



What’s New?

A Deep Dive into What News Articles are Talking About

Adrianne Van

Brian Vennemann

Andrew Taylor

Nick Videtti

IST 736 Text Mining Final Project

Spring 2023

Adrianne Van ([pvan01@syr.edu](mailto:pvan01@syr.edu))

Brian Vennemann ([bvennema@syr.edu](mailto:bvennema@syr.edu))

Andrew Taylor ([ataylo27@syr.edu](mailto:ataylo27@syr.edu))

Nick Videtti ([nvidetti@syr.edu](mailto:nvidetti@syr.edu))

IST 736 Final Project

Text Mining on News Articles

NewsAPI, Document Vectorization, Naïve Bayes, Decision Trees, and Support Vector Machines

June 2023

|  |  |  |
| --- | --- | --- |
| What’s New? A Deep Dive into What News Articles are Talking About | | |
| **Introduction** | In the world today, even instant information doesn’t seem to be fast enough. Social media, phone applications, texts, emails, and notifications have given and continue to give an extremely high standard for speed and convenience that is often-times an unreasonable expectation. Waiting for a web page to load on dial-up internet 20 years ago took minutes, but now anything more than a few seconds will usually indicate some sort of technical error or limitation. Slow information is not much better than no information at all in today’s world. Lack of immediate information can and will give anybody a significant disadvantage over those with immediate information. Every aspect of life has been affected by this phenomenon, even the once-almighty newspaper.    While it may not always seem like it, given the continuous decline of the once abundant newspaper, there is exponentially more news than there has ever been in history. This is mainly because the internet has taken over as the main platform for delivering news, which allows for a seemingly unlimited amount of news sources and real-time updates to any story. It is no longer the only option to rely on a handful of publishers selling daily newspapers at the stand or on the shelf with only a dozen pages. In fact, those who still use newspapers as their main source of news are, ironically enough, less informed than those who embrace the real-time news that is all over the internet. Instant news from all over the world is available at any smart phone owner’s fingertips, and being the first to report a story is as important as ever to the prestige of news sources.    Unfortunately, all ground-breaking technologies and/or services result in some people looking to exploit them. In the case of news, this means the spread of false or biased information, which is often used to manipulate or persuade people into believing false narratives. There are also satirical news sources that are not malicious but are sometimes mistakenly interpreted as real news. All of this can change the opinions and values of society, which is extremely dangerous when it is not based on undisputed facts. Because of this, it is as important as ever to be informed about which news sources are reliable, what the reliable news sources are talking about, and what the overall sentiment of the news is.    Something that the internet has not changed in the world of news is the importance of keeping up with news itself. Regardless of the volume and frequency of information, the news tells the reader about the world around them. This may sound obvious, but it is a powerful statement, as many do not appreciate the importance of this. From something as relaxed as conversations with friends to something as serious as a job interview, not being up to date on the news can make a person seem like they are out of touch or intellectually inferior, whether that is actually true. Having a feel for the state of the world can also better prepare somebody for the world so they are not blindsided, or simply left behind, by society. Ignorance is not always bliss.    To gauge the topics and the overall sentiment of the news, many online news articles will be analyzed from a reliable source, NewsAPI. NewsAPI is a place to find many news articles from many different reliable sources. While these may not all necessarily be free of bias, they should all be factual and a good representation of news and the world. There will be some research into article topics, article sentiment, and what is being talked about in each topic. |
