International Journal of Technology and Management



Polarity and Similarity Measures Towards Classifying an Article on Food Insecurity

IJOTM
ISSN 2518-8623
DOI: 10.67610/ijotm.5.2.011
Volume 5. Issue II
pp. 1-10, December 2020
ijotm.utamu.ac.ug
email: ijotm@utamu.ac.ug

Andrew Lukyamuzi
Mbarara University of Science and Technology
Email: andrewlukyamuzi@gmail.com

John Ngubiri
Uganda Technology and Management University (UTAMU)
Email: : jngubiri@utamu.ac.ug

Washington Okori
Uganda Technology and Management University (UTAMU)
Email: :

Abstract

Polarity measure has been applied to text mining tasks such as sentiment analysis and text classification. At the same time similarity measure has also been applied to text mining tasks such as summarization, classification and information retrieval. Limited studies if any have used both measures as predictors in a Machine Learning task to automatically identifying articles with conversations on food insecurity in a news feed. These conversations on food insecurity can be generated into trends. Now these trends can guide stakeholders in taking appropriate action depending on food insecurity situation. The proposed strategy relates to a study that used polarity measures for tweets on food prices to predict actual food prices which could also provide proxy information on food insecurity to similarly guide the stakeholders. However our study is on blending polarity and similarity as handcrafted predictors in Machine Learning to automatically label articles on food insecurity using Machine Learning. To explore the proposed strategy, the study used articles from Monitor site (www.monitor.co.ug), a ugandan news media. Promising findings were obtained with KNN and N-BAYES classifiers which had AUC measures of 0.931 and 0.927 respectively compared to random classifier of 0.91. Future work considers blending these handcrafted features with automatic features from deep learning to explore performance improvement.

Key words: *Text similarity, Polarity measure, Food Insecurity*

Introduction

Baseline methods in Machine Learning rely on manipulation of word vocabularies to extract patterns from text data. Traditional Machine Learning uses word counts on these vocabularies based using schemes such as (Term Frequence) TF and (Term Fequence-Inverse Document Frequence (TFIDF) during pattern extraction. These counts are based on bag of words. TF and TFIDF are however unable to capture word meanings (Kowsari et al., 2019). Bag-of-word model used however ignores word order and this restricts learnability of algorithms as this plays important role in some situations. For example the statement ' Paul love Mary' is different from 'Mary loves Paul' much as bag-of-words treats these sentences having same meaning. Deep learning techniques have introduced advance techniques to curb existing challenges such as this. Deep Learning use wordembeddings as text representation and these simultaneously integrate semantic information. Wordembeddings can also operate on word bag-of-words. To address challenges of bag-of-word model, Deep Learning uses advanced algorithms to capture word order. These algorithms include but not limited to: Long Short Term Memory (LSTM) and Rercurrent Neural Netowrks (RNN). The result has been better performance of with wordembeddings and consequently with Deep Learning. Beyond baseline methods, enhancements techniques have been introduced in both traditional Machine Learning and Deep Learning. These include but not limited to : (1) use WordNet to enhance text classification (Ranwez, Duthil, Montmain, & Augereau, 2013), (2) coding and handling of word negation to enhance performance (Elagamy, Stanier, & Sharp, 2018), and (3) use of ensemble learning (Al-Ash, Putri, Mursanto, & Bustamam, 2019).

Polarity measure and similarity if introduced as external features in baseline Machine Learning can potentially enhance text classification. Polarity aims at measuring negativity, neutrality and positivity of a text. This has had important application of tasks such as opinion analysis, sentiment analysis, and review scoring/rating (Bhardwaj, Narayan, & Dutta, 2015; Brahimi, Touahria, & Tari, 2016; Khedr, Salama, & Yaseen, 2017; Pang, Lee, Rd, & Jose, 2002). In nature most text are opinionated ranging from positivity to negativity. Our argument is that this tendency can be harnessed to study some situation if appropriate home work is done. This where we now introduce ourarguement: food insecurity have tendencies of negativity and we can exploit this fact to model a classifier on this topic or otherwise. On the other hand, similarity measure establishes similarity between texts. Similarity measure have been applied to tasks such plagiarismdetection, information retrieval and summarization (Wh Gomaa & Fahmy, 2012; Jayakodi, Bandara, Perera, & Meedeniya, 2016; Rahutomo, Kitasuka, & Aritsugi, 2012). The arguement raised is that if a suitable corpus on food insecurity is selected, a text on food insecurity is expected to generate high similarity with selected corpus. It is on the basis of this argument that this study explores how reliable is similarity measure in classifying if an article is on food insecurity or not.

