PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

NewsAnalyticalToolkit: an online natural language processing platform to analyze news

McCann, Ian, Tahmassebi, Amirhessam, Foo, Simon Y., Erlebacher, Gordon, Meyer-Baese, Anke

Ian McCann, Amirhessam Tahmassebi, Simon Y. Foo, Gordon Erlebacher, Anke Meyer-Baese, "NewsAnalyticalToolkit: an online natural language processing platform to analyze news," Proc. SPIE 10653, Next-Generation Analyst VI, 106530P (27 April 2018); doi: 10.1117/12.2304646

SPIE.

Event: SPIE Defense + Security, 2018, Orlando, Florida, United States

NewsAnalyticalToolkit: An Online Natural Language Processing Platform to Analyze News

Ian McCann[a], Amirhessam Tahmassebi[a,*], Simon Y. Foo[b], Gordon Erlebacher[a], and Anke Meyer-Baese[a]

[a]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA
[b]Department of Electrical and Computer Engineering, FAMU-FSU College of Engineering, Tallahassee, Florida 32310-6046, USA

ABSTRACT

In today's increasingly divided political climate there is a need for a tool that can compare news articles and organizations so that a user can receive a wider range of views and philosophies. NewsAnalyticalToolkit allows a user to compare news sites and their political articles by coverage, mood, sentiment, and objectivity. The user can sort through the news by topic, which was determined using Natural Language Processing (NLP) and Latent Dirichlet Allocation (LDA). LDA is a probabilistic method used to discover latent topics within a series of documents and cluster them accordingly. Each news article can be considered a mix of multiple topics and LDA assigns a set of topics to each with a probability of it pertaining to that topic. For each topic, a user can then discover the coverage, mood, sentiment and objectivity expressed by each author and site. The mood was determined using IBM Watsons ToneAnalyzerV3, which uses linguistic analysis to detect emotional, social and language tones in written text. The analyzer is based on the theory of psycholinguistics, a field of research that explores the relationship between linguistic behavior and psychological theories. The sentiment and objectivity scores were determined using SentiWordNet, which is a lexical database that groups English words into sets of synonyms and assigns sentiment scores to them. The features were combined to plot an interactive graph of how opinionated versus how analytical an article is, so that the user can click through them to get a better understanding of the topic in question.

Keywords: Natural Language Processing, Sentiment Analysis, Tone Analyzing, Probabilistic Modeling, Latent Dirichlet Allocation.

1. INTRODUCTION

The history of NLP might date back to the 1950s when Alan Turing proposed a criterion of intelligence in his paper which was titled "Computing Machinery and Intelligence".[1] In summary, NLP focuses on the interactions between human language and computers. In this way, computers are able to analyze, understand, and derive meaning from human language based on the goals determined by developers. Tasks that can be organized and structured by NLP including automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.[2]

In 2014, Manning et al.[3] developed the Stanford CoreNLP Natural Language Processing Toolkit. The Stanford CoreNLP, is a JVM-based annotation pipeline framework, which provides most of the common core NLP steps including tokenization, sentence splitting, part-of-speech tagging, morphological analysis, named entity recognition, syntactic parsing, and conference resolution. This project was started in 2006, and led to a multi-site project in 2009, and was finally published as an open source frame work in 2014. In addition to the Stanford NLP, there are different NLP libraries that are available such as Apache OpenNLP, Natural Language Toolkit (NLTK), and MALLET.

* Corresponding Author: Amirhessam Tahmassebi
E-mail: atahmassebi@fsu.edu
URL: www.amirhessam.com

Next-Generation Analyst VI, edited by Timothy P. Hanratty, James Llinas, Proc. of SPIE Vol. 10653, 106530P · © 2018 SPIE · CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2304646

The aim of this project is to provide users with tools to examine political news and the outlets that produce it. In a political climate growing ever more divided, we believe it is important have at one's disposal the tools that enable the comparison of news articles and organizations.

