Sentiment Analysis with Word Embedding: The Case of Double-track Education System in Ghana*

[1]O. B. Deho and [1]W. A. Agangiba
[1]University of Mines and Technology (UMaT), Tarkwa Ghana

Deho, O. B. and Agangiba, W. A. (2019), "Sentiment Analysis with Word Embedding: The Case of Double-track Education System in Ghana", *Ghana Journal of Technology,* Vol. 3, No. 2, pp. 51 - 57.

Abstract

This paper applies the concept of sentiment analysis for the determination of polarities (positivity, neutrality or negativity) of sentiments borne in the views expressed by Ghanaians regarding the newly introduced double track system in Second Cycle Schools in Ghana. These views are sourced from tweets (twitter posts). Accurate analysis of sentiments depends largely on the context of word usage. Most sentiment analysis approaches however ignore context when predicting sentiments; thereby leading to loss of context. In this paper, the loss of context is avoided with the use of the concept of Word embedding. Word embedding is a context-preserving technique which embeds the contextual information of data in the form of vectors before analysis of sentiment is done. An overall model accuracy of 76% was achieved using this technique. Our model's accuracy outdoes similar works such as Garg's (2016) work with an accuracy of 72%. The results from this work may help the Ghana government to get well informed on how the citizenry reacted to the reform of the educational system as well as help those at the helm of affairs to know how to roll out policies in the near future.

Keywords: Word2Vec, Word Embedding, Classifier, Sentiment Analysis

1 Introduction

Second cycle education in Ghana has undergone series of reforms and reviews in an attempt to find a model that best fits the needs of the citizenry. Students who complete basic education and meet the basic entry requirements are supposed to be given admission to their respective preferred second cycle institutions. This however is not the case since so many children who qualify for secondary education are denied the chance for various reasons. According to Reports from the Presidential Commission on Education Reforms in Ghana, inability to afford school fees, inadequate infrastructure, inability to meet minimum entry requirements and lack of alternative tracks for students with different interests and abilities are the main reasons why many junior high school graduates were unable to access second cycle education (Annon., 2004). Provisions in the Chapter 5 of Ghana's 1992 Constitution, Clause 1(b) states that "secondary education in its different forms, including technical and vocational education shall be made available and accessible to all by every appropriate means, and in particular, by the progressive introduction of free education".

As a result, Ghana's government in January, 2017 rolled out a free Senior High School (SHS) policy - an initiative which seeks to do away with financial barriers hindering access to second cycle education. The free SHS policy caused enrolment to skyrocket by 33.2% (Partey, 2018) as students who otherwise could have not gone to SHS due to financial constraints could now do so. The surge in enrolment put much pressure on the existing inadequate infrastructure. An attempt to sustain and keep the free SHS policy running then led to the formulation of the double-track education system. The double-track educational system kick started in September 2018. This policy requires that the entire students and staff be divided into two tracks, such that, while one track is in school, the other is on vacation. This decision by the Ghana government has been received by the Ghanaian citizenry with mixed-reactions. Many Ghanaians have expressed their sentiments in relation to the reform of the educational system.

For this research, views expressed on twitter regarding the educational reform are used. This is because the advent of internet and social media has created a platform for people to express their views on various issues: product reviews, governmental policies and so on. Research shows that an average Ghanaian spends 3 hours 30 minutes browsing the internet on his or her phone out of which 3 hours 13 minutes is spent on social media (Annon., 2016). According to Annon. (2018a), a little over 10 million of Ghanaians which constitutes 35% of the Ghanaian population are active internet users. Digital in 2018's survey states that there are 5.6 million Ghanaians who are active social media users. The most used social media platforms in Ghana are WhatsApp, Facebook, YouTube, FB Messenger, Instagram, Google+, Skype, Snapchat, Twitter, LinkedIn, Pinterest and Viber in decreasing order of usage (Annon., 2018b). Contents shared on these social media platforms provide an invaluable source of information which can be leveraged to gain insights and make

**Manuscript received December 05, 2018*
Revised version accepted March 22, 2019

decision on issues and policies (Pang and Lee, 2008).

