CS224N Project Proposal: Measuring Impact and Consumption of Local News in a Changing Environment

Anish Saha, Austin Pennington, Yu-Chen Tuan

February 13, 2020

1 Key Information

Title: Measuring Impact and Consumption of Local News in a Changing Environment
Team Members: Anish Saha, Austin Pennington, Yu-Chen Tuan
Project Type: Custom Project
Project Mentor: Shoshana Vasserman
External Collaborators: Gregory Martin

2 Introduction

We propose to work on 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 2 decades, from the beginning of the year 2000 to end of the year 2019.

Leveraging a the baseline metadata alongside several synthetic features extracted using various state-of-the-art Natural Language Processing algorithms, **the main goal of our project [1]** is to develop a predictive model that is able to analyze the content of an article, common contemporary trends or entities, and other variables to output whether a given article is "impactful." In the future, the implications of this project will be used as the input for an economics-based model to improve our understanding of the political consequences of changes in the economics of news media, as well as to inform future choices about alternative funding, centered around big data.

3 Related Works

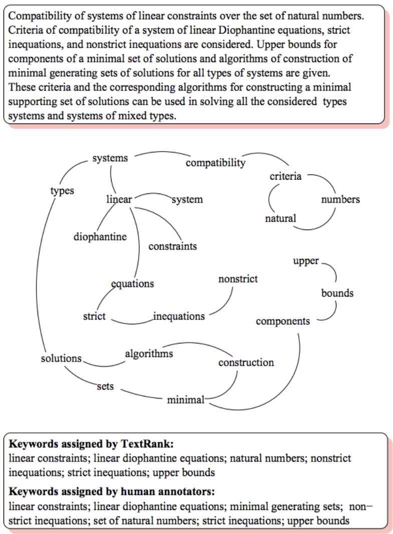
TextRank: Bringing Order into Texts
Rada Mihalcea and Paul Tarau
Department of Computer Science University of North Texas
Published in 2004, Association for Computational Linguistics
https://web.eecs.umich.edu/ mihalcea/papers/mihalcea.emnlp04.pdf

In order to tackle the classification task in this project, **the primary NLP tasks we will be addressing [2]** are topic detection, sentiment analysis, graph/network analysis (to perform entity detection), and content complexity analysis. These will be used to extract synthetic features from the corpus that will later be used for classification purposes.

This paper takes an unsupervised approach to deriving and extracting semantic meaning, specifically as regards to keyword and sentence extraction. This paper is motivated primarily by examining the frameworks and underlying structures behind algorithms like Kleinberg's HITS algorithm and Google's PageRank, used primarily in Web search technology (and identified as a "paradigm shift"), notable for their focus on a graph-based approach to ranking the importance of nodes in a graph using graph-global information. The authors seek to apply this paradigm to natural language processing (NLP) fields, translating the focus on global information, i.e. the entire document, as opposed to node-specific information, which would correspond to simply looking at windows and words. The paper develops such a model, called the TextRank Model, in the hopes that, following the aforementioned paradigms, it becomes competitive in the space of keyword and sentence extraction in NLP.

The authors utilize undirected graphs to represent text, using the graph's structural features to represent interrelatedness between words and other "text entities"; nodes primarily exist as these "text entities", words or collection of words with semantic significance, like sentences or phrases. Edges are used to demonstrate semantic or lexical connections between vertices, which in the paper, they hold to be either weighted or unweighted, directed or undirected, contingent upon the application. Fundamentally, the TextRank Model is built on a system of mutual voting developed in previous literature (like Google's PageRank by Brin and Page, 1998), where vertices and text units "vote" up and down one another's significance. These votes are cast through edge links, and a vertex's score is increased with the number of votes (connections) it has. [MT04]

As stated before, this approach is expanded away from web-based page ranking towards semantic connections/word-unit extraction. TextRank then proceeds, as regards keyword extraction, by preprocessing text: tokenizing, only considering single words, etc. Then, co-occurrence within a window results in connections and edges being constructed between those tokenized units (now vertices in the graph). Then scores are assigned in accordance with the aforementioned ranking algorithm, and vertices are sorted in order of ascending magnitude for their scores. An example of such a connected graph is as follows, from the paper:



Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

Keywords assigned by TextRank:
linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Keywords assigned by human annotators:
linear constraints; linear diophantine equations; minimal generating sets; non−strict inequations; set of natural numbers; strict inequations; upper bounds

Figure 2: Sample graph build for keyphrase extraction from an *Inspec* abstract

The paper's approach is entirely unsupervised, and requires no training, validation, or test data, only the base text. The model is evaluated against "the total number of correct keywords, as evaluated against the set of keywords assigned by professional indexers, and the mean number of correct keyword", and is run exclusively on a set of 500 abstracts from papers listed on Inspec, an indexing database for scientific papers.