We have developed tools as an online framework, "`www.NewsAnalyticalToolkit.com`" that allow a user to compare news sites by their mood, sentiment, and objectivity toward certain topics. A user can pick for reading an analytical piece, opinionated piece, neither or both on a certain topic. This allows one to be informed about the political topics in the news, how they are covered and in what way by different outlets. The hope is that a user can view all sides of an issue and come to their own conclusion about how they feel towards it. We have extracted the topics from a weeks worth of news articles and determined their bias score using Equation 1, where the positive, negative and objective values were determined by the sentiment library and the word probability is the probability the word pertained to that topic and was determined by LDA.

$$BiasScore = [PositiveSentiment + NegativeSentiment] \times [1 + ObjectiveScore] \tag{1}$$

2. MATERIALS & METHODS

In this section, we explain the details of the development of the database, including how the information of the database was parsed.

2.1 Data

The data consisted of articles gathered from the Rich Site Summaries (RSS) feeds of twelve different websites. Those sites with their associated RSS links were CNN[1], ABC[2], FOX[3], NYT[4], Reuters[5], Washington Post[6], Huffington Post[7], Esquire[8], Rolling Stone[9], CBS[10], FiveThirtyEight[11], and The Washington Times[12]. Every hour, the articles linked from each RSS feed were scraped and saved to a Mongo database on an Amazon Web Services (AWS) server. The data was then converted to a file with a comma-separated-value (CSV) format and stored on a S3 bucket. Additionally, all past articles from the Wall Street Journal (WSJ) were available, so they were scraped as well for possible use in future projects.

The first step in the pre-processing was to remove stop-words. The stop-words are words such as "the", "a", "I", "him", etc. Next, we created bi-grams, tri-grams and quad-grams. These terms are new words that are combinations of words that are commonly juxtaposed. An example of a quad-gram extracted from the articles is "FBI Director James Comey". This would help with topic model and for interpretation when viewing the words in a topic is desired. Next, we lemmatized the words. This involved removing inflectional endings, thus returning it to its base form. An example is changing the word "Working" to "Work". This helps with topic modeling and interpretation. Finally, we separated quotes and tweets from an article to derive what is called "Sentiment Texts" to be used together with sentiment analysis. The purpose of this step was to only consider words written by the authors themselves when applying sentiment analysis. As a future project we will examine how different news sites use quotes and tweets.

In addition to this, when performing topic modeling there were words that can appear in all topics other than stop-words. Since, we have examined political articles, there were several words that appeared often, such

[1] https://www.cnn.com/
[2] http://abcnews.go.com/
[3] http://www.foxnews.com/
[4] https://www.nytimes.com/
[5] https://www.reuters.com/
[6] https://www.washingtonpost.com/
[7] https://www.huffingtonpost.com/
[8] https://www.esquire.com/news-politics/
[9] https://www.rollingstone.com/politics
[10] https://www.cbsnews.com/politics/
[11] https://fivethirtyeight.com/politics/
[12] https://www.washingtontimes.com/news/politics/

as "Trump", "President", "Election", "Politics", etc. These words can reduce interpretability of topics, so we decided to remove words that appeared in 50% or more of the articles and words that appeared in less than 20% of them. The latter step was to ensure mis-spelled words or obscure words were not included in the topic modeling.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative three-level Bayesian model proposed by Beli et al.[4] It is usually used for collection of discrete data modeled as a finite mixture over a set of features.[5–8] An infinite mixture set of feature probabilities has been used to model each feature in turn.[9] It turns out that the feature probability is an explicit representation of the reduced data. In a joint distribution model, α and β are smoothing parameters that are sampled only once.[9] However, parameter θ will be sampled once each generation. In addition, variables z and w are sampled once for each feature to calculate the posterior distribution for each parameter.[10,11] Figure 1 presents the graphical model representation of the LDA model.

$$p(z, \theta, \beta | w, \alpha, \eta) = \frac{p(z, \theta, \beta | \alpha, \eta)}{p(w | \alpha, \eta)} \tag{2}$$

In other words, LDA is a probabilistic method used to discover latent topics within a series of documents and cluster them accordingly. LDA was used in this project to determine topics in political news sources. Each news article can be considered a mix of multiple topics and LDA assigns a set of topics to each along with a probability of it pertaining to that topic. Each topic has a set of words together with probabilities of being related to it. Articles with a high frequency of words that have high probabilities of being in a topic will themselves have a high probability of belonging that topic. The assumption is that the articles cover a small set of topics and the topics use a small set of words frequently.