| **Analysis** | **data preparation and cleaning**  The data preparation and cleaning process utilizes an assortment of user-defined functions:   1. data\_pull(apikey, topic, date\_start, date\_end) 2. VectorizeModelCrossVal(dataframe, text\_column, label\_column, cross\_val\_folds = 5) 3. VectorizeLDA(dataframe, text\_column, n\_grams = 1, tfidf = False, boolean = False, vec\_in\_title = False, title = 'LDA Model Topics', LDA\_topics = 5, top\_words = 15) 4. wayback(x) 5. vader\_sentiment(text) 6. categorize\_sentiment(sentiment, neg\_threshold=-0.05, pos\_threshold=0.05)   News articles will be pulled from NewsAPI using a Python user-defined function, *data\_pull*. The function will have parameters for an API key to access the API, topic so that specific topics can be queried from the API, and two other parameters for the start and the end of the desired date range. The function reads in news articles based on the aforementioned parameters and cleans the data as it comes in. Regular expressions are used to remove numbers and special characters and then the data is appended to a CSV file with data for two columns, one for the topic and one for the article itself. The *data\_pull* function works in tandem with *wayback*. *Wayback* takes in a topic and applies a 30-day filter to utilize the *data\_pull* function to retrieve 30 days' worth of articles on the specified topic. The resulting data frame from the data pulled from NewsAPI is seen below in *figure 1*.    Figure 1 - News Topics Data Frame  First, the text data will be manipulated and further cleaned. This will include three major steps. The first step is to use NLTK’s word tokenizer to split the text into words. Once that has been done, the words will be cleaned. Cleaning will involve including only phrases without numbers, excluding NLTK’s English stop words, excluding any phrases without letters, converting all words to lowercase, removing words that are three character and less, and stemming words using NLTK’s PorterStemmer. The final step will be to combine all the cleaned words into a single string that can be used as the new text column. The cleaned data set can be seen in *figure 2* below.    Figure 2 - Cleaned Data Set  Next, the text data will be vectorized and then joined with corresponding labels. This will be done using four different vectorizers: CountVectorizer, binary CountVectorizer, TfidfVectorizer, and binary TfidfVectorizer. The result is 4 different vectorized data frames, each of which will be used in 6 different algorithms, except for one algorithm. Bernoulli Naive Bayes requires Boolean or binary vectors, so only the two binary vectorizer data frames will be used for that algorithm. The other five algorithms will be Multinomial Naive Bayes, Decision Trees, Support Vector Machine with a RBF Kernel, Linear Kernel, and Polynomial Kernel.  Each of these models will generate metrics for minimum accuracy, maximum accuracy, and average accuracy across the cross validation. For each model, confusion matrices will be generated. Lastly, the best model, based on the average accuracy metric, will be displayed. In total, this function will create 4 vectorized data frames, 22 models, 66 metrics, 22 confusion matrices, and one result for the best model.  Preparation of models will be executed utilizing the *VectorizeModelCrossVal* and *VectorizeLDA* functions. Data in the form of a Pandas data frame will be passed into *VectorizeModelCrossVal*. Other function parameters will include a name for the column containing the text, a column containing the label or what to predict, and the number of folds to use in the cross validations that will be used to test the models for accuracy. The cross-validation folds parameter will be defaulted to 5, or in other words, 5-fold cross validations.  *VectorizeLDA*, will create a Latent Dirichlet Allocation (LDA) based on the data frame that is fed in. Other function parameters will include the column name for the text data, and number of n-grams to use during document vectorization. These parameters determine whether to use TFIDF or Boolean vectorization, vectorizer type is included in the title of the visualization, the title of the visualization, the number of topics to generate, and the number of words per topic. All these other parameters are optional except for the text\_column parameter.  First, the data will be cleaned the same way it was in the *VectorizeModelCrossVal* function, but will also exclude the phrase “char”, which seemed to be highly prevalent but useless in the results during testing. Once the data is verified clean, a vectorized data frame will be created using the specified vectorizer. From there, LDA will be performed on the data and a visualization will be created based on the parameter choices for the title, number of topics (default of 5), and words per topic (default of 15).  The sentiment of the article will also be recorded but will need to be created since it is not pulled from NewsAPI. Through a combination of *vader\_sentiment* and *categorize\_sentiment*, the sentiment will be scored (employing NLTK’s VADER Sentiment Intensity Analyzer) and then categorized for analysis. Scores greater than 0.5 will be considered positive, scores less than -0.5 will be considered negative, and all scores in between will be considered neutral. The NLTK VADER sentiment analyzer factors both polarity (negative or positive) and the intensity strength of each word and emotion. The model utilizes its dictionary to map and score the sentiment of the text. The data frame, in *figure 3* below, reflects the sentiment polarity scores and sentiment determined for each article assigned from the NLTK Vader function.    