. Secondly, we make the assumption that the topic of the article that will be described in the first n-words, is also the topic of the article in the in the whole article. We decided to use $n = 100$ due to computational restriction caused by copyright restrictions on moving the article data to different machines; higher dimensionality embeddings were not computationally feasible on the provided NewsBank virtual machine. As such, for each article we extract the first 100 words and generate word embeddings from these excerpts using Word2Vec. We then pass those embeddings through a dimensionality reduction algorithm to generate a reduced matrix, which is then passed into a k-Means clustering algorithm ($k = 5$ is chosen, as most articles in our dataset originate from one of five different major sections in the dataset). This maps each word to a cluster, and running over the corpus of truncated articles generates a list of labels corresponding to each word in an article. From this, we average those articles' labels and assign them as article-level labels, that we treat as a proxy for article topics. These cluster labels are then used as a categorical feature variable to improve the predictive analytic capabilities of our classification models.

Extracting event-related information as features for each article is an important step in determining the value of articles. [6] However, all research in events extraction utilize LDC corpus which would require further computational and financial resources. Nevertheless, exploring this avenue could be a future enhancement that would improve the performance of our classifiers. Therefore, instead of extracting event-related information, we decide to extract the best possible alternative. Named entities mentioned in each article (such as influential people, large organizations, and significant locations) seemed like a great option to help improve the model's language understanding of each article. We used a framework called `DeepPavlov`, which works on top of Google's pre-trained `BERT-Large` model to perform Named Entity Recognition. Using this methodology, we extracted the most frequent named entities in each article. These were then passed into a target encoder (so as not to introduce an incredibly high dimensionality categorical variable); our model then leveraged these common named entities as a categorical feature to better inform the model's decision on what named entities exemplify an "impactful" article.

Next, the content and headline columns are removed. The entire dataset is then split into the training set, the validation set, and the test set. The training data, alongside all of the synthetic features we extracted, are then passed into four different classification models: a Logistic Regression Classifier, a Random Forest Classifier, a Gradient Boosting Classifier, and a Multi-Layer Perceptron Classifier. Finally, these models are used to generate predictions on both the validation ans test sets, yielding performance metrics that we can use to analyze the efficacy of our classification models.

5 Experiments

5.1 Data

For this project, the dataset we will be using is the NewsBank corpus, containing millions of articles from local newspapers. Due to copyright-related legal restrictions with porting over the data to different machines, we decided to work with a subset of the entire corpus. As such, we randomly sampled 2000 award-winning articles and 8000 typical articles from the dataset. The awards considered include all local-level journalism awards and Pulitzer Prizes from all categories. We developed an XML parser 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:

Publication Date	Author(s)	Word Count	Section	Award
01-06-2002	Jane Ada	982	Politics	1
03-02-2008	John Doe	189	Obituary	0

5.2 Evaluation method

For the purposes of the evaluation of our results, we will include the following performance metrics: `Accuracy` (percent of correctly classified samples), `Recall` (true positive rate), `Precision` (evaluation metric combining true positive and false positive rates), `F1 Score` (evaluation metric combining both precision and recall scores), and the `AUC-ROC` 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. We aim to optimize the `Recall` score for our models, as the research group is primarily focused on identifying the distinguishing qualities of "impactful" (award-winning) journalism.

5.3 Experimental details

First, we 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 randomly sampled a subset of 10,000 articles from the dataset. After preprocessing and cleaning the data and performing some preliminary feature selection, the metadata columns utilized in our baseline models were as follows:

`author`, `section`, `publication_date`, and `word_count`
Target Variable: `award` (`1`: award winner, `0`: typical article).

Using this smaller dataset and a typical `80-20` train-test set split, we utilized a Logistic Regression Classifier and a Random Forest Classifier to attain results for our baseline models. Following this we utilized text analysis, Word2Vec embeddings with k-Means ($k = 5$) clustering, and BERT-NER to extract the following synthetic features based on the content of each article:

`quantitative_data_index`, `accountability_index`, `logistical_record_index`,
`anecdotal_data_index`, `investigation_index`, `breaking_news_index`,
`word2vec_cluster_label`, and `bert_ner_entity`.

Further technical details about the feature extraction procedure for these features can be found in the **Approach** section of this document. Appending these features, the dataset is now passed into four different classification models (Logistic Regression Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Multi-Layer Perceptron Classifier) to generate the results for our final models.