This paper however, is based on Twitter posts (tweets). There have been various tweets concerning the introduction of the double-track educational system. These tweets have various sentiments borne in them, some being positive, others being neutral or negative. In this paper, we derive insights from the tweets which may be useful to the Ghanaian government in knowing what the majority of Ghanaians are saying by means of sentiment analysis.

Meanings of words or sentences to a large extent depend on the context of usage. This issue being a very dicey one, it is very prudent to use algorithms and techniques that are very accurate and efficient so as to ensure misclassification of tweets are reduced to the barest minimal. In order to improve overall accuracy as well as preserve contextual information, we use a cutting-edge, context preserving technique called word embedding (Word2Vec) (Rong, 2016) in this paper.

1.1 Literature Review

Sentiment analysis is a natural language processing task which involves classification of a piece of text as one that carries a positive, neutral or negative sentiment. The approaches to sentiment analysis are the lexicon-based approach and the machine learning approach. The lexicon-based approach measures the polarity and subjectivity of a textual data against a database (lexicon) of emotional values of words which have been prerecorded by researchers. Different approaches to creating lexicons have been proposed, including manual and automatic approaches (Turney *et al*., 2010). In lexicon-based approach, a piece of text is represented generally as a bag-of-words. A combining function such as average or sum is used to determine the overall sentiment of a text. Lexicon-based approach is easy to implement, but has a downside of disrupting word order and discards semantic information (Turney *et al*., 2010). In machine learning methods, sentiments are classified by applying a machine learning algorithm in the form of a classifier to a piece of text (Pang and Lee, 2002). Sentiment analysis can be done at the document-level (Turney, 2002), sentence level (Hu and Lui, 2004) or aspect level (sentiment about specific aspects of an entity) (Wilson *et al*., 2005).

There have been lots of research done in this area, but one notable thing that strikes through most of these works is that the bag-of-words model (BOW) is mostly used for text representation. According to Shamseera and Sreekanth (2016), the BOW is at best, good for topic-based text classification and not sentiment analysis. The BOW loses contextual information (a key requirement in accurate sentiment classification) by disrupting word order and discards contextual information. Garg (2016), in his work, coupled the BOW model with an ensemble classifier made up of Naïve Bayes classifier, MultinomialNB classifier, Bernoulli classifier, Stochastic Gradient Descent Classifier (SGDC) and Support Vector Classifier (SVC) which resulted in an average accuracy of 72%. The result in Garg's work stems from the deficiencies of BOW model. To improve accuracy in our work, we used word embedding (Word2Vec) coupled with random forest classifier for sentiment analysis.

Word Embedding emerged from the Natural Language Processing (NLP) research field which is an intersection of Computer Science, Artificial Intelligence, Machine Learning and computational linguistics with a long history (Chopra *et al*., 2013). Word embedding is a text mining technique of establishing relationship between words in textual data (Corpus). The syntactic and semantic meanings of words are realized from the context in which they are used. The concept of distributional hypothesis suggests that words occurring in similar context are semantically similar (Sahlgren, 2008). Count based embeddings and prediction-based embeddings are the two broad approaches to word embedding. The count-based embeddings however, just like the traditional bag-of-words model, does poorly at preserving contextual information in textual data (Sunil, 2017). The prediction-based embeddings try to predict a target word given a context word. Word2Vec and Global Vectors for Words Representation (GloVe) are the most commonly used prediction-based algorithms. GloVe, an unsupervised learning algorithm for obtaining vector representation of words developed by researchers at Stanford (Pennington *et al*., 2013) does very well at context preservation.

The Word2Vec model was developed by Tomas Mikolov and his colleagues at Google in 2013. The Word2Vec model uses a shallow neural network with a single hidden layer to embed high quality word vectors (Mikolov *et al.*, 2013). It is an algorithm that is trained to predict target word from the context of its neighboring words. It uses one-hot-encodings of the corpus as the input for the input layer of the neural network. The Word2Vec neural network has a weight matrix between the input layer and the hidden layer and another one between the hidden layer and the output layer. The weight matrix of dimension m*n where m is the size of the dictionary and n that of the hidden layer is in essence the word vector for the predicted target word. Word2Vec is able to learn analogy and perform tasks like predicting the relationship of: *King-man* + *woman* = *queen*, implying that "*man*" is to "*woman*" as "*king*" is to "*queen*". The

Word2Vec algorithm comes in two “flavors”: the continuous bag-of-words (CBOW) model and skip-gram (SG) model. The continuous bag-of-words model predicts a target word, given a context of words. The skip-gram model flips the CBOW architecture around by predicting the context of a given word. The architecture of the CBOW and SG models are shown in Fig. 1 and Fig. 2 respectively. The skip-gram model coupled with negative sampling outperforms the CBOW making it the preferred choice for this research work.

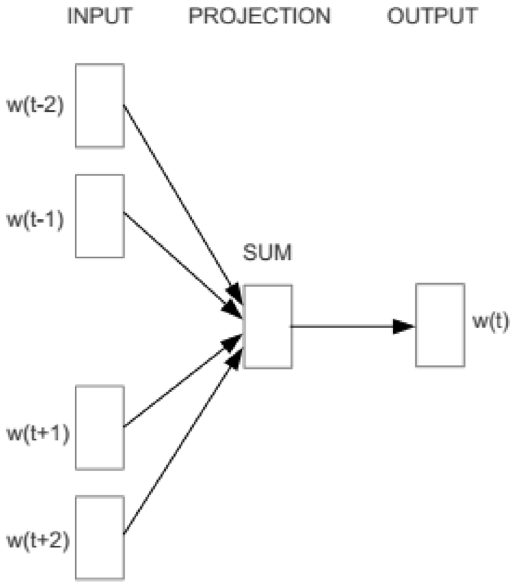


Fig. 1 CBOW Architecture (Mikolov *et al.*, 2013)

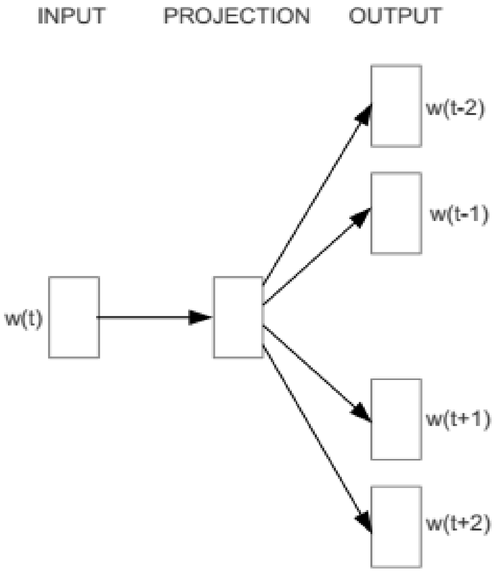


Fig. 2 SG Architecture (Mikolov *et al.*, 2013)

2 Resources and Methods Used

The methods and resources that were used for this work are organized in a step-wise manner as shown in Fig. 3.

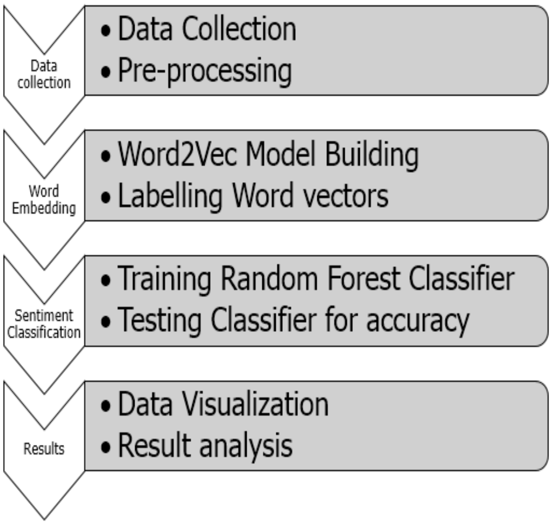


Fig. 3. Steps in Sentiment Analysis

2.1 Data Collection and Preprocessing

2.1.1 Data Collection

The data used for this project was collected from twitter. Twitter is a micro-blogging website with massive patronage. With the passage of every second there are millions of tweets being posted from across the globe. Twitter allows its users a space of 140 characters to write their Twitter posts (tweets). Before one can stream tweets, one must first have a twitter account. In order to communicate with the twitter streaming Application Programming Interface (API) to stream and save tweets, one must login with his twitter account credentials to apps.twitter.com. This allows the creation of a mini twitter application. The created application comes with consumer keys and secret, access tokens and access secrets. These keys are used by the twitter authentication handler to allow access to twitter data. Tweepy, a python library is used in this work to help python to communicate with the Twitter API in order to stream the tweets. Twitter allows access to tweets that posted on the day of access up to tweets from seven days ago. The tweets for this project was collected on four different days when the topic “double-track education system” was trending. The streamed tweets were then saved to a JavaScript Object Notation (JSON) file format.

2.1.2 Preprocessing

The saved tweets come with lots of metadata which do not hold any value as far as this project is concerned. These data are just noise which if not removed would just use up memory space leading to slow computational speed. Symbols like the “#”, “@”, “html tags”, “URLs”, “tweet ids”, “geo-locations” and other metadata had to be cleaned to make the tweets fit for feature extraction. To clean the tweets, BeautifulSoup4, a python library was used to clean the tweets. The final dataset used was derived after the cleaned tweets had been tokenized, stemmed and rid of stop words.

2.2 Word Embedding

2.2.1 Word2Vec Model Building

The skip-gram model of the word2vec algorithm is used for this work. The word2vec implementation of the Gensim python library was trained on the processed data. For better word embedding, certain hyper-parameters were given utmost consideration. These parameters are:

i. Training algorithm;
ii. Dimensionality;
iii. Context window; and
iv. Sub-sampling

The training algorithm used for this project was negative sampling (an optimization algorithm that causes each training sample to update only a small percentage of the model's weight) as it proved to be computationally efficient compared to hierarchical softmax. A dimension of 300 was assigned to the hidden layer of the neural network as it resulted in better word embedding. A context window of 10 was used as it was the prescribed context window for skip-gram models (Mikolov *et al.,* 2013). The sub-sampling rate of 1e-3 was used to counter the imbalance between rare and frequent words in the data set. Due to the size of the data used, the minimum count was set to 1 so that every word in the corpus was considered during training. The word2vec model was trained on the processed data with the aforementioned hyper-parameter settings and saved in a file data format. The word vectors produced by the word embedding model have m*n dimension where m is the size of the dictionary and n is the size of the hidden layer.

2.2.2 Labeling Word Vectors

The word vectors were split into training and testing set. 70% of the word vectors were used as training sample and the rest for testing. VADER (Valence Aware Dictionary for sEntiment Reasoning), a sentiment analysis engine was used to determine the polarities of the various tweets. The VADER sentiment engine uses Vader lexicon which contains lexical features (words) and their sentiment intensities (Hutto *et al.,* 2014). The classification accuracy of VADER engine beats some of the prominent machine learning models as it combines qualitative analysis and empirical validation by using human raters from Amazon Mechanical Turks and the wisdom of the crowd. VADER calculates the overall sentiment polarity by summing up the intensity of each word in the tweet while taking into consideration five heuristics;

Punctuations

The presence of punctuations in tweet affects the intensity of the polarity. The text "I like it !!" sounds more positive than "I like it". VADER captures these subtleties in polarity assignment.

Capitalizations

Capitalizations also somewhat affect sentiment intensity. The text "I LIKE it" sounds more positive than "I like it". VADER takes this into account by incrementing or decrementing the sentiment score of a word depending on whether the word is positive or negative respectively.

Degree modifiers

The presence of modifiers in a text also add some value to the sentiment intensity. Consider "this is a very wise decision" and "this is sort of wise decision". The modifier in the former increase the intensity of wise while that of the latter decreases the intensity. VADER maintains a dictionary of boosters and dampeners to handle modifiers. The effect degree modifiers depend on the proximity of modifier and the word being modified.

Shift in polarity due to "but"

The word "but" is often used to connect two clauses with contrasting sentiments. However, the dominant sentiment is usually the latter. VADER handles polarity shifts by implementing a "but" checker which reduces sentiment valences of words before the "but" by 50% and increase those after by 150% of their values.

Examining the tri-gram before a sentiment laden lexical feature to catch polarity negation

Tri-gram in this context refers to a set of three lexical features. VADER maintains a list of negator words. VADER handles negation by multiplying the sentiment score of sentiment bearing word by empirically determined value of -0.74

In this work, we use VADER to determine the polarities of the raw tweets after which these polarities were assigned to 70% of the respective word vectors as the training data set. The labeled word vectors were used to train a random forest classifier.

2.3 Sentiment Classification

2.3.1 Random Forest Classifier Training

Random forest classifier is an ensemble classifier that is made up of many decision tree classifiers. Random forest classifiers are used for both classification and regression tasks. The forest it builds are usually based on the idea that a combination of learning models increases the overall result. Random forest randomly selects subsets of the training set and creates a set of decision trees from those subsets thereby reducing overfitting. The aggregated score from the different decision trees decides the final class of the test object. It can be represented mathematically by:

$$g(x) = f_o(x) + f_1(x) + f_2(x) + \ldots\ldots. \qquad (1)$$

Where the final model $g(x)$ is the aggregated value of the base models $f_{i.}$ Each base model is a decision tree classifier. This works well as the aggregated value of many decision trees reduces the noise of an individual decision tree thereby giving more accurate results. Some parameters that are tuned in the random forest classifier are the number decision trees generated and decision tree related parameters like minimum split, split criteria etc. The accuracy of the classification depends largely on the number of decision trees used, the larger the number of the trees in the forest, the more accurate the classification. The random forest classifier is a supervised classification algorithm which somewhat works like Naïve Bayes algorithm and Support Vector Machine (SVM). The decision trees form a set of rules for prediction based on the targets and features of the training dataset. There are two stages in the random forest algorithm, the creation of the random forest and making of prediction from the random forest. In this research work, we use a random forest classifier fitted with 100 decision tree classifiers to train the word vectors.

2.3.2 Testing

After training the classifier with 70% of the training data, it was tested with 30% unlabeled word vectors for sentiment prediction. The accuracy of the system was determined to be 76% and validated with some standard validation metrics like the F1 Measure, Precision and Recall. The results of the accuracy and validation tests are given in Table 1.

Table 1 Accuracy and validation metric scores of the Random Forest Classifier

Results			
Metrics	**Negative**	**Neutral**	**Positive**
F1	0.7843	0.8059	0.40000
Precision	0.9090	0.6750	1.0000
Recall	0.6896	1.0000	0.2500
Accuracy	0.7656		

3 Results and Discussion

3.1 Results

The Table 1 shows the classification report of the random forest classifier for each sentiment class. The precision rate, F1-score, recall and overall accuracy of the classifier for the various sentiment classes is shown in the Table 1. An overall accuracy of 76% shows that the classifier did well in predicting the sentiment polarities which is due to the quality of word vectors that were produced by the skip-gram model. The confusion matrix shown in Fig. 4 gives a detailed view of how the classifier did the classification with respect to the true positives, false positives, true negatives and false negatives. The Fig. 5 is a pie chart that was used to substantiate the various percentages of tweets that bore positive, neutral or negative sentiments. From the Fig. 5, it could be inferred that more than half of the tweets were negative which could be thought of as citizens who had opposing views as against those of the government. The neutral sector denotes those tweets which either were not so clear with their stance on the issue or probably were simply indifferent about the implementation of the double-track educational system. The positive sector denotes those tweets that could be thought of as those that are in alignment with the government’s stance. Juxtaposing the results from this research and those reported by the media, Anim-Appau, (2018) and Partey (2018), we could say that the result from this work are not far from the truth.

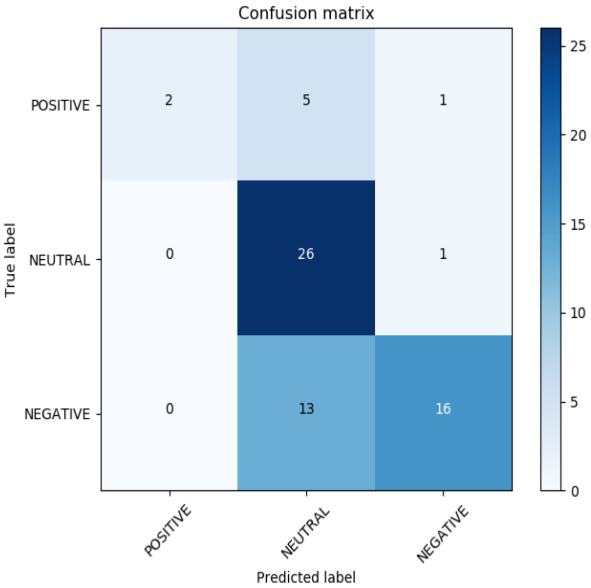


Fig. 4 Confusion Matrix of Random Forest Classifier

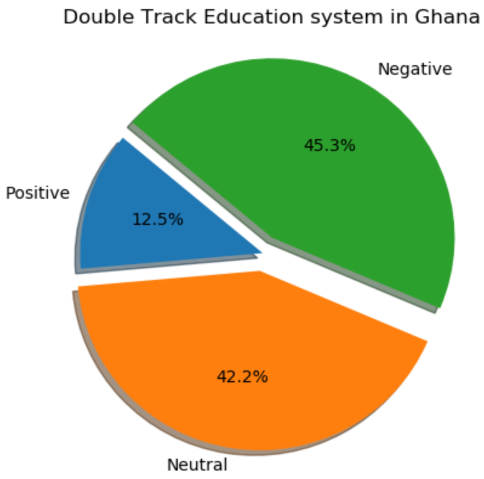


Fig. 5 Pie Chart Showing Sentiment Polarity Distribution

3.2 Limitations

Sentiment analysis of tweets faces some limitations due to reasons such as 140-character limit per post. The limit on number of characters forces users to resort to abbreviations, slangs which somewhat affects context learning. Sarcasm detection has also been another limitation of this work.

4 Conclusions and Recommendations

4.1 Conclusion

This work was undertaken with the primary goal of accurately predicting the sentiment polarity of tweets concerning the implementation of the double-track education system in Ghana using a technique called word embedding to preserve the contextual information of word usage in the tweets which in the end was achieved. Most existing systems ignored context which resulted in less accurate results. An accuracy of 76% and scores from other validation metric scores shown in Table 1 goes to reason that, the technique employed in this work has helped in accurately predicting the sentiment polarities which otherwise would have been much less accurate should an approach such as the bag-of-words model be used for the text representation. The information gathered from this project is a great asset that could help the Ghana government realize how the double-track education system is being responded to by the citizenry as well as give the government idea on policy making and subsequent roll out of policies.

References

Anim-Appau, F. (2018), “Double-track system: Disadvantages outweigh advantages”, *https://www.myjoyonline.com/news/2018/July-26th/double-track-system-disadvantages-outweigh-advantages-educationist.php*. Accessed: October 1, 2018.

Annon. (2004), “White Paper on the Report of the Education Reform Review Committee*”*, *Ministry of Education Youth and Sports*, Accra, Ghana, pp. 1-2.

Annon. (2016), “Digital in 2016”, *https://wearesocial.com/special-reports/digital-in-2016*. Accessed: October 16, 2018.

Annon. (2018a), “Internet Users Statistics for Africa- Africa Internet Usage, 2018 Population Stats and Facebook Subscribers”, *https://www.i-nternetworldstats.com/stats1*.htm. Accessed: October 16, 2018.

Annon. (2018b), “Digital in 2018”, *https://digitalreport.wearesocial.com/*. Accessed: October 16, 2018.

Chopra, A., Prasha, A. and Sain, C. (2013), “Natural Language Processing”, *International Journal of Technological Enhancement and Emerging Engineering & Research*, Vol. 1, No. 4, pp. 131-134.

Garg, P. (2016), “Sentiment Analysis of Twitter Data using NLTK in Python”, *Thapar University*, Patiala, India, 50 pp.

Hu, M., Lui, B. (2004), “Mining and summarizing customer reviews”, *University of Illinois at Chicago, Illinois*, USA, pp. 1-5.

Hutto, C. and Gilbert, E. E. (2014), “VADER: A parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”, ICWSM-14, *Eight International AAAI Conference on Weblogs and Social Media*, Michigan, USA, pp. 1-10.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), “Efficient Estimation of Word Representation in Vector Space”, *Google Inc, Mountain View*, USA, 12 pp.

Partey, A. P. (2018), “Implications of a double-track school calendar on SHS”, *https://www.myjoyonline.com/opinion/2018/July-23rd/implications-of-a-double-track-school-calendar-on-shs.php/*. Accessed: October 1, 2018.

Pang, B. and Lee, L. (2002), “Thumbs up? Sentiment Classification using Machine Learning Techniques”, EMNLP 2002, *Proceedings of the ACL-02 Conference on Empirical methods in natural language processing*, Stroudsburg, USA, Vol. 10, pp. 79-86.

Pang, B. and Lee, L. (2008), “Opinion mining and sentiment analysis”, *Foundations and Trends in Information Retrieval*, Hanover, USA, 135 pp.

Pennington, R., Socher, R. and Manning, D. C. (2013), “GloVe: Global Vectors for Word Representation”, *Stanford University, Stanfor*d, USA, 12 pp.

Shamseera, S. P. and Sreekanth, E. S. (2016), “Word Vectors in Sentiment Analysis”, *International Journal of Current Trends in*

Engineering & Research (IJCTER), Vol.2, No. 5, pp. 594-598

Sunil, R. (2017), "An Intuitive Understanding of Word Embeddings: From Count Vectors to Word2Vec", *https://www.analyticsvidhya.com-/blog/2017/06/word-embeddings-count-word2veec/*. Accessed: January 10, 2018.

Turney, P. (2002), "Thumbs up or thumbs down? Semantics orientation applied to unsupervised classification of reviews", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistic*, Stroudsburg, USA, pp. 417-424.

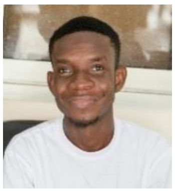
Turney, P. and Pantel, P. (2010), "From Frequency to Meaning: Vector Space Models of Semantics", *International Journal of Artificial Intelligence Research*, Vol. 37, No. 1, pp. 141-188.

Wilson, T., Wiebe, J. and Hoffman, P. (2005), "Recognizing contextual polarity in phrase level sentiment analysis", *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, pp. 347-354.

Rong, X. (2016), "Word2Vec Parameter Learning Explained", *arXiv:1411.273v4*, 21 pp.

Sahlgren, M. (2008), "The Distributional Hypothesis", *Italian Journal of* Linguistics, Vol. 20, No. 1, pp. 20-21.

Authors



Deho Oscar Blessed is a Teaching Assistant at the Computer Science and Engineering Department of University of Mines and Technology (UMaT), Tarkwa, Ghana. He holds a Bachelor's degree in Computer Science and Engineering from UMaT. His current research interest includes data mining and social media analytics.

W. A. Agangiba is currently a lecturer at the Department of Computer Science and Engineering in University of Mines and Technology (UMaT), Tarkwa. He obtained his MSc and BSc degrees in Information Systems and Technologies from the Tver State Technical University, Tver, Russia. He is an Associate Member of the Institute of Electrical and Electronics Engineers (IEEE) and a Professional Member of Association of Computing Machines (ACM). His research interests include high level programming languages, Data Structures and algorithms, Data Analytics, Web and Mobile Technologies, Expert Systems and Logics of Computer Science.

.