Figure 3 - Sentiment Data Frame  **data exploration**  Once our finalized labeled data frame was created, it is time to explore the data and topics through word clouds. A for loop is used to create a word cloud for each topic of interest, as seen in *figure 4* below. Each topic’s data was placed in a data frame and then plotted. The resulting word clouds are depicted in *figures 5-9* below.  A screen shot of a computer code  Description automatically generated with low confidence  Figure 4 – Word Cloud Generation  *Figure 5* contains the most common words for the politics topic. Words such as election, political, government and party are common when discussing politics. The words American, Congress, and Biden hint that the dataset skews towards more American news, even when efforts were made to include as many international sources as possible. This may be a limitation of the NewsAPI dataset due to the limited number of well-known foreign news organizations that publish news in English.    Figure 5 - Word cloud of Topic: Politics  *Figure 6* on the following page is the word cloud of the health topic. The two most common words were mental and world. This could be attributed to the current mental health crises that are affecting western nations as well as the remnants of the coronavirus pandemic. ‘World’ could also be related to various initiatives by international health organizations such as the Red Cross and the World Health Organization.  A picture containing text, font, screenshot, graphic design  Description automatically generated  Figure 6 - Word cloud of Topic: Health  *Figure 7* is the word cloud of the sports topic. Some expected common words include game, team, final, series, betting, and league. There are surprising common words that would expand across all news article topics, such as will, one, new, first, time, etc. The words that are expected under this topic would describe sport team names, games being played, activities in recent games and major sport leagues.    Figure 7 - Word cloud of Topic: Sports  *Figure 8* on the following page is the word cloud of the weather topic. Some expected common words include temperature, wind, storm, summer, forecast, sunny, hot, rain, etc. The words that are expected under this topic would describe the seasons, temperatures, climate conditions and updated forecasts.    Figure 8 - Word cloud of Topic: Weather  Figure 9 is the word cloud of the climate topic. Some expected common words include climate change, powered, liter, temperatures, etc. The words that are expected under this topic would describe the long-term changes to weather patterns and atmosphere and the factors that are influencing these changes.  A picture containing text, font, screenshot, typography  Description automatically generated  *Figure 9 - Word cloud of Topic: Climate*  **models and methods**  Support Vector Machine (SVM) (linear (lin) / radial basis function (RBF) / polynomial (poly)), Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes, LDA, and Decision Trees were selected for this analysis. SVM is a deep learning algorithm that processes and sorts the data into the two subgroups by analyzing patterns and applying hyperplanes to differentiate the groups. Naïve Bayes applies conditional probabilities and assumes each parameter (tokenized words) are independent of each other to attempt to classify the movie reviews. Bernoulli applies Boolean logic to the parameters to develop a distribution across the reviews to attempt to classify the reviews. Decision trees attempt to classify the data by analyzing all possible outcomes of a decision using a Boolean approach of whether an event took place. Everything is stemmed from the root word (word the algorithm deemed most important) and branches off into level after level of leaves. All these models are supervised due to the use of the labels and are fed the training data (via k-fold cross validation, where k is a function parameter) to train each model.  **TRAIN/TEST CODE EXAMPLE – Multinomial Naïve Bayes with Count Vectorizer**    LDA is an unsupervised algorithm that takes a Bayesian approach to apply probabilities and conditional dependencies of unobserved groups to classify and group text into topics. The model was then built to create five unique topics from the cleaned data and find the top 15 words for each topic.  **TRAINING CODE EXAMPLE – Latent Dirichlet Allocation**    **ANALYSIS GOALS**  The first goal of the models listed on the following page is to predict the news article topic based on the text of the articles. The models will categorize the article into one of the following topics: politics, climate, health, sports, and weather. The model with the highest accuracy will also be reflected in the results. Given the prediction results of the model, the text of the articles will confirm if specific words or combinations of words are typically utilized to discuss each topic. The second goal of the models is to predict the sentiment of each news article based on the text of the articles. Based on the words utilized, can the models accurately determine if the general sentiment of each article is positive, negative, or neutral?   * Support Vector Machine (SVM) (lin / RBF / poly) * Multinomial Naïve Bayes (MNB) * Bernoulli Naïve Bayes * LDA * Decision Trees |
|  |  |
| **results** | **technical results**  **TOPIC PREDICTIONS**  All models, *figures 10-22* below, were based on 5% of the total data set due to memory allocation issues and generally had a low prediction accuracy of news article topics. The accuracy for all models ranged from approximately 25% to 30%. The SVM (RBF Kernel) with the Boolean TFIDF vectorizer had the highest accuracy of 28.6%. The models particularly didn’t predict the climate and weather topics since both topics are discussed interchangeably. Weather changes are frequently discussed to describe long term impacts to climate and vice versa. The word clouds earlier and sentiment analysis also showed common words that all topics shared, which most likely factored into the low accurate predictions of the topics. The positive and negative news articles also contained many neutral words that would have influenced how the models were predicting the topics.    Figure 10 - Bernoulli/Boolean Topic Results    Figure 11 - DT/Boolean Topic Results    Figure 12 - DT TFIDF Topic Results    Figure 13 - DT Boolean Topic Results    Figure 14 - SVM RBF/Boolean Topic Results    Figure 15 - SVM RBF/Boolean TFIDF Topic Results    Figure 16 - SVM Lin/Boolean Topic Results    Figure 17 - SVM Lin/Boolean TFIDF Topic Results    Figure 18 - SVM Poly/Boolean Topic Results    Figure 19 - SVM Poly/Boolean TFIDF Results    Figure 20 - Topic Model Accuracy Summary    Figure 21 - Topic Model Summary Statistics    Figure 22 - Best Topic Model  **SENTIMENT PREDICTIONS**  The sentiment models, *figures 23-35* below, had a higher prediction accuracy range than the topic models. The accuracy range for the sentiment predictions were 39% to 42%. The SVM (RBF Kernel) with the count vectorizer had the highest accuracy of 42.4%. The low accuracy across the models is most likely driven by individual words that could be positive or negative. Certain words individually can provide false negatives or positives. The accuracy of the sentiment predictions could be improved next time by integrating bigrams or ngrams.    Figure 23 - MNB/Boolean Sentiment Results    Figure 24 - MNB/Boolean TFIDF Sentiment Results    Figure 25 - Bernoulli/Boolean Sentiment Results    Figure 26 - DT/Boolean Sentiment Results    Figure 27 - DT/Boolean TFIDF Sentiment Results    Figure 28 - SVM RBF/ Boolean Sentiment Results    Figure 29 - SVM RBF/ Boolean TFIDF Sentiment Results    Figure 30 - SVM Lin/Boolean Sentiment Results    Figure 31 - SVM Lin/Boolean TFIDF Sentiment Results    Figure 32 - SVM Poly/Boolean Sentiment Results    Figure 33 - SVM Poly/Boolean TFIDF Sentiment Results    Figure 34 - Sentiment Model Summary Statistics    Figure 35 – Best Sentiment Model  **LATENT DIRICHLET ALLOCATION**  The LDA model summary, seen in *figure 36*, reflects five general topics the model identified based on the news articles. However, the topics are not quite aligned to the actual news article topics and the same words are categorized across the LDA topics. Topics 1, 2, 4 and 5 all have weather, health, and sport. Topic 3 is the only distinguished topic with words that are mostly associated with technology.    Figure 36 - LDA Topics  **Sentiment Analysis**  Based on the sample data set of the news articles pulled, *figure 37* depicts that there is nearly an equal number of positive and negative news articles across all topics. Less than a quarter of the news articles are categorized as neutral.    Figure 37 – Total Sentiment Analysis  As seen *figure 38* on the following page, the sports topics had the highest number of negative news articles compared to the other topics. Health has a nearly equal ratio of positive and negative sentiment. Weather has a slightly higher number of positive news articles, but the topic also has a nearly equal number of negative and positive sentiment. Politics has a slightly higher number of positive than negative sentiment. Climate is the only topic that surprisingly did not share the same sentiment metrics as weather and other topics. Climate had a nearly equal amount of positive and neutral sentiment and the least amount of negative sentiment. Overall, the sentiment ratio between positive and negative is expected to vary across all topics depending on current events and issues.    Figure 38 - Sentiment by Topic  *Figure 39* below shows the most common words in positive news articles. The expected common words for positive articles are good and like. However, breaking, free, time and year are some of the highest common words for positive articles and could be considered more neutral. Due to a higher number of these neutral words, this will influence the model’s sentiment predictions since these words can contribute to positive or negative sentiment.    Figure 39 - Positive Common Words  *Figure 40* on the following page shows the most common words in negative news articles. The expected common words for negative articles are abuse, spam, fraud and warning. However, words including, data, services, protect, track and maintain are some of the highest common words for negative articles and could be considered more neutral. Due to a higher number of these neutral words, this will influence the model’s sentiment predictions since these words can contribute to positive or negative sentiment.    Figure 40 - Negative Common Words  *Figure 41* below shows the most common words in neutral news articles. All common words listed below would be expected under neutral articles. The most common words in neutral articles are powered, finished, speed and liter.    Figure 41 - Neutral Common Words |
| **conclusions** | Determining the inherent topics of articles through algorithmic means is a challenging proposition and as shown in this study can be error prone and produce results no better than flipping a coin to guess the topic. One of the reasons theorized is the shared patterns used in the English language. Words and sentences can have wildly different meanings based on the context in which they are used. The good news is that consumers still retain the most powerful tool to parse through garbage articles for more relevant ones. That tool is our innate ability as human beings, and more specifically, our mastery of the English language, to understand that context matters when reading about an event. This context can be found in the article or by combining knowledge from several different topic areas to arrive at a conclusion that provides the most information for lifestyle or policy change.  The importance of the context of an article must be carefully considered, especially when readers are asked to feel or act in a particular way. With such articles, it is imperative that readers have access to as much external information relevant to the topic. This provides an opportunity for news aggregation sites to improve their audience retention and gain new ones by building a reputation of strongly supported well researched, and contextualized articles. One example could be a news article on a climate topic that provides links throughout the article to articles on the political and societal effects of any personal or political action taken in support of or against the area of concern. If a reader is reading about new carbon capture storage breakthroughs, the site could provide links to articles on policy changes that incentive such a technology or the inherent dangers that may come with fielding such new technology.  Although algorithmic tools failed to provide a clear solution for topic delineation, they must still be considered for other aspects of the new aggregation process. The biggest place they can have an impact on is in the process of providing relevant subtopics or related topics that provide additional context on an article. With new tools such as ChatGPT, analysis of large data sets has been made accessible to those without the resources to invest in the infrastructure to support their own models. However, as found in this analysis, users must be wary of the results produced by these models and must always be alert to inaccuracies and falsities.  Furthermore, sentiment is significant to understanding and even shaping the public’s general opinion of various topics in the news. The sentiment of a topic is partially determined by whether the current event and issue itself is positive, negative, or neutral. However, the news media can use sentiment to influence the public’s perspective or opinion of current events and issues in the news. Since the news contains a varying degree of bias depending on the news source, readers should seek articles from multiple news sources to understand the general sentiment of the topic. A general sentiment shared within the public can impact the people and entities being discussed. |
|  |  |