6 Results

Baseline Classification Models - Here are tables highlighting our baseline results:

Table 1: Baseline Classification Models - Accuracy

Classifier	Validation set	Test set
Logistic Regression	76.69%	90.30%
Random Forest	95.13%	96.75%

Table 2: Baseline Classification Models - Other Performance Metrics *[Test Set]*

Classifier	Precision	Recall	F1	AUC
Logistic Regression	0.999	0.030	0.058	0.515
Random Forest	0.789	0.920	0.850	0.946

Final Classification Models - Here are tables highlighting the final results of our models:

Table 3: Final Classification Models - Accuracy

Classifier	Validation set	Test set
Logistic Regression	76.55%	90.25%
Random Forest	**99.20%**	**98.75%**
Gradient Boosting	98.95%	98.70%
Multi-Layer Perceptron	79.70%	91.65%

Table 4: Final Classification Models - Other Performance Metrics *[Test Set]*

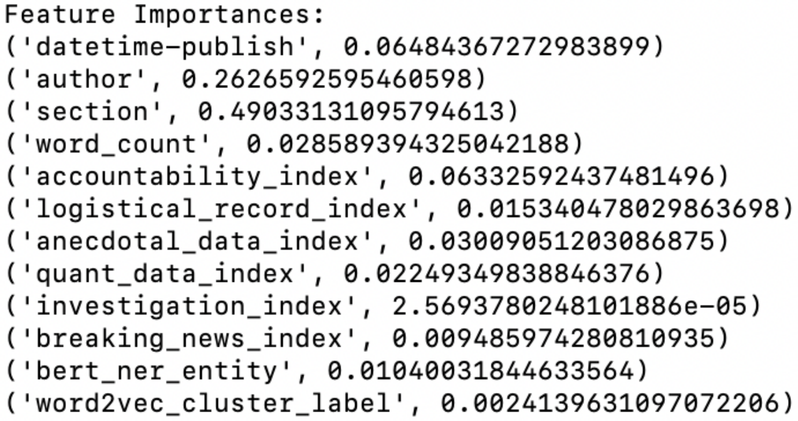
Classifier	Precision	Recall	F1	AUC
Logistic Regression	0.999	0.025	0.049	0.513
Random Forest	**0.900**	**0.985**	**0.940**	**0.986**
Gradient Boosting	0.910	0.965	0.937	0.977
Multi-Layer Perceptron	0.923	0.180	0.301	0.589

7 Analysis

7.1 Results Analysis

Overall, our final classification models were able to outperform the baseline classification models by a significant margin, as can be seen from the performance metrics tables in the **Results** section. Evidently, tree-based classification models performed the best for the purposes of "impactful" article classification, with the Random Forest Classifier achieving an `accuracy` `98.75%` on the test set, a `recall` score of `0.985`, and an `AUC-ROC` value of `0.986`. The Gradient Boosting Classifier performed quite admirably as well, achieving an `accuracy` `98.70%` on the test set, a `recall` score of `0.965`, and an `AUC-ROC` value of `0.977`. These results show a significant improvement over the baseline classification model performance metrics, where the Random Forest Classifier attained an `accuracy` `96.75%` on the test set, a `recall` score of `0.920`, and an `AUC-ROC` value of `0.946`.

The Logistic Regression classifier was very ineffective as it was unable to detect many "impactful" articles and labelled most articles as typical. As such the model only achieved an `AUC-ROC` value of `0.513`. Meanwhile neural-based approach using the Multi-Layer Perceptron (initialized with the following parameters: a `hidden_layer_size` of `100`, the `relu` activation function, an `alpha` value of `0.0001`, a `batch_size` of `200`, a `learning_rate` of `0.001`, and `10000` training iterations) performed very poorly, achieving an `AUC-ROC` value of only `0.589`. Further hyper-parameter tuning and other optimizations may have improved the performance of these models, but the tree-based models significantly outperformed these models in terms of `accuracy, recall`, and `AUC-ROC`, the important performance metrics for the purposes of this classification task.



```
Feature Importances:
('datetime-publish', 0.06484367272983899)
('author', 0.2626592595460598)
('section', 0.49033131095794613)
('word_count', 0.028589394325042188)
('accountability_index', 0.06332592437481496)
('logistical_record_index', 0.015340478029863698)
('anecdotal_data_index', 0.03009051203086875)
('quant_data_index', 0.02249349838846376)
('investigation_index', 2.5693780248101886e-05)
('breaking_news_index', 0.009485974280810935)
('bert_ner_entity', 0.01040031844633564)
('word2vec_cluster_label', 0.0024139631097072206)
```