The sentence extraction task proceeds similarly, with sentences replaced as the vertices and co-occurrence refined as "similarity" between sentences, where similarity itself is a metric of content intersection, e.g. the number of shared , non-trivial tokens between the sentences. The sentence extraction is evaluated now on 567 news articles, whereby TextRank's sentence extraction generates summaries of said articles. Taken as a whole, the authors claim that the global, unsupervised TextRank is closest to how humans naturally summarize, compared to other supervised approaches; it relies only on the text at hand to generate a summary. It is easily adaptable to longer or shorter passages and longer or shorter summaries, and the unsupervised aspect means it is adaptable to other content domains and languages.

Given these claims, it would be interesting to see TextRank applied to sentence extraction in languages with different semantic grammars and cultures, like East Asian languages; furthermore, the sole testing on keywords from scientific journals is interesting, but straightforward due to the structured and rigid nature of these papers; it would be interesting to see some incorporation of probabilistic models of semantic extraction (word clustering) so that it accounts for synonyms and analogous phrases and idioms.

This paper was selected because of its clear relevance towards the space of feature extraction in news articles; the testing of their algorithm on news articles aside, the extraction of clear keywords and importance sentences from an article are eminently important in our task of extracting endogenous features from news articles, and to the field of feature extraction as a whole. Familiarity with TextRank, PageRank, and similar graph-based algorithms is of seminal importance to our methodology and planned approach, and this paper serves as a theoretical foundation for deeper research into adjacent algorithms (like PageRank). [LPW]

4 Data

For this project, **the dataset we will be using [3]** is the NewsBank corpus, containing millions of articles from local newspapers, appended to a dataset of manually scraped articles from The New York Times and The Guardian US. 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. From this we will extract the 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	Sentiment	Award
New York Times	Jane Adams	01/06/2002	982	+0.3213	1
Oakland Tribune	John Doe	03/02/2008	189	-0.1243	0

One challenge of this project is the computation needed with the size of this dataset. Since the full dataset is comprised of more than 5 million articles, we will use randomly sampling to yield a relatively small development set of 100,000 articles (80,000 training samples + 20,000 test samples), modified to have more evenly distributed award-winning article ratio (since the dataset is heavily skewed right, as a vast majority of articles did not win awards). We will then optimize our methods of feature engineering/selection, model selection, and hyperparameter tuning to yield a final model that will work best on the entire dataset in terms of effectively detecting all "impactful" articles.

5 Methods

There are several factors determining which article is the most influential article. The earlier the article is published, it is likely it is more influential. If an event or fact is mentioned early in the news, then referenced in another news at a later time, the news that was published earlier is likely to be more influential than the latter. Furthermore, if a new unlikely or unknown event is published in a well-documented way, this article will likely draw more attention and be considered "impactful" journalism. For this project, the **neural methods we plan to explore [4]** are the use of deep learning via LSTM neural networks; they will help us to analyze and derive synthetic features from the text data in the corpus by solving supervised learning tasks such as topic detection.

The NewsBank corpus has millions of articles. The articles can be visualized as many large directed graphs in the corpus, where each graph represents a relevant news subject group. A subject group may or may not have an edge connecting to another subject group. Within each subject, each node represents an article and each directed edge represents a reference to the event stated in another article. A graph can grow very large depending on how many News articles referencing to the events in the earlier news articles.

The news articles can be generally classified into two categories: local news and global news. However, there are many categories of topics for articles, such as sports, finance, politics, and fashion. In the first step, we will use deep learning to perform topic detection in order to cluster relevant news subjects into groups.

The PageRank algorithm can be applied to each subject group to help predict which article could be the most influential. The way how the algorithm works intuitively is if a node is referenced more by the subsequent nodes, the node being referenced will have a higher PageRank score. Because the earliest published article is represented as a sink node in the graph, the PageRank score of that sink node is likely to be high.

An edge in the graph represents an event reference between two nodes. Therefore, to build the graph, the NLP plays a critical role in order to extract the events and facts from the text. We will apply deep learning techniques to extract features relating to these named entities from the corpus.

Finally, once these synthetic features have been extracted, we will use impactful article annotations in the NewsBank corpus and classify articles using deep learning based on all data collected, coupled with the synthetic metadata features extracted in the earlier steps, ensembled with the PageRank algorithm, to achieve the most effective models.

5.1 Baselines

In terms of **the baseline models we plan to use [5]**, two simple predictive models are used: logistic regression and random forest — trained on the smaller sample dataset with just two synthetic feature columns of additional metadata: `article length` (computed by counting the number of total words in the article, excluding standardized stopwords) and `article sentiment` (computed using Python's NLTK library).

6 Evaluation

6.1 Oracle

The Oracle for this project would be attaining 100% accuracy on the dataset and perfect AUC of 1.0 — in other words, a predictive model that predicts the correct label for all articles, identifying all award-winning ("impactful") articles while simultaneously avoiding any false positives.

6.2 Performance Metrics

For the purposes of the **evaluation of our results [6]**, we will include the following performance metrics: Accuracy, False Positive Rate (Type I Error), False Negative Rate (Type II Error), F1 Score (evaluation metric combining both precision and recall scores), and most importantly, the AUC (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.

References

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

[MT04] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *EMNLP*, 2004.