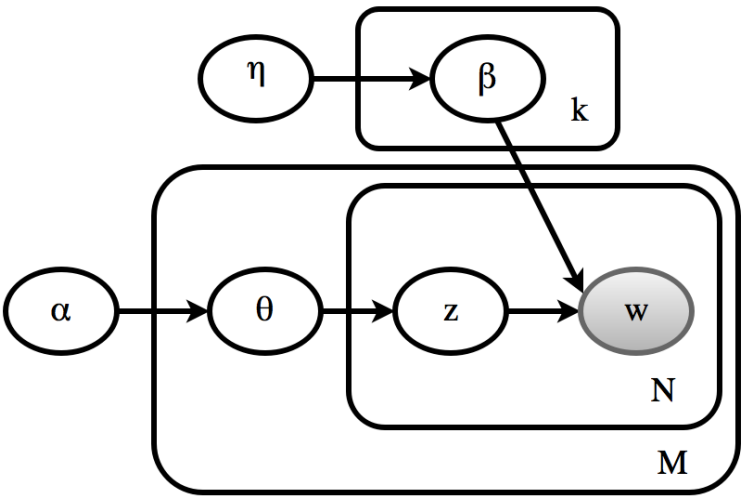

Figure 1. A schematic representation of the LDA model.[4,10,11]

LDA requires to input how many topics are desired prior to running the model. Since we were using all the articles that can be extracted from the news source and the number of topics discussed can be arbitrary, the number of topics needs to be determined. There were several attempts to approach this decision in a scientific manner, but ultimately we considered several models with varying topics and picked the one that had the most distinguishable topics. We believe that this was an acceptable solution because the attempts to scientifically find a number of topics showed that there was a wide range of acceptable values. The first attempt was to use Hierarchical Dirichlet Process (HDP)[12] LDA, which is an extension of LDA designed to address the issue of not knowing the number of topics beforehand. The goal of using HDP was to deduce the numbers of topics from the

data. The HDP was used to calculate 150 topics and the probability that each of those topics had in regards to the articles was determined accordingly. Unfortunately, an obvious choice for the number of topics using this method has not been made. All topics had less than 10% probability and decreased at a slow rate to around 5%. The hope was to see a steep drop in the probability and then an "Elbow", where the probability leveled off which would lead to a natural choice for the number of topics to prescribe. The second attempt to resolve the discussed issue was using a "Coherence" metric. The idea is that some topics were more coherent than others or in other words, some topics were more interpretable than others. To determine a topic's coherence, a model using the data collected from people who were told to rate a topic's coherence was created.[13] This metric was used on the LDA models which were created from topic numbers from 5 to 80. The coherence value did not vary much between any of the models.

2.3 SentiWordNet

SentiWordNet[14,15] is a lexical database that groups English words into sets of synonyms and assigns sentiment scores to them. These scores are positive sentiment, negative sentiment and objectivity. They are valued from 0 to 1. For all sentiment words found in the sentiment texts, the sentiment scores of all synonyms for that word and their mean value were calculated. Next, the summation of the words' scores was calculated for each article to get the positive, the negative, and the objective scores for each article. Finally, a "Sentiment Score" was defined as shown in equation 3 which could be used to relate these articles:

$$SentimentScore = [PostiveSentiment + NegativeSentiment] \times [1 - ObjectiveScore] \tag{3}$$

This conclusion was drawn to sum up the positive and the negative sentiment since a word can have both a positive and a negative score. In addition to this, it should be noted that an article's positive and negative scores were found to be linearly related. It is believed that if an author puts more sentiment or emotion into an article, the positive and negative scores tend to go up. Thus, the sentiment score should reflect this issue instead of the negative and positive scores canceling each other throughout subtraction. Next, the subjectivity was considered to be one minus the objectivity. In fact, if the author was objective, their sentiment score should be low. Thus, the scores were multiplied by the subject score.