Figure 1: Random Forest Classifier: feature importances

As can be seen from the figure above, the most important features in terms of influencing the predictions of the model were the baseline features, `author, section, publication_date`, and `word_count`. However, this is because the baseline models had relatively high performance to begin with. It turns out that a bulk of the improvements in performance with these models can be attributed to the new synthetic features we added. More specifically, the synthetic features that added the most value were: `quantitative_data_index, accountability_index, logistical_record_index, anecdotal_data_index`, and `bert_ner_entity`. Unfortunately, `investigation_index` and `word2vec_cluster_label` were relatively insignificant in terms of feature importance for the classification model. Nevertheless, it is clear that the synthetic features we developed were successful in improving the performance of the baseline article classifier, without causing any issues related to increased dimensionality, such as over-fitting.

7.2 Error Analysis

Although our models achieved relatively high performance metrics, there are still a few sources of error that may have artificially increased or decreased the performance of our classification models. One such source of error lies the limitations of our dataset. Due to the fact that copyright-related legal issues restricted our use of the data to a NewsBank virtual machine with relatively low memory capacity, we were unable to work with the full scale of the dataset, which had several million articles and only a few thousand award winning article — a more representative distribution of the dataset would have much fewer award-winning articles. This would mean utilizing a dataset where less than `0.25%` of articles truly winning awards. Since only 10,000 articles are randomly sampled, it is likely that our models do not account for all the variability in the millions of articles within the NewsBank corpus. There were also many issues with the article content text data such as spelling errors, punctuation inconsistencies, and formatting/spacing errors that were not fixed in our data cleaning step, and this may have been another source of error. As such, future steps should include a more rigorous data collection and cleaning methodology prior to performing such analyses.

Moreover, due to the limitations of the NewsBank virtual machine, the BERT-based named entity recognition feature and the Word2Vec-based k-Means cluster label feature were only able to leverage a small subset of the words in the content (the analyses only leveraged the first 50 and 100 words of each article, respectively). Another source of error lies in the fact that the BERT model used in the first analysis leveraged a pre-trained model, which means it was not optimized for the purposes of our analysis. Given more powerful computational resources, it would be possible to develop a more context-specific model that could better detect the entities within articles that are more impactful in journalism. Given further development time to train our natural language processing models, we may also have been able to leverage more complex deep learning models to extract features and enhance the performance of our classification models.

8 Conclusion

In conclusion, the use of synthetic features extracted using text analysis, a Word2Vec-based k-Means clustering analysis, and a BERT-based Named Entity Recognition analysis proved to be successful in significantly improving the performance of classification models for detecting "impactful" (award-winning) articles. Using these features, the Random Forest Classifier model improved by `2%` in terms of `accuracy`, by `0.065` in terms of `recall` score, and by `0.040` in terms of `ROC-AUC` value.

However, the computational limitations of our project caused due to copyright-related legal restrictions forced us to work with a very small randomly sampled subset of the entire dataset and prevented us from developing more complex deep learning models to extract more useful features based on the content of each article. As such, while our classification models were largely successful in terms of performance metrics, there is still much room for optimization.

With more computational resources and development time, next steps would include scaling our classification models, text analysis feature extraction steps, and deep learning methodologies to handle the full content data of the millions of articles in the entire dataset, developing a context-specific custom entity recognition model, experimenting with different k values for the Word2Vec-based k-Means clustering analysis model, and exploring topic detection models and time series based trend analysis models leveraging other deep learning models to further enhance the performance of our classification models. Another optimization would be to explore the prospect of media event detection and time series analyses to distinguish "trend-setting" journalism to help further improve the capabilities of our classification models.

Nevertheless, we hope that our work will be leveraged in the future as the project scales up to handle data on a much larger scale. While the grander goals of the research group — developing models reliant on big data to help researchers to create economically viable publication models, foster a more productive journalism culture, and better understand of the aspects of journalism that influence people the most — are still far from reality, we hope that our research project will provide a robust starting point for future developers and researchers.

References

[1] H. T. Kung Miriam Cha, Youngjune Gwon. Language modeling by clustering with word embeddings for text readability assessment. 2017.

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

[4] Jason P.C. Chiu and Eric Nichols. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[6] Walker, Christopher, et al. *ACE 2005 Multilingual Training Corpus*. LDC2006T06. Web Download. Philadelphia: Linguistic Data Consortium, 2006.