Proceedings of Indian River Lagoon Symposium 2020

Opinion Mining of Newspaper Articles using Natural Language Processing: Pilot Test Using Texts on Indian River Lagoon

Manit Bhusal[(1)], Juan Calderon[(1)], and Hyun Jung Cho[(2)]

[(1)]Department of Computer Engineering, Bethune-Cookman University, 640 Dr. Mary McLeod Bethune Blvd, Daytona Beach, FL 32114
[(2)]Department of Integrated Environmental Science, Bethune-Cookman University, 640 Dr. Mary McLeod Bethune Blvd, Daytona Beach, FL 32114

Keywords Artificial Intelligence, Deep Learning, Indian River Lagoon, Natural Language Processing, Neural Network, Opinion Mining

Introduction

Background and Study Goal. News articles play a critical role in developing personal or societal visions and opinions on a given issue or event by creating positive or negative reactions (Rameshbhai and Paulose 2019). These reactions can vary based on the subject matter and on the biases and beliefs of the authors and readers. For example, articles covering environmental issues can be viewed with skepticism, particularly when the issues are politicized. With the progress of the internet age, the abundance of data available, and the advances in computational capability, Artificial Intelligence (AI) technology using Natural Language Processing (NLP) has been applied in opinion mining using texts (Pak and Paroubek 2010, Kouloumpis et al. 2011).

The Indian River Lagoon (IRL) system in Florida has experienced increasing phytoplankton blooms associated with elevated nutrient levels stemming from various sources (Lapointe et al. 2015). Numerous media articles have been published that document the blooms, discuss associated environmental and economic impacts through interviews with local experts and stakeholders, track potential sources and triggers, and predict future trends. These articles contain both facts and opinions about the algal blooms and their impacts on the IRL and surrounding communities. Assuming the article contents would influence the public's opinion and judgement over the environmental issue, it is important to develop a tool for opinion mining (e.g. facts vs. opinion; subjective vs. objective; positive, neutral, vs. negative) of the large text datasets.

This paper presents a study that classified sentences from newspaper articles on the health of the IRL in relation to recent algal blooms. The study goal was to test if

Corresponding Author: Hyun Jung Cho, choh@cookman.edu

AI technology can be utilized to detect any subjectivity from the sentences that can lead to biased interpretations.

Relevant Technological Applications. Machine-learning (ML) technology, a branch of AI, is used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant search results (LeCun et al. 2015). NLP is the ML technology linked with human natural language, including texts; and its ultimate objective is to read, decipher, understand, and make sense of the human languages in a manner similar to that of a human brain. NLP combined with Deep Learning have been making technological breakthroughs such as text-prediction, text-correction, text analysis, speech-to-text, text-to-speech, and voice commands. The NLP-based approach has been tested in opinion mining to enhance sentiment classification (Kim et al. 2014; Kanakaraj and Guddeti 2015, Sun et al. 2016, Cenni et al. 2017, Rameshbhai and Paulose 2019). Mining opinions and sentiments from texts and natural language requires deep learning of the language rules. Effective systems have been designed for common tasks like bag-of-words variants for information retrieval (Kanakaraj and Guddeti 2015), syntactic information (e.g., part-of-speech tagging, chunking, and parsing) or semantic information (e.g., word-sense disambiguation, semantic role labeling, named entity extraction, and anaphora resolution; Collobert et al. 2011).

AI-based approaches are beginning to be used to perform automated text analyses related to environmental issues. Jacques and Connolly (2016) analyzed Twitter messages to better understand why a certain sector of the population refuses to believe in the imminent emergency described by the climate change discourse.

Methods

This research tests a method of opinion mining using Deep Learning and NLP to analyze online newspaper articles on topics related to the IRL algal blooms by classifying the sentences into objective or subjective. The method uses four classification models that are based on Artificial Neural Networks (ANN): Convolutional Neural Network (CNN), CNN with Glove Word Embedding (GloVe), Long Short-Term Memory (LSTM), and LSTM with GloVe. The methodology is divided into 4 stages: (1) preprocessing data, (2) encoding and labeling data, (3) ANN models training, and (4) training validation and database adjustment.

One of important factors that affects the ML algorithm performance is the database used for the training process. The training database should contain a full array of samples of classified sentences and should not be biased toward a specific category. No appropriate database was available at the time of this study which contained relevant environmental texts tagged with objectivity vs. subjectivity. Therefore, a combination of datasets from different sources was used as a training database. For objective sentence training datasets, sentences from encyclopedias were used; and for subjective sentences, datasets from Twitter.com and movie reviews were used, taking advantage of the resources that are rich in opinion-based sentences (Kouloumpis et al. 2011). Since the model is trained to classify sentences from media articles regarding local issues, the Daytona Beach News Journal (https://www.news-journalonline.com/) was used as a primary online news source for the test database.

Pre-processing the Data. Cleaning data is an important step to improve the performance of the trained model. The cleaning steps included lowercasing each word, removing stop words (e.g. a, an, the, is, am, are, was, were, has, have, etc.), as well as removing mentions, URLs, and symbols. Stop words could affect the performance of the trained classifier, however, words like "not", "never" are kept in the

whitelist. This approach of removing the stop words and keeping the negation terms proved useful in improving the efficiency of the model (Pak and Paroubek 2010).

As an example, two sentences, denoted by S_1 and S_2, are shown below:

S_1= "*Earth is our mother, our common home. Earth is covered with about 71% of water. Most of the water found on the earth is salty.*"
S_2= "*earth mother, common home. earth covered about 71% water. most water found earth salty*"

where, S_1 are the sentences from the original text and S_2 is the cleaned version of the sentences.

Encoding and Labeling Data. The cleaned data were processed through an encoder function which transforms the words into numbers. One way of encoding the words is to give a number to each word based on the frequency of its occurrence. Additionally, each word is ranked according to its order of appearance in the text. Below is an example: the sentence S_2 is encoded using the calculation of the frequency and ranking as depicted next.

Word	Frequency	Ranking
Earth	3	1
water	2	2
mother	1	3
common	1	4
home	1	5
covered	1	6
about	1	7
71	1	8
most	1	9
found	1	10
salty	1	11

The encoded version of S_2 is: [[1, 3, 4, 5], [1, 6, 7, 8, 2], [9, 2, 10, 1, 11]], where each vector represents each part of S_2 separated by a comma. Since there is a variable length of data in both vectors, the vectorial calculations become complex. So, the encoded sentences are padded with zeros to yield vectors of equal length, which facilitates the calculation. In this example, the padded vectors are [[0, 1, 3, 4, 5,], [1, 6, 7, 8, 2], [9, 2, 10, 1, 11]]. The target classes or labels are the output values for each input, which are encoded using the one-hot encoding technique. In this study, the encoding for the target classes is:

Objective Sentence: [1, 0]
Subjective Sentence: [0, 1]

ANN Models Training. The encoded sentences and labeled data were split into training and testing sets, with 90% of the data in the training set and 10% in the testing set. In addition, 10% of the training set was used to validate the model as the training process was underway.

ANN, the main unit of the classification model, is based on the principle of how human neurons work. In the past, NLP methods were used alone for text classification; however, such methods had limitations for short texts. The use of CNN helps to overcome these limitations via advanced text representation models (Wang et al. 2017). In addition, Long Short-Term Memory models (LSTM) have outperformed other models for sequential data like text (Nowak et al. 2017). In this study, CNN and LSTM were trained with a pre-trained GloVe word embedding vector to create four models that were compared for accuracy: CNN, CNN with GloVe, LSTM, and LSTM with GloVe. The training process was performed by 3 epochs per model. An epoch is a variable factor defined as one iteration of going through the entire dataset.

Training Validation and Database Adjustment. After a model was trained, it was tested using the testing data. The accuracy of the model was calculated to evaluate the performance in the training

Table 1. Confusion matrix of the 4 trained models.

Predicted Values	Actual Values							
	CNN Model		CNN with GLOVE		LSTM		LSTM with GLOVE	
	Objective	Subjective	Objective	Subjective	Objective	Subjective	Objective	Subjective
Objective	2423	52	2403	72	2446	67	2477	36
Subjective	69	2413	53	2429	72	2382	72	2382

process. Then, the trained model was used to classify newspaper sentences. The newspaper articles were broken down into sentences. The tokenized sentences were then passed through the pre-processing and encoding in a similar way described for the training data. The encoded sentences were passed through the neural network model to be classified into objective or subjective sentences.

Results and Discussion

The four models were implemented and evaluated through the training and testing processes. The confusion matrix was calculated for every classification model (Table 1). Four values were calculated in the confusion matrix: True Positive (TP= (Objective, Objective)), False Positive (FP= (Subjective, Objective)), False Negative (FN= (Objective, Subjective)), and True Negative (TN= (Subjective, Subjective)). Using the values from the confusion matrix, the precision (Equation 1), recall (Equation 2), and accuracy (Equation 3) of each model were calculated (Table 2).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

All four models scored over 0.97 (out of 1.00) in precision, recall, and accuracy, with the model based on LSTM with GloVe showing the top score. In this study, limiting the model to learn on a particular topic made it possible for the classifier to score high for precision, recall, and accuracy. However, the classifier was trained using the environmental encyclopedia sentences for the objective

Table 2. Results of accuracy assessments of the four models.

Model	Precision	Recall	Accuracy
CNN	0.979	0.972	0.976
CNN with GLOVE	0.971	0.978	0.975
LSTM	0.973	0.971	0.972
LSTM with Glove	0.986	0.983	0.984

sentence label and Twitter and movie review datasets for subjective label. Selection of discipline-specific appropriate sentimental words is needed and can be improved through the development of an environmental-specific dictionary via the NLP (Kim et al. 2014). Still, it can be concluded that the model performance was satisfactory, promising that performance will improve if a well-classified environmental dataset is available and a domain-specific dictionary is developed.

Classification of a sentence into objective or subjective is a mere start to developing a computerized system for analyzing textual data focused on environmental topics. Future works include sentimental analysis, and the combination of sentimental analysis and objectivity classification. Then, the complete system of sentimental analysis, objectivity classification, opinion mining, and fake detector can help to make a powerful computerized system that can analyze any textual data. Furthermore, a system can be developed allowing a machine to scan a newspaper, generate the main topics and issues described in the newspaper, and display the sentiment (optimistic vs. pessimistic) toward the topic. This new classification will be based on the use of additional sentiment analysis models in charge of classifying the subjective sentences provided by the current classification system.

Acknowledgements This publication was made possible by a Section 319 Nonpoint Source Management Program Implementation grant from the U.S. Environmental Protection Agency through an agreement/contract with the Nonpoint Source Management Section of the Florida Department of Environmental Protection as well as the National Oceanic and Atmospheric Administration, Office of Education Educational Partnership Program award (NA16SEC4810009). Funding for graduate research assistantship was funded by the National Oceanic and Atmospheric Administration (NOAA) through the NOAA- Center for Coastal and Marine Ecosystems graduate assistantship.

Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Environmental Protection Agency, Florida Department of Environmental Protection, nor U.S. Department of Commerce, National Oceanic and Atmospheric Administration.

References

Cenni D, Nesi P, Pantaleo G, Zaza I. 2017. Twitter Vigilance: a multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis. The 2017 IEEE SmartWorld, pp. Pp. 1–8 *in* Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation.

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. 2011. Natural language processing (almost) from scratch. Journal of Machine Learning Research 12:2493–2537.

Jacques PJ, Knox CC. 2016. Hurricanes and hegemony: A qualitative analysis of micro-level climate change denial discourses. Environmental Politics 25:831–852.

Kanakaraj M, Guddeti RMR. 2015. NLP based sentiment analysis on Twitter data using ensemble classifiers. Pp. 1-5 *in* The 3rd International Conference on Signal Processing, Communication and Networking.

Kim Y, Jeong S, Ghani I. 2014. Text opinion mining to analyze news for stock market prediction. International Journal of Advances in Soft Computing and its Applications 6: 2074–8523.

Kouloumpis E, Wilson T, Moore J. 2011. Twitter sentiment analysis: The good the bad and the omg! Pp. 538 –541 *in* Proceedings of the Fifth International Conference on Weblogs and Social Media.

Lapointe BE, Herren LW, Debortoli DD, Vogel MA. 2015. Evidence of sewage-driven eutrophication and harmful algal blooms in Florida's Indian River Lagoon. Harmful Algae 43:82–102.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature 521:436–444.

Nowak J, Taspinar A, Scherer R. 2017. LSTM recurrent neural networks for short text and sentiment classification. Pp. 553–562 *in* International Conference on Artificial Intelligence and Soft Computing.

Pak A, Paroubek P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation 10:1320–1326.

Rameshbhai CJ, Paulose J. 2019. Opinion mining on newspaper headlines using SVM and NLP. International Journal of Electrical & Computer Engineering 9:2088–8708.

Sun S, Luo C, Chen J. 2016. A review of natural language processing techniques for opinion mining systems. Information Fusion 36:10–25.

Wang J, Wang Z, Zhang D, Yan J. 2017. Combining knowledge with deep convolutional neural networks for short text classification. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence 350:2915–2921.

Submitted: August 18, 2020
Accepted: December 12, 2020