Figure 2. Basic emotions: sadness, joy, fear, disgust and sadness which were personified by Disney PIXAR in the movie "Inside Out"[13].

2.4 ToneAnalyzerV3

IBM' Watson has a Tone Analyzer[14] service that uses linguistic analysis to detect emotional, social and language tones in written text. The analyzer is based on the theory of psycholinguistics,[16] a field of research that explores

[13] http://movies.disney.com/inside-out
[14] https://www.ibm.com/watson/services/tone-analyzer/

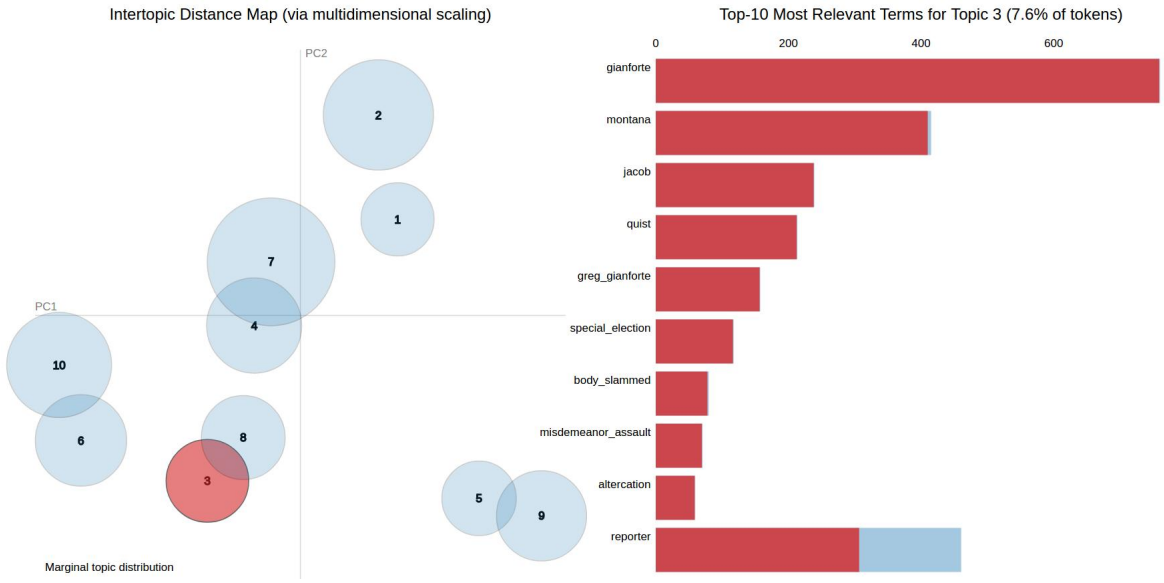

Figure 3. Left Figure: Intertopic distance map via multidimensional scaling considering the marginal topic distribution employing the first and second principal components (PC1 and PC2). Right Figure: Top-10 most relevant terms for topic 3 (7.6% of tokens).

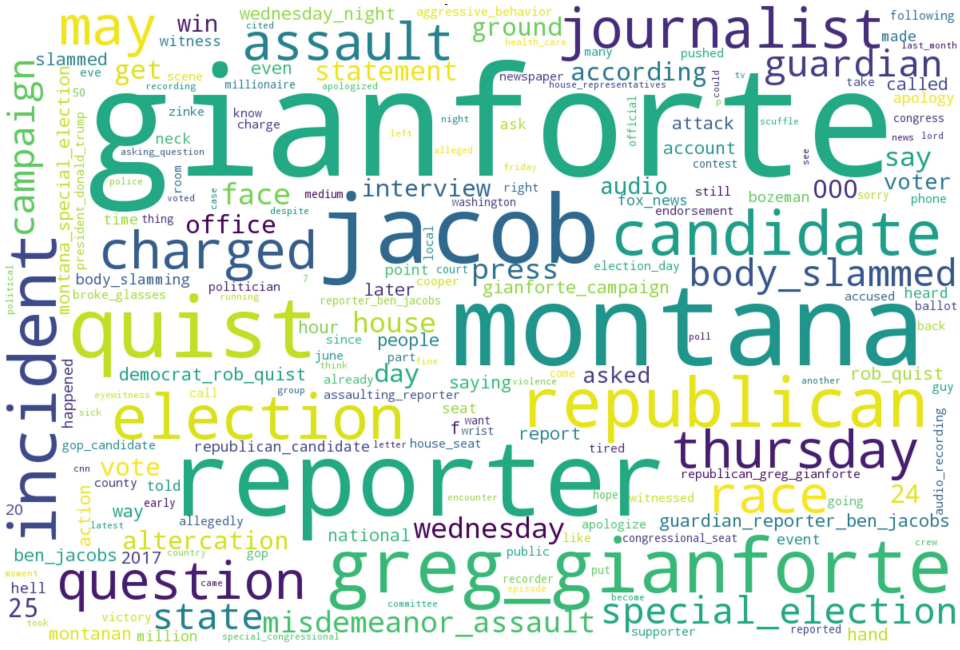

Figure 4. Topic 30 world cloud.

the relationship between linguistic behavior and psychological theories. The service uses linguistic analysis and the correlation between the linguistic features of written text and emotional, social, and language tones to develop scores for each of the aforementioned tone dimensions. The emotional tones are sadness, joy, fear, disgust, and sadness, which are the basic emotions that are classified by the psychologist, Paul Ekman[17,18] and were later personified as the characters shown in Figure 2 in the Disney PIXAR movie "Inside Out"[13]. IBM's emotional tones were derived from an IBM classification framework[15]. The emotional tones were used in this project to

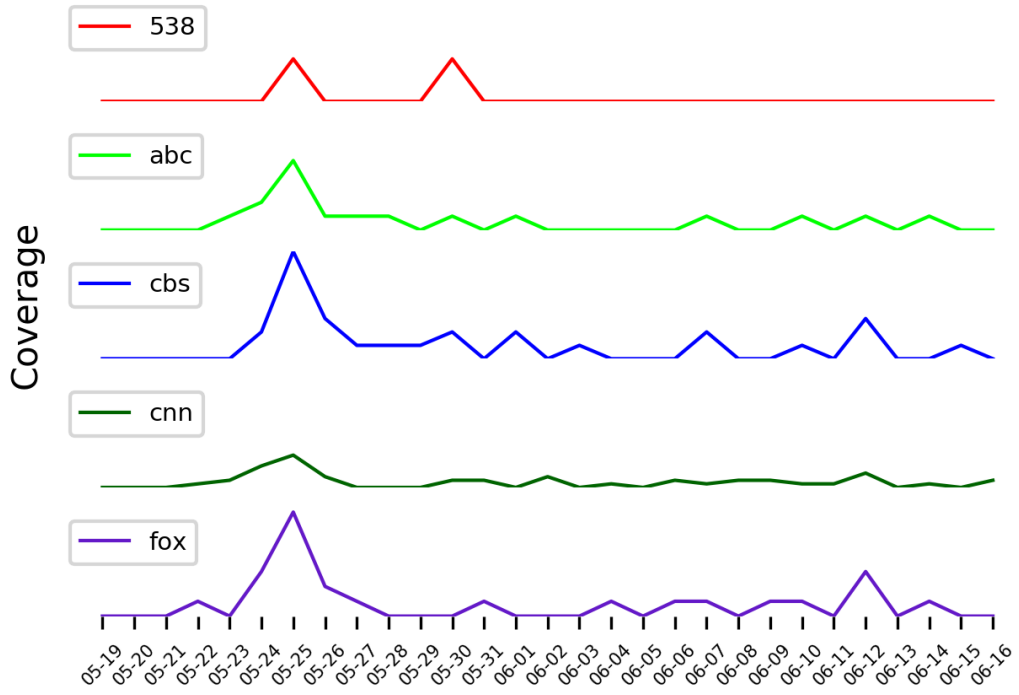

Figure 5. Coverage of topic 30 by site.

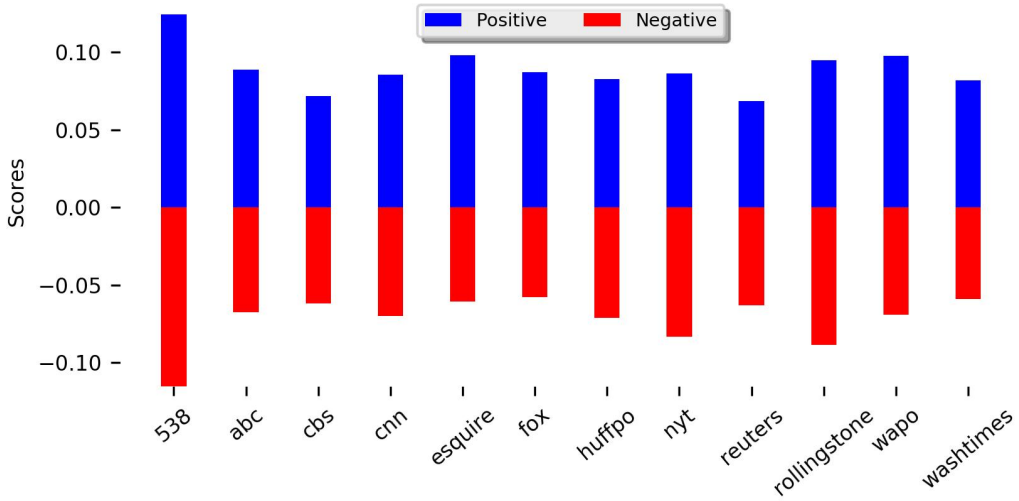

Figure 6. Positive/Negative scores by site for topic 30.

determine the mood of a topic. Each article was run through the Tone Analyzer and for each topic were summed together to derive mood scores for each site. Figure 7 shows an illustration of the mood scores by site.

3. RESULTS & DISCUSSION

To illustrate the performance of the framework, "Topic 30" was chosen and all the figures are related to this topic. Topic 30 concerns Greg Gianforte, who won the U.S. House special election on May 25th, 2017. The day

[15] https://console.bluemix.net/docs/services/tone-analyzer/index.html

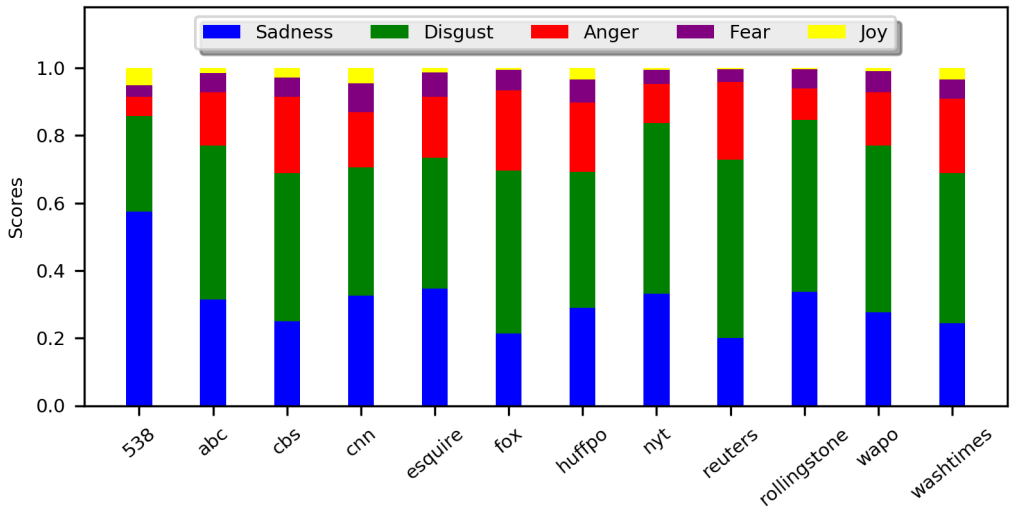

Figure 7. Stacked mood sources by site for topic 30.

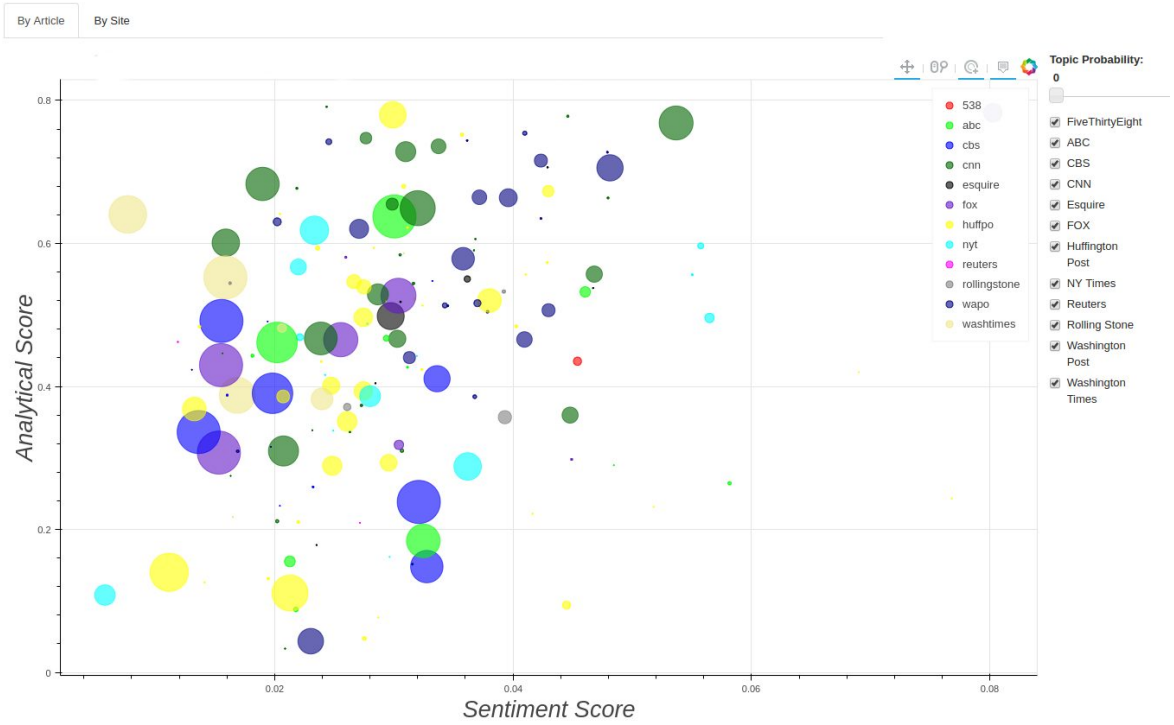

Figure 8. Analytical score by sentiment score for topic 30.

before being elected, he was accused of body-slamming a reporter and on June 12, 2017, he pleaded guilty to the assault charge.

Figure 3 presents a visualization of the topics generated using LDA for 10 topics discussed the week from May 21st, 2017 to May 28th, 2017. This was done to better illustrate topic 30 from the 55 topics that were created in the main LDA model used in this project. Figure 3 (Left) illustrates the intertopic distance map via multidimensional scaling considering the marginal topic distribution employing the first and second principal components (PC1 and PC2). In addition to this, Figure 3 (Right) illustrates the top-10 most relevant terms for

topic 3 which contains 7.6% of the tokens. Figure 4 shows a word cloud constructed from the words found in the topic based on their probabilities. You can see there are many large words in the cloud also included in the top relevant terms previously discussed.

Figure 5 presents the coverage by date of the sites that mostly covered this topic. One clearly sees a rise in coverage beginning near the 24th (the day of the assault) and peaking on the 25th. There is another spike on June 12th (the day of his trial). Moreover, to present the topic sentiment, Figure 6 shows the average positive and negative sentiment scores by site for Topic 30.

Figure 7 shows a stacked bar chart of the five emotional tone scores by site. We found that the common mood was disgust. In addition, we used Bokeh[16] (a Python interactive visualization library) to present interactive sentiment figures. Figure 8 shows a single snapshot, however, the interactive version is available on the website. In this way, all the articles related to Topic 30 were illustrated based on their analytical scores versus their sentiment scores. The size of the circle indicates the probability that the article relates to this topic and the slider on the top right allows the user to set a probability threshold that the article must reach to be shown. Furthermore, the check-boxes allow the user to select which news websites to view. In addition, clicking on any of the circles opens the related article for the user to read. There are also two tabs "By Article" and "By Site". Figure 8 presents the plot by article. However, by clicking the "By Site" tab, the user can examine how each site scored on average, as well as, their standard deviation.

4. CONCLUSIONS

NewsAnalyticalToolkit allows a user to compare news sites and their political articles by coverage, mood, sentiment, and objectivity. The platform was constructed from Natural Language Processing, and Latent Dirichlet Allocation. The mood was determined using IBM Watsons ToneAnalyzerV3, the analyzer was based on the theory of psycholinguistics, and the sentiment and objectivity scores were determined using SentiWordNet. The features were combined to plot an interactive graph of how opinionated versus how analytical an article is so that the user can click through them to gain a better understanding of the topic in question.

5. SUPPLEMENTARY MATERIALS

All of the developed codes are available at `http://www.NewsAnalyticalToolkit.com/About`.

6. ACKNOWLEDGEMENTS

This work is partially supported by ONR Grant "Gulf of Mexico Spring School (GMSS) Deep Learning Workshop".

7. CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this article.

[16] https://bokeh.pydata.org/

REFERENCES

[1] Hutchins, J., "The history of machine translation in a nutshell," *Retrieved December* **20**, 2009 (2005).

[2] Jurafsky, D., "Speech and language processing: An introduction to natural language processing," *Computational linguistics, and speech recognition* (2000).

[3] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D., "The stanford corenlp natural language processing toolkit," in [*Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*], 55–60 (2014).

[4] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent dirichlet allocation," *Journal of machine Learning research* **3**(Jan), 993–1022 (2003).

[5] Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H., Schmaal, L., Goudriaan, A. E., and Meyer-Baese, A., "An evolutionary approach for fmri big data classification," in [*Evolutionary Computation (CEC), 2017 IEEE Congress on*], 1029–1036, IEEE (2017).

[6] Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H., Schmaal, L., Goudriaan, A. E., and Meyer-Baese, A., "fmri smoking cessation classification using genetic programming," in [*Workshop on Data Science meets Optimization*], (2017).

[7] Tahmassebi, A., Gandomi, A. H., and Meyer-Bäse, A., "High performance gp-based approach for fmri big data classification," in [*Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*], 57, ACM (2017).

[8] Tahmassebi, A., "ideeple: Deep learning in a flash," in [*Disruptive Technologies in Information Sciences*], **10652**, International Society for Optics and Photonics (2018).

[9] Blei, D. M., "Probabilistic topic models," *Communications of the ACM* **55**(4), 77–84 (2012).

[10] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G., "API design for machine learning software: experiences from the scikit-learn project," in [*ECML PKDD Workshop: Languages for Data Mining and Machine Learning*], 108–122 (2013).

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[12] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M., "Sharing clusters among related groups: Hierarchical dirichlet processes," in [*Advances in neural information processing systems*], 1385–1392 (2005).

[13] Röder, M., Both, A., and Hinneburg, A., "Exploring the space of topic coherence measures," in [*Proceedings of the eighth ACM international conference on Web search and data mining*], 399–408, ACM (2015).

[14] Esuli, A. and Sebastiani, F., "Sentiwordnet: a high-coverage lexical resource for opinion mining," *Evaluation* , 1–26 (2007).

[15] Baccianella, S., Esuli, A., and Sebastiani, F., "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.," in [*LREC*], **10**(2010), 2200–2204 (2010).

[16] Mandelbrot, B., "Information theory and psycholinguistics," *BB Wolman and E* (1965).

[17] Ekman, P. E. and Davidson, R. J., [*The nature of emotion: Fundamental questions.*], Oxford University Press (1994).

[18] Ekman, P., "An argument for basic emotions," *Cognition & emotion* **6**(3-4), 169–200 (1992).