The study aims at blending similarity and polarity in a text classification task, an area which has received limited attention. The few studies (such as in Feng et al., 2013; Mao, Xiao, & Mercer, 2015; Tidke, Mehta, Rana, & Jangir, 2020) that attempted this blending, have ignored applying the same strategy to a Machine Learning task of automatically identifying articles on food insecurity from news feed. This study attempts to close this gap. The work closely relates to a study by Surjandari, Naffisah, & Prawiradinata, (2015) which exploited tweet polarities predict food price increase. The study extends this work (in Surjandari et al., 2015) still in the direction of studying food patterns. However this study is on blending polarity and similarity measures in a task of classifying either a text is on food insecurity or not and more so using news articles insteady of tweets.

Food insecurity is an attractive situation to study due unforgettable sufferings it has inflicted on human kind. It will be remembered the between 1343 and 1345, 43 million Europeans perished because of famine (Mellor & Gavian, 1987). In the same report between 16 and 64 million people in China faced similar catastrophe in the period from 1959 to 1961. Even in this current era of technological advancement food insecurity has been not completely conquered. According FAO (2020) about 821 million people is experiencing some form of food insecurity. When it comes to developing countries, unfortunately food

insecurity tends to hit harder as observed by (Clover, 2003; FAO, 2020). Uganda as a developing country is also not immune to the challenge. Between July 2012 to August 2014, approximately 60,000 households (400,000 individuals) were identified as food insecure in Northern Uganda due to climate changes(USAID, 2017). Other incidences of food insecurity have been reported and investigated by researchers such as; (Bahiigwa, 1999) for 1997-198, (FEWSNET-Uganda, 2009) for 2016 and (Umana-Aponte, 2011) for 1980. Recently IPC, (2017) reported that about 10.9 million in Uganda were victims of food insecurity. It is on this basis that Uganda was used as case study to investigate the proposed strategy.

To explore this concept, news articles from Monitor website were used as experimental data sets. These were retrieved based on 'food' as keyword. The study was on developing a reliable classifier to correctly identify/classify news feed on food insecurity from the rest. Article on food insecurity can be generated into trends to reveal the prevailing circumstances of food insecurity. This is a potential source of information to guide stakeholders towards appropriate action.

This study brings forth the following contributions:

(i) Explores blending polarity and text measures as potential handcrafted features in a Machine Learning task of automatically identifying articles on food insecurity from a news feed.

(ii) It builds on an investigation which has explored use of tweet sentiment/polarity to predict food prices (Surjandari et al., 2015), a situation which potentially provide proxy information on food insecurity. This study however is on blending polarity and similarity measures in a Machine Learning task of automatically identifying articles from news feed. Additionally the study uses specifically news articles as opposed to using tweets.

The paper is organised into six sections. Section I is introduction, section II reviews related work, section III is on methodology, section IV presents results , section six discusses the results. The paper ends with section VI which presents conclusion and future work.

Related work

This section reviews related work which has informed the study. The purpose is to highlight which areas have been adopted and those areas where this study has contributed to close gaps identified.

Text polarity measures:

Polarity measure is the negativity or neutrality or positivity of a text. Three options have been proposed in generating polarity measure; lexicon (rule based) approaches, Machine Learning approaches and hybride approach (Heerschop et al., 2011; Krishnamoorthy, 2017). Lexicon based approach lies on established polarity on a dictionary of words. These words are summed up to compute the overal polarity of a text. The challenge is however the overal polarity generated with this method does not always ream with human judgement on the overal polarity of a text. Studies have generated promising results with lexicon based approach in tasks such as stock prediction, sentiment analysis, product review, tweet analysis, movies review scoring. For more details see studies for example Brahimi et al., 2016; Gupta, Singh, & Singla, 2020; Heerschop et al., 2011; Hernández, Lorenzo, Simón-Cuevas, Arco, & Serrano-Guerrero, 2019; Pang, Lee, & Vaithyanathan, 2002; Rıfkı Aydın, Güngör, & Erkan, 2020. Machine Learning approach use models such as K-Nearest Neighbors(KNN), Support Vector Machine (SVM) and Naïve Bayes (N-BAYES) algorithms to model polarity measure. The offer tendencies of more accurate results than lexicon approach. Blended (also called mixed) approaches are capable of generating benefits both in lexicon approach and Machine Learning approach but at the same time supressing the limitations in either Machine Learning approach or lexicon approach. In text mining, black box approaches have proposed libraries to generate text polarity such as Textbob. Our study opted to explore text polarities generated with this library due to its simple of use.

Text Similarity

Text similarity can be categorized into; string based similarity, corpus based similarity, and knowledge based similarity (WH Gomaa & Fahmy, 2013; Hajeer, 2012; Vijaymeena & Kavitha, 2016). String based similarity establish similarity by comparing individual characters in the text compared. This is mainly between words and words. It is applicable in tasks such spell checking and suggestions and gene sequence matching. Several String based similarity measures have been proposed and include but not limited to: Cosine, Euclidean, Jaro and Jaccard similarity measures (Vijaymeena & Kavitha, 2016). Corpus based similarity is on computing similarity between words using relatedness establish using a large corpora (Vijaymeena & Kavitha, 2016). This has been applied to tasks such as question/answering, and analogy analysis. Knowledge based similarity measure similarity between words and concepts. Knowledge based similarities utilize tools such as ontologies capable of providing the required similarity. In WordNet as an ontology was applied to text mining to study medical rcords (WH Gomaa & Fahmy, 2013; Vijaymeena & Kavitha, 2016).

As traditional Machine Learning is evolving in deep Learning, new/improved text similarity techniques have followed suit. In traditional Machine Learning a computer scientist tend to manually engineer features as predictor in a data science task. Algorithm in this category include KNN, SVM, and N-BAYES. Deep Learning is modern approach and is towards dependence on automatic feature extraction. Covolutional Neural Networks (CNN) is an early algorithm in the initial stages of deep Learning. It began mainly as an image processing algorithm though it has been extended to text mining studies such as in (Mandelbaum & Shalev, 2016). Other dominant algorithms in deep learning include Recurrent Neural Network (RNN), Longest Short Term Memory (LSTM), and Gated Relation Network (GRN). Now let us switch back to similarity measure in Deep Learning. Similarity techniques introduced include Word Movers Distance, word vector-based Dynamic Time Warping (wDTW) and word vector-based Tree Edit Distance (wTED) (Zhu, Klabjan, & Bless, 2017). Our study choose to explore using cosine measure as a widely implemented measure as observed in (Rahutomo et al., 2012).

Blending similarity and polarity measures

Current technological development have given rise to massive generation of data sent/uploaded to site such as weblogs, social networks, and news sites. This massive data is unfortunately mostly unstructured and this makes it challenging for utilization.Text mining aim at devicing mechanisms to harness useful patterns from this kind of data sets. Data mining techniques have take route in areas such identification of social networks groups, opinion mining, information retrieval, topical trends, stock trends. A long this direction in (Tidke et al., 2020) researchers have proposed use of similarity and sentiment/polarity analysis to organise tweets in groups/classess. This has vital application in various tasks such as identification of social groups and tracking topical trends.

Advancement in technology have generated opportunities for several users to contribute their view towards a chosen topic of interest. In Wikipedia, this possibility facