Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 3 (2021) 477–483

Journal of the Nigerian Society of Physical Sciences

Performance Study of N-grams in the Analysis of Sentiments

O. E. Ojo*, A. Gelbukh, H. Calvo, O. O. Adebanji

[a]*Instituto Politécnico Nacional, Natural Language and Text Processing Laboratory, Centro de Investigacion en Computación, CDMX, Mexico*

Abstract

In this work, a study investigation was carried out using n-grams to classify sentiments with different machine learning and deep learning methods. We used this approach, which combines existing techniques, with the problem of predicting sequence tags to understand the advantages and problems confronted with using unigrams, bigrams and trigrams to analyse economic texts. Our study aims to fill the gap by evaluating the performance of these n-grams features on different texts in the economic domain using nine sentiment analysis techniques and found more insights. We show that by comparing the performance of these features on different datasets and using multiple learning techniques, we extracted useful intelligence. The evaluation involves assessing the precision, recall, f1-score and accuracy of the function output of the several machine learning algorithms proposed. The methods were tested using Amazon, IMDB, Reuters, and Yelp economic review datasets and our comprehensive experiment shows the effectiveness of n-grams in the analysis of sentiments.

DOI:10.46481/jnsps.2021.201

Keywords: n-grams, machine learning, deep learning, sentiment analysis

Article History :
Received: 3 July 2021
Received in revised form: 13 September 2021
Accepted for publication: 14 September 2021
Published: 29 November 2021

Communicated by: J. Ndam

1. Introduction

Machine learning and deep learning architectures are the bane of many Natural Language Processing (NLP) research works. To address a variety of tasks, including sentiment analysis, several machine learning and deep learning architectures have been proposed. Uncommon and unfamiliar words used in information and knowledge exchange can influence different aspects of life including marketing, education, governance, etc. As an integral part of the internet, the digital media platforms facilitates meaningful information and knowledge exchange with a list of other network users. Data collection and reviews, with diverse views and opinions about events, is gaining more impact and fast becoming an attraction for researchers and generating significant computational challenges. Effective wide-ranging mining of information from text helps to discover useful knowledge of vital significance. Computers can detect, interpret and produce the sentiments (or tags) of a text, thereby improving government and private companies' operations, recognizing possible threats, minimizing crime, and improving public services.

The main objective of this research is to observe the efficacy of unigrams, bigrams and trigrams as characteristics of a word sequence and to predict a tag for sentiment classification. The target is to extract words or phrases behind the tags and to use the machine learning and deep learning methods to classify the data whilst measuring the accuracy of the classification. Using n-grams to study the opinion of people, we will be able to see their strength by tagging them. As they contribute to conceptual

*Corresponding author tel. no: +525560590794
Email addresses: olumideoea@gmail.com (O. E. Ojo), gelbukh@gelbukh.com (A. Gelbukh), hcalvo@cic.ipn.mx (H. Calvo), olaronke.oluwayemisi@gmail.com (O. O. Adebanji)

characterization, we use machine learning and deep learning models on the text.

In machine learning, two major techniques are adopted: supervised learning [1, 2, 3, 4] and unsupervised methods [5]. Supervised methods have a training dataset with manually defined tags, and they learn the characteristics that match the tags from the training data. Gelbukh and Kolesnikova [2, 6] developed methods that allowed the automatic sorting of word combinations into pre-established categories relating to automated collocation classification. Gambino and Calvo [7] considered the use of text-learned opinions using NLP techniques to distinguish tags. On the other hand, unsupervised systems are more flexible across different kinds of texts and domains. Different machine learning algorithms have been used in past works [6, 2, 4], and also neural network models [1] to gain the knowledge of how to predict the sentiments of text [3, 8]. Pre-trained models are very helpful in classifying text and other NLP activities.

To extract the keywords behind the text's feelings, the models will be pre-trained on the training data and the accuracy of prediction of the models will be measured, registered and compared. The remaining part of the paper is organized as follows: Section 2 deals with the background and relevant works, Section 3 describes the approach used in this study, while Section 4 shows the features we experimented. Sections 5 and 6 provides the information about the machine learning and deep learning algorithms used and our experimental findings with the discussion of results in section 7. Section 8 gives conclusion about the work.

2. Background and Related Work

Different works have been carried out in the field of sentiment analysis [1, 7, 3, 4]. The social media and other digital media platforms, as an accepted means of communication, has flourished thereby aiding intelligence gathering and information dissemination. Machine learning techniques have shown good results in analysing sentiments in text [6, 2, 4] and other tasks such as part of speech recognition (PoS) [9], named entity recognition (NER) [10], etc. Linear statistical models, such as random-field (CRF) and Hidden Markov (HMM) fields, are NLP approaches used for sequence tagging with a long history of excellent performance. However, adapting these models to new tasks in new domains or languages is challenging.

The combination of categorical grammar, annotation, acquisition of lexicons and semantic networks was used by Pekka et al. [11] to analyze the feelings of the text and to define the tags of the text. They investigated how the overall phrase structured data and domain-specific language usage could aid in the detection of semantic orientations in financial and economic news.

In [12], the use of syntactic n-grams (Sn-grams) to incorporate syntactic knowledge into machine learning algorithms proved successful. Sn-grams were utilized as a baseline for authorship identification, replacing standard n-grams of words, POS tags, and characters.

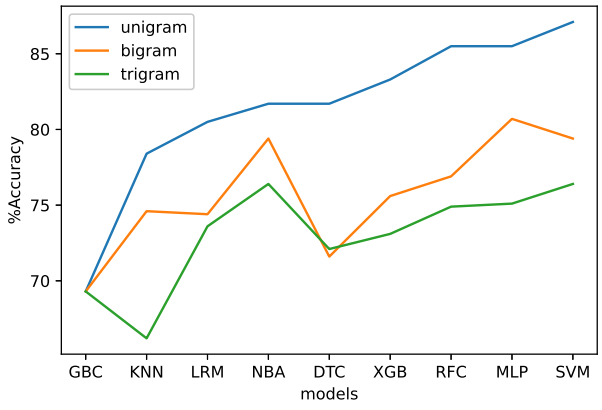

Figure 1. Distribution of n-grams accuracy scores in the Models for the first dataset

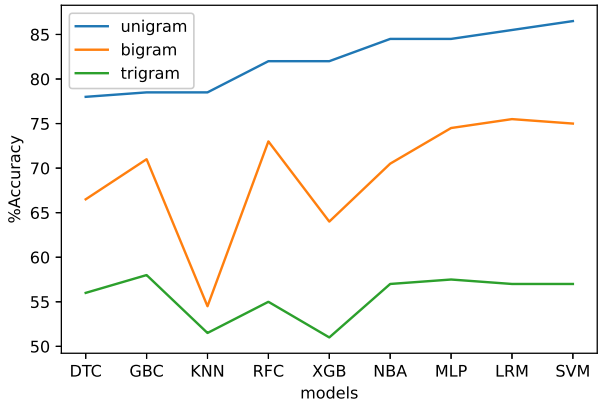

Figure 2. Distribution of n-grams accuracy scores in the Models for the second dataset

[13] have used text-CNN for extraction of text features with LSTM architecture in addition to the unimodal input functions. In a multimodal sentiment analysis task, they explored and analyzed the performance of three deep-learning-based architectures and recorded their results.

On a social media data baseline, [14] explored the efficiency of deep neural network models of different complexity based on character n-grams. The training was done with augmented data and pseudo-labeled samples, and the accuracy result was enhanced.

[15] also used classifiers to predict sentence tags using an objective function to infer similarity between sentences. A new objective function was used to train many classifiers to make predictions at the instance level, promoting smoothness of inferred instance-level labels while keeping group-level label constraints in place.

Table 1. Precision, Recall, f1 score and accuracy of the classifiers trained on the first dataset.

Model	n-grams	Precision	Recall	F1 score	Accuracy
Logistic Regression	Unigram	0.85	0.69	0.72	80.5%
	Bigram	0.81	0.59	0.58	74.4%
	Trigram	0.86	0.57	0.54	73.6%
Support Vector Machine	Unigram	0.88	0.81	0.83	87.1%
	Bigram	0.79	0.70	0.72	79.4%
	Trigram	0.78	0.64	0.65	76.4%
Naive Bayes	Unigram	0.84	0.72	0.75	81.7%
	Bigram	0.85	0.54	0.48	71.6%
	Trigram	0.85	0.51	0.43	69.8%
Gradient Boosting	Unigram	0.35	0.50	0.41	69.3%
	Bigram	0.35	0.50	0.41	69.3%
	Trigram	0.35	0.50	0.41	69.3%
Decision Tree	Unigram	0.79	0.77	0.78	81.7%
	Bigram	0.66	0.64	0.65	71.6%
	Trigram	0.67	0.61	0.61	72.1%
Random Forest	Unigram	0.90	0.77	0.80	85.5%
	Bigram	0.79	0.64	0.66	76.9%
	Trigram	0.82	0.60	0.59	74.9%
K-Nearest Neighbors	Unigram	0.75	0.72	0.73	78.4%
	Bigram	0.70	0.65	0.66	74.6%
	Trigram	0.62	0.63	0.62	66.2%
XGBoost	Unigram	0.85	0.75	0.77	83.3%
	Bigram	0.80	0.61	0.62	75.6%
	Trigram	0.75	0.58	0.56	73.1%
Multi-Layer Perceptron	Unigram	0.84	0.81	0.82	85.5%
	Bigram	0.78	0.74	0.76	80.7%
	Trigram	0.76	0.62	0.62	75.1%

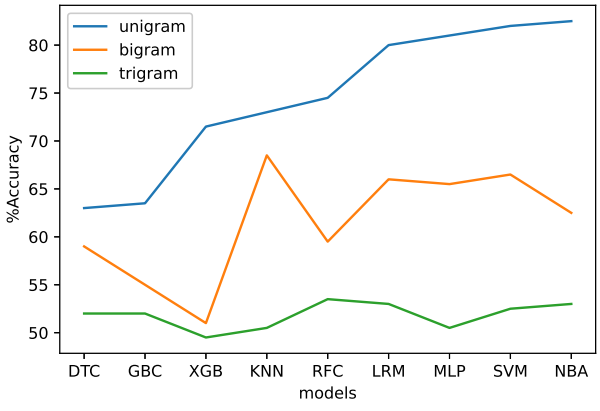

Figure 3. Distribution of n-grams accuracy scores in the Models for the third dataset

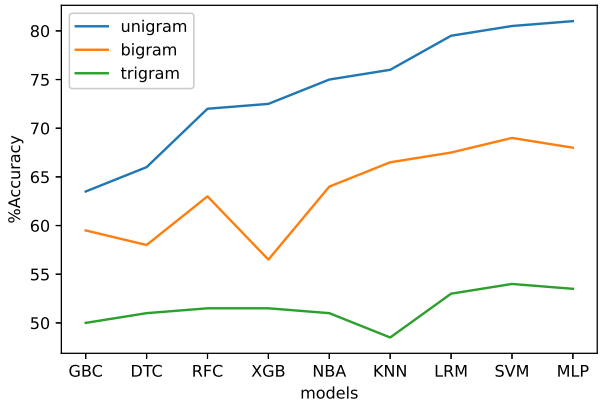

Figure 4. Distribution of n-grams accuracy scores in the Models for the fourth dataset

3. Approach

To recognize patterns and regularities in data, the machine learning and deep learning algorithms use learned patterns to predict new observations. We pre-processed the text data before we applied the different learning algorithms on the text data. We tokenized the data into words and n-grams and generated a vocabulary of all the special n-grams that occurred in the document. Using the term frequency-inverse document frequency (tf-idf) technique, data features were rescaled. We used supervised learning and we compared the results.

For our work, we chose to filter out uncommon, non-informative

Table 2. Precision, Recall, f1 score and accuracy of the classifiers trained on the second dataset.

Model	n-grams	Precision	Recall	F1 score	Accuracy
Logistic Regression	Unigram	0.86	0.85	0.85	85.5%
	Bigram	0.76	0.75	0.75	75.5%
	Trigram	0.63	0.57	0.51	57.0%
Support Vector Machine	Unigram	0.87	0.86	0.86	86.5%
	Bigram	0.75	0.75	0.75	75.0%
	Trigram	0.63	0.57	0.51	57.0%
Naive Bayes	Unigram	0.86	0.84	0.84	84.5%
	Bigram	0.73	0.71	0.70	70.5%
	Trigram	0.65	0.57	0.51	57.0%
Gradient Boosting	Unigram	0.79	0.79	0.78	78.5%
	Bigram	0.73	0.71	0.70	71.0%
	Trigram	0.72	0.58	0.50	58.0%
Decision Tree	Unigram	0.78	0.78	0.78	78.0%
	Bigram	0.68	0.67	0.66	66.5%
	Trigram	0.73	0.56	0.46	56.0%
Random Forest	Unigram	0.82	0.82	0.82	82.0%
	Bigram	0.75	0.73	0.72	73.0%
	Trigram	0.69	0.55	0.45	55.0%
K-Nearest Neighbors	Unigram	0.80	0.79	0.78	78.5%
	Bigram	0.76	0.55	0.43	54.5%
	Trigram	0.53	0.51	0.41	51.0%
XGBoost	Unigram	0.82	0.82	0.82	82.0%
	Bigram	0.71	0.64	0.61	64.0%
	Trigram	0.75	0.51	0.36	51.0%
Multi-Layer Perceptron	Unigram	0.85	0.84	0.84	84.5%
	Bigram	0.75	0.74	0.74	74.5%
	Trigram	0.64	0.57	0.52	57.5%

content by extracting n-grams to make the algorithms more intelligent. We identified some common techniques used in recent studies [6, 2, 4, 16], namely Decision Tree Classifier (DTC), Gradient Boosting Classifier (GBC), Naive Bayes Algorithm (NBA) and Random Forest Classifier (RFC). Others are K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGB), Support Vector Machines (SVM), Logistic Regression Model (LRM) and the Multi-Layer Perceptron (MLP) Classifier. For sentiment analysis, these models have been extensively tested, and provided accurate results when working with various dataset types. The words were patterned for parsing, such that every n-gram consists of n terms and are tagged accordingly. The accuracy of these methods often differ widely in validation, ranging from using small samples to a wide array of tagged data.

4. Experiments

The data used for this analysis consist of a collection of four related economic and financial market reviews selected from multiple texts that have been tagged with positive, negative and neutral classes. These four datasets, extracted from different digital media platforms, have been selected because they contain explicit economic sentiments from which the machine and deep learning algorithms can learn. We have used the Reuters dataset in Pekka et al.[11], containing subjective sentences from economic review, and the IMDb, Amazon, and Yelp datasets in Kotzias et al.[15], which contains text sentences from reviews of products, movies, and restaurants. The first dataset contains reviews and tags for products sold on amazon.com while the second dataset contains the sentiment dataset for IMDb movie reviews. The third and fourth datasets have a collection of texts about economic and restaurant reviews respectively. The text were splitted into training and testing data. Using the training set, the machine and deep learning algorithms were trained to understand, extract and evaluate subjective information from the data with n-grams as features.

Basically, after fitting the training data to the models, we used the various models to predict the tags of the test data. Using the training set to train the algorithm, we translated the data into numeric form, while the test set was used to evaluate the performance of the machine and deep learning models. The machine and deep learning algorithms learnt from the training data, passing the features and tags as parameters. The models predicted the outcomes, while the precision, accuracy and f1 score were obtained using the n-gram features within the model. To keep a list of the word vectors, we transformed the text array into a TF-IDF function matrix and a vocabulary was created. A

Table 3. Precision, Recall, f1 score and accuracy of the classifiers trained on the third dataset.

Model	n-grams	Precision	Recall	F1 score	Accuracy
Logistic Regression	Unigram	0.80	0.80	0.80	80.0%
	Bigram	0.66	0.66	0.66	66.0%
	Trigram	0.55	0.53	0.48	53.0%
Support Vector Machine	Unigram	0.82	0.82	0.82	82.0%
	Bigram	0.67	0.67	0.66	66.5%
	Trigram	0.54	0.53	0.48	52.5%
Naive Bayes	Unigram	0.84	0.82	0.82	82.5%
	Bigram	0.70	0.62	0.59	62.5%
	Trigram	0.60	0.53	0.43	53.0%
Gradient Boosting	Unigram	0.64	0.64	0.63	63.5%
	Bigram	0.56	0.55	0.54	55.0%
	Trigram	0.54	0.52	0.44	52.0%
Decision Tree	Unigram	0.63	0.63	0.63	63.0%
	Bigram	0.60	0.59	0.58	59.0%
	Trigram	0.57	0.52	0.42	52.0%
Random Forest	Unigram	0.75	0.74	0.74	74.5%
	Bigram	0.61	0.59	0.59	59.5%
	Trigram	0.56	0.54	0.48	53.5%
K-Nearest Neighbors	Unigram	0.73	0.73	0.73	73.0%
	Bigram	0.69	0.69	0.68	68.5%
	Trigram	0.51	0.51	0.46	50.5%
XGBoost	Unigram	0.72	0.71	0.71	71.5%
	Bigram	0.51	0.51	0.47	51.0%
	Trigram	0.48	0.49	0.38	49.5%
Multi-Layer Perceptron	Unigram	0.81	0.81	0.81	81.0%
	Bigram	0.66	0.66	0.65	65.5%
	Trigram	0.52	0.51	0.39	50.5%

machine and/or deep learning algorithm can then directly be used on the encoded vectors. The classification and evaluation of the different meanings of the text was carried out and we compared them to each other.

The n-grams offered an indication of the words that could affect the tags of the text. We extracted the n-gram distribution such as unigrams, bigrams, and trigrams for use in the different models, thereby making the learning algorithms more intelligent for proper prediction. We applied the machine and deep learning algorithm on the text for classification and the accuracy score for all models used in the experiment were calculated.

5. Results and Discussion

In this study, we present a performance review of special n-gram based evaluation of a sequence labeling task using different learning algorithm. We introduced machine learning and deep learning techniques to analyze the sentiments in the data for better and faster decision making, and we were able to compare the output of the techniques implemented, thus adding to the state-of-the-art literature on tasks of sentiment analysis. These algorithms were applied on the datasets to predict the tags and to classify it accordingly using the n-grams features.

The performance of the n-grams in the different machine and deep learning approach was calculated using the overall accuracy measurement. For a comparative performance evaluation of each system in terms of predicting the tags correctly, we present the results for the nine methods used for precision, recall, accuracy and F1-score calculation. Tables 1-4 shows the model classification values of the models and Figures 1–4 depicts the distribution of the values on a line graph.

The macro-averaged f1, recall, precision and accuracy scores of the various models used are shown in Tables 1-4. The findings indicate that the SVM and the MLP models generally improved the effectiveness of the classification. The results also reveals that the DTC, GBC, KNN and the XGB failed to perform well in the classification task. In the comparative analysis, using the different methods of machine learning and n-gram approaches on the datasets, results were better compared and the effectiveness of the n-gram features were recorded.

The n-gram features gave a very good performance for all learning algorithms with the unigrams performing better than the bigrams and trigrams in the classification task. The SVM, LRM, RFC, NBA and the MLP models are the most reliable for all of the n-gram features. In the first dataset (see Figure 1), RFC, MLP, SVM had maximum scores among the models. The SVM, LRM, and MLP models gave the highest output for all n-gram functions on the second dataset (see Figure 2). On the third dataset, NBA, SVM and MLP are with the highest results

Table 4. Precision, Recall, f1 score and accuracy of the classifiers trained on the fourth dataset.

Model	n-grams	Precision	Recall	F1 score	Accuracy
Logistic Regression	Unigram	0.80	0.80	0.79	79.5%
	Bigram	0.68	0.68	0.67	67.5%
	Trigram	0.57	0.53	0.45	53.0%
Support Vector Machine	Unigram	0.81	0.81	0.80	80.5%
	Bigram	0.69	0.69	0.69	69.0%
	Trigram	0.59	0.54	0.46	54.0%
Naive Bayes	Unigram	0.75	0.75	0.75	75.0%
	Bigram	0.70	0.64	0.61	64.0%
	Trigram	0.63	0.51	0.36	51.0%
Gradient Boosting	Unigram	0.64	0.64	0.63	63.5%
	Bigram	0.60	0.59	0.59	59.5%
	Trigram	0.50	0.50	0.37	50.0%
Decision Tree	Unigram	0.66	0.66	0.66	66.0%
	Bigram	0.59	0.58	0.57	58.0%
	Trigram	0.57	0.51	0.38	51.0%
Random Forest	Unigram	0.72	0.72	0.72	72.0%
	Bigram	0.64	0.63	0.63	63.0%
	Trigram	0.55	0.52	0.41	51.5%
K-Nearest Neighbors	Unigram	0.77	0.76	0.76	76.0%
	Bigram	0.69	0.67	0.66	66.5%
	Trigram	0.45	0.48	0.38	48.5%
XGBoost	Unigram	0.73	0.72	0.72	72.5%
	Bigram	0.59	0.56	0.53	56.5%
	Trigram	0.55	0.52	0.41	51.5%
Multi-Layer Perceptron	Unigram	0.81	0.81	0.81	81.0%
	Bigram	0.68	0.68	0.68	68.0%
	Trigram	0.59	0.54	0.45	53.5%

(see Figure 4) while on the fourth dataset, SVM, LRM and MLP performed best on the test dataset (see Figure 4). The models used with the feature classification techniques shows the effectiveness of n-grams for sentiments tagging and the most reliable methods of classification.

6. Conclusion

An important problem in the analysis of sentiments is being able to determine the contextual labels or tags of words and phrases. We addressed this problem in this study by successfully introducing various machine learning and deep learning approaches to produce the labels or tags of economics and financial reviews text using n-grams as features. Modeling was performed using different pre-processing techniques in texts, converting the text into vectors, and applying various machine learning and deep learning techniques on the different datasets. The use of multiple classifiers in this analysis led to a better evaluation efficiency than any individual classifier. The findings recorded in this study suggests that the support vector machine and multi-layer perceptron neural networks were the best options for achieving successful results, because they efficiently and effectively classify the sentiment tags behind the sentence in the text. The unigram model, which is an n-gram analysis representation at low level, has a greater predictive potential compared to the bigram and trigram models. While high-level n-gram representations account for the complexities of the human language, their use in predicting consumers' choices is less efficient than low-level n-gram representations in these economic reviews.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico and grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico. The authors also wish to thank Malo Pekka and his colleagues for sharing the Financial Phrasebank Dataset [11] and the University of California, Irvine machine learning repository of databases for the datasets.

References

[1] H. Gómez-Adorno, I. Markov, G. Sidorov, J. Posadas-Durán, M. A. Sanchez-Perez, & L. Chanona-Hernandez, "Improving feature representation based on a neural network for author profiling in social media texts", Computational Intelligence and Neuroscience **2016** (2016) 1638936.

[2] O. Kolesnikova & A. Gelbukh, "A study of lexical function detection with word2vec and supervised machine learning", Journal of Intelligent & Fuzzy Systems **39** (2020) 1993.

[3] S. Poria, E. Cambria, & A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network", Knowledge-Based System **108** (2016) 42.

[4] O. E. Ojo, A. Gelbukh, H. Calvo, O. O. Adebanji, & G. Sidorov, "Sentiment detection in economics texts", Advances in Computational Intelligence @ MICAI 2020 **12469** (2020) 271.

[5] T. Lugo-Garcia,A. Gelbukh, & G. Sidorov, "Unsupervised learning of word combinations for syntactic disambiguation", Avances en la Ciencia de la Computación. Proceedings of the Workshop on Human Language Technologies at the 5th Mexican International Conference on Computer Science, ENC-2004 (2004) 311.

[6] A. Gelbukh & O. Kolesnikova, "Supervised machine learning for predicting the meaning of verb-noun combinations in Spanish" MICAI 2010. Lecture Notes in Artificial Intelligence **6438** (2010) 196

[7] O. Juárez Gambino & H. Calvo, "Predicting emotional reactions to news articles in social networks", Computer Speech & Language **58** (2019) 280.

[8] H. Gómez-Adorno, R. Fuentes-Alba, I. Markov, G. Sidorov & A. Gelbukh, "A convolutional neural network approach for gender and language variety identification", Journal of Intelligent & Fuzzy Systems **36** (2019) 4845.

[9] P. Pakray, A. Pal, G. Majumder, & A. Gelbukh, "Resource building and parts-of-speech (pos) tagging for the mizo language", 14th Mexican International Conference on Artificial Intelligence, **MICAI 2015** (2015) 3.

[10] S. N. Galicia-Haro, A. Gelbukh, & I. A. Bolshakov, "Identification of composite named entities in a spanish textual database", 9th International Conference on Applications of Natural Languages to Information Systems, Salford, UK **3136** (2004) 395.

[11] M. Pekka, A. Sinha, P. Korhonen, J. Wallenius, & P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts", Journal of the Association for Information Science and Technology **65** (2014) 782.

[12] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh & L. Chanona-Hernández, "Syntactic n-grams as machine learning features for natural language processing", Expert Systems with Applications **41** (2014) 853.

[13] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh & A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines", IEEE Intelligent Systems **33** (2018) 17.

[14] S. T. Aroyehun & A. Gelbukh, "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling", Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (**TRAC-1**) Santa Fe, USA (2018) 90.

[15] D. Kotzias, M. Denil, N. de Freitas & P. Smyth, "From group to individual labels using deep features", **KDD 2015** (2015) 597.

[16] V. Athanasiou & M. Maragoudakis, "A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: A case study for modern Greek", Algorithms **10** (2017) 34.