Know before you read

An article classifier

BSAN 6200: Text Mining and Social Media Analytics

Carla Mariana Fera, Rongxing (Vincent) Chen

May 6th, 2020

Table of Contents

Background

Objective

This report's main objective is to provide a convenient method in article classification for web browsers such as Chrome, Firefox, or Safari. By using the classifier, users could know what promotional tones the articles are holding before reading them. It could be very convenient for the users especially those who are frequently searching for information online.

In addition, businesses that have their websites online can take advantage of this article classifier and use it to classify their own articles automatically. In this sense, they would avoid having inappropriate articles according to their content.

Our purpose in this paper is to apply some of the techniques that we have learned in Text Analytics so far. Applying these techniques in a real-world case scenario can give us a better sense and understanding of how to work with the data.

Lastly, but not least our expectations are to be able to provide a new solution to an existing problem. There is a demand for condensing and presenting the data in a way that is convenient to the user. With our classifier, the stakeholders involved can perform this type of task in a professional manner.

In other to assess which articles have a promotional tone and which ones not, we need to define what a good article is. According to Wikipedia, A good article (GA) is an article that meets the good article criteria, passing through the good article nomination process successfully. The good article criteria consist of:

1. Well written: clear, concise, and understandable.
2. Verifiable with no original research:
3. Broad in its coverage: it stays focused on the topic without going into unnecessary detail.
4. Neutral: it represents viewpoints fairly and without editorial bias, giving due weight to each.
5. Stable: it does not change significantly from day to day because of an ongoing edit war or content dispute.
6. Illustrated, if possible, by media such as images, video, or audio.

Problem Statement

As the Internet becomes more and more convenient to access with computers, phones, pads, people could check on the information they want any time in any place. However, the Internet is not a paradise, there are too many articles online that are provided with a promotional tone. They are not like the advertisements waiting to be clicked, which could be blocked by some ad-block extensions within the web browser like Chrome. Web browser users could be easily misguided by those articles or wasted too much time before realizing the nature of those articles.

Now, everything could be changed. With our algorithm provided, users can know immediately if an article is being promotional and what type of promotional tone it has before they start reading the articles. It would be greatly beneficial to any web browser users especially the users who are used to searching for useful information online. For very frequent users on researching, they might even be willing to pay for this useful tool.

There is some research done in this field, but we found that none of these investigations have actually applied the classifier to a real-world scenario. For instance, in “Detecting Promotional Content in Wikipedia” by Bhosale, Vinicombe & Mooney (https://www.aclweb.org/anthology/D13-1190.pdf), they use a similar approach; however, there is no business sense combined with their findings. A similar situation happens with “Bert-Based Promotional Words Detection Classifier Development” by Zikun Lin (http://www.cs.columbia.edu/~jrk/NSFgrants/videoaffinity/Interim/19x_zikun.pdf). We decided to take this classifier one step further and provide users and businesses with a real tool that can be used when encountering promotional articles.

Methodology

Data

Our data comes from Kaggle. One of the datasets has more than 23,000 Wikipedia articles, tagged by Wikipedia with possible promotional tones (advertisement, conflict of interest, fan’s point of view, press release, resume). Another dataset contains more than 30,000 “good” Wikipedia articles, without any promotional tones. Both are very large datasets, and thus, very beneficial for our analysis and modeling. Wikipedia articles are also the most popular articles to read for researching. It sets up a good example of “good articles” or “promotional articles” for us to reference. The link of datasets from Kaggle and promotional tone classification from Wikipedia are listed below:

- https://www.kaggle.com/urbanbricks/wikipedia-promotional-articles
- https://en.wikipedia.org/wiki/Category:Articles_with_a_promotional_tone

(Note: the tags/labels for classification used in Wikipedia are changed during this project, once we are able to update the datasets with the new tags/labels, the analysis and algorithm from this report could be using the tags Wikipedia is using for today.)

After analyzing the characteristics of these two datasets with sentiment analysis and topic modeling, we combine these two datasets together into a big dataset with seven main columns:

1. Text: text from Wikipedia articles
2. Promotional: 1 means promotional tone, 0 means no
3. Advert: 1 means yes for advertisement tone, 0 means no
4. Coi: 1 means yes for conflict of interest, 0 means no
5. Fanpov: 1 means yes for fan's point of view, 0 means no
6. Pr: 1 means yes for a press release, 0 means no
7. Resume: 1 means yes for resume, 0 means no

We will use the "text" column as the independent variable for our model and the rest six columns as our dependent variable.

Process

The text data is clean enough and requires no further processing. This credit should go to Wikipedia and the dataset contributor in Kaggle.

As we already know, these labels given in the datasets are promotional, Advert, Coi, Fanpov, Pr, and Resume. Before we actually build the classifier, we are going to explore the characteristics of the texts which are under these labels. This process does not only help in

building the classifier, but it also helps the users to understand what it means when a certain type of classification is given.

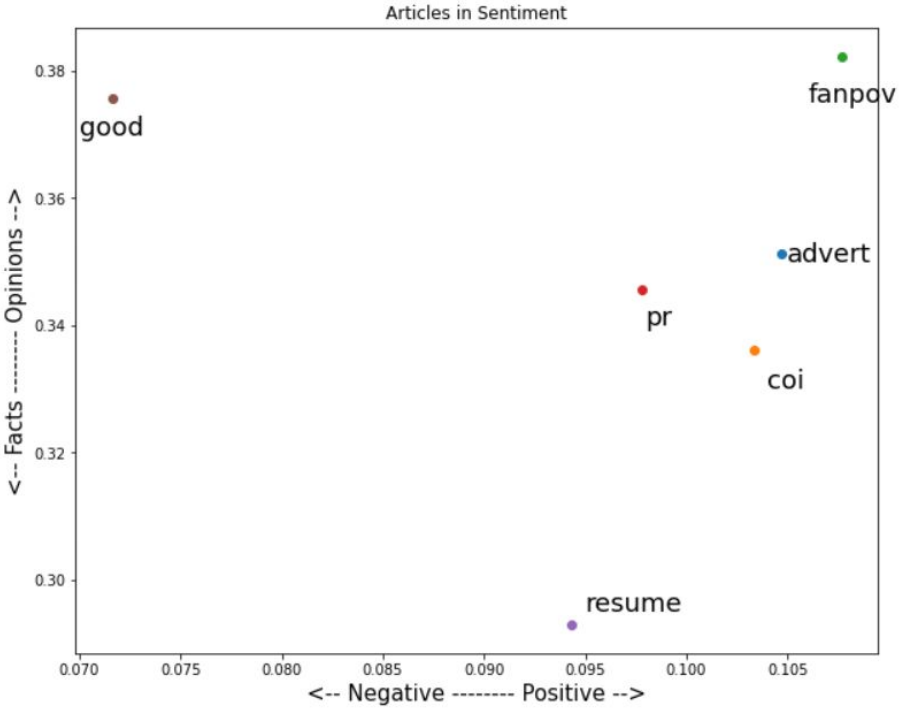
In further analysis, we first run a sentiment analysis on the datasets. A sentiment score is assigned on polarity and subjectivity for each text. Next, by grouping out the articles, we are able to see the average sentiment scores for each type of article. It helps in understanding the sentiment differences among the good articles and different types of promotional articles.

We also run topic modeling on our promotional dataset. The purpose of running topic modeling is to assess which are the most common topics in which promotional tone occurs. There are several steps we took in order to be able to run our model. Firstly, we did some data preparation. We tokenize each article into a list of words, removing punctuations and unnecessary characters all together and we perform bigram and trigram models. Following, we proceed to remove stopwords from the bigrams and do lemmatization keeping only noun adjectives, verbs, and adverbs. Finally, we created our term-document frequency table and we run our model on this one.

Last but most important, we developed the algorithm that could employ the text data and generate the outputs for the six outputs we want. A CountVectorizer is used to convert data to a matrix of token counts, so we could use TfidfTransformer to transform the matrix into a normalized tf-idf representation. In the end, we are employing a random forest model to make the prediction.

Results

Sentiment Analysis



Above is the chart that shows different tones of articles from the sentiment analysis. The x-axis stands for polarity, measuring the text as being more positive or negative. The y-axis stands for subjectivity, measuring the text as being more subjective or objective.

The good articles have a polarity score of 0.07, which means they are usually neutral in wording, not being very positive or negative. Their subjectivity is 0.375, a little bit lower than 0.5, implying they are more facts-oriented.

We could see that the promotional articles are a lot more positive than the good articles, this is easy to guess since promotional articles need more positive tones to be “promotional”.

However, to our surprise, the promotional articles are usually more fact-oriented than the good articles, mostly because they are trying to be persuasive as they can.

Among the promotional articles, they are not very different in the polarity, although we can see that resume-like articles are being most negative in wording and fan's-point-of-view articles are being most positive. There are great variances in promotional articles when it comes to subjectivity. Fan's-point-of-view articles are at about the same level of the good articles, while the resume-like articles are very objective in tone. The other three types of articles are in the middle of the two.

Topic Modeling

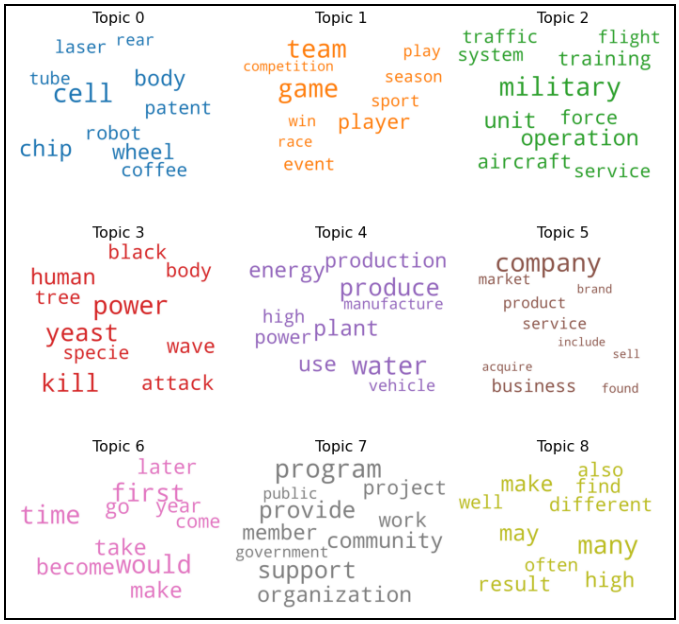
Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

For the purpose of our research, we applied this technique in order to asses which are the topics that mostly occur between the datasets. Specifically, we have applied Latent Dirichlet Allocation Model from Gensim in Python.

The results of this model will give us a set of topics that contains a set of words for each one of them. The main idea is that each topic is clearly different from each other, in order to separate them more intuitively.

Some of the main findings of this model for each dataset are:

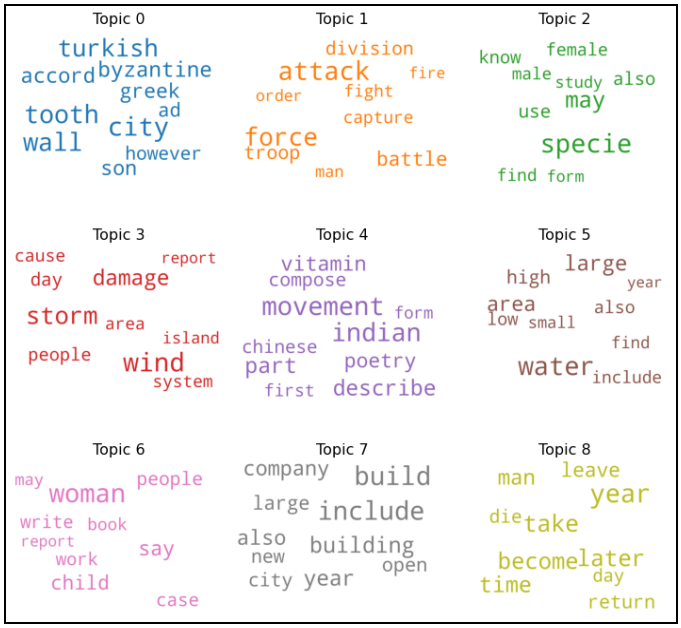
- **Promotional Articles**



Overall, we found 20 main topics in our data. The WordCloud above is only showing 9 of these 20 topics. To be able to arrive at this total number of topics, we built a function that created different LDA models with different numbers of topics. For each of these models, the function calculated a coherence score, which measures the degree of semantic similarity between high scoring words in the topic. Consequently, we determine that the model with the highest coherence score had the optimal number of topics.

All the topics in the visualization above contain some type of promotional tone in their articles. The topics are Science/Robotics, Sports, Military, History, Power Supplies, Business, Releases, Society and Results. Moreover, each WordCloud represents the 10 most common words used in each type of topic. Knowing which topics are prone to having a promotional tone can facilitate the earlier detection of these ones.

- **Good Articles**



The same process was done for the good_articles dataset. The visualization above contains 9 topics among the 20 total topics in the model. We defined these topics as Culture, War, Science, Weather Conditions, Cultural Movements, Lands, Books, Companies and Historical Events.

With this new set of topics, users can assess which are the articles' topics that are less prone to have a promotional tone on them. In addition, we can consider which words, either substantives or verbs, are most commonly used when writing a good article that has no advertisement on it.

Classifier

The accuracy of the model we finally got could reach to 0.81, which is high enough to recognize most articles online for our web browser users. If we input the article to this model, we could get a detailed classification for it. A resume text file from the internet was tested with the classifier, and we got the output as a promotional article, with a tone of advertisement. It makes sense because resumes are just like advertisements for ourselves.

There are several advantages to this classifier:

- **Convenient**

It is a very easy and useful classifier. All we need is to input the text data we want to analyze to the classifier, then in a very short period of time, we could get the output claiming this article is promotional or not, and what type of promotional tone it has. If it is developed like an extension in web browsers like Chrome, it is possible to pop up the classifications as soon as the users get into the website, right before the users start reading the article. With those classifications, users could decide if they want to continue to read the article or not. It could save a lot of time for web browser users.

- **Accurate**

The accuracy could reach to 0.81, a high number without deeper complicated tuning. It means that the classifier is good enough for users to judge the quality of articles they intend to read.

Limitations

Data Representativeness

The data used for building the model is only coming from Wikipedia, probably the most popular and professional website for reading articles for research-use. Given this, there is some concern that the articles collected are too “good” for being a representation of all of the articles online. If that is true, the bar for being a “good” article might be too high and seems biased and unfair for all other articles online, as a bar we set with the training on our classifier.

Imbalanced Data

We did inspect the sample size based on the categories, and we found that our data is imbalanced. For example, most of the promotional articles are in the category of advertisement, which could hurt our model prediction, but that is also a question in the real world. Most promotional articles do look like advertisements. In the end, we decided to keep going with our model with the raw data because there might be some features that can be captured only with the whole dataset; there is also a time limitation to prepare the project.

Black-Box Model

As we all know, the Random Forest model is a black-box model. It is difficult to asses which single decision tree is supporting the final result and which decision tree is not. There is not much space for tuning the model to best fit the text data. We believe there might be a better understanding of the text data and better classification results if we could apply more advanced word embedding methods and neural networks to our classifier. Probably, it is not the best

model theoretically that we could have applied, but it is the best model we can find in the market.

Running Time Complexity

When dealing with datasets that are lengthy, a complexity comes involved. Coded programs, in general, and especially text analysis models, take a long time to run; making the process slow and inconvenient. In our case, we had time constraints, which meant to fulfill this project upon the deadline. This fact causes that we had to accommodate to the situation that we were facing as best as we could. For instance, for the good articles dataset in topic modeling, we had to run our created function with 25 maximum topics because the program was crashing when we use any number above this one. In contrast, in the case of the promotional dataset, this problem has not occurred even when using 40 topics. These situations are needed to be handle as best as we can, despite the circumstances, which sometimes are out of our hands.

Recommendations

For Practitioners and Stakeholders

Stakeholder/Practitioner	Recommendations/Implications
CEO of Web Browsers / Director of Quality (i.e.: Google Chrome, Firefox, Safari, etc.)	• Recognize the complexity of articles from the Internet • Use our model to classify articles online for users' convenience • Develop an extension within the web browsers • Increase customer satisfaction/product quality by providing customers with a useful tool • Attract more customers/users by introducing new sub-products
Marketing Analysts	• Improve interests in text analysis • Develop articles that can capture the user's attention by using the correct tone • Realize that people are seeking useful information
Websites' Owners / Administrators	• Avoid using articles that are misleading or have promotional tones • Implement this tool in articles to assess written communication tone • Recognize users' real needs

Resources

https://www.kaggle.com/paoloripamonti/twitter-sentiment-analysis/data

https://www.kaggle.com/ndrewgele/omg-nlp-with-the-djia-and-reddit/comments#133158

https://en.wikipedia.org/wiki/Category:Articles_with_a_promotional_tone

https://www.kaggle.com/urbanbricks/wikipedia-promotional-articles

https://www.kaggle.com/benhamner/clinton-trump-tweets

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#3importpackages

https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/#9.-Word-Clouds-of-Top-N-Keywords-in-Each-Topic

https://python-graph-gallery.com/11-grouped-barplot/

https://monkeylearn.com/blog/introduction-to-topic-modeling/

Individual Contribution

Carla Mariana Fera

Project Paper
- Background: Objective, Problem Statement
- Methodology: Process (Topic Modeling)
- Results: Topic Modeling (Promotional Articles, Good Articles)
- Limitations: Data Representativeness, Black-Box Model, Running Time Complexity
- Recommendations

Python Notebook
- Data Exploration
- Data Cleaning
- Topic Modeling (Promotional, Good Articles)

Presentation
- Background (Problem Statement, Objectives, Key Goals)
- Methodology & Results (Data Information, Process, Data Exploration & Cleaning)
- Machine Learning Models (Topic Modeling)
- Recommendations & Implications

Vincent Chen

Project Paper
- Background: Objective, Problem Statement
- Methodology: Data, Process (Sentiment Analysis, Classifier)
- Results: Sentiment Analysis, Classifier
- Limitations: Data Representativeness, Imbalance Data, Black-Box Model
- Recommendations

Python Notebook
- Data Exploration
- Data Cleaning
- Sentiment Analysis
- Classifier

Presentation
- Background (Problem Statement)

- Machine Learning Models (Sentiment Analysis, Classifier)
- Advantages
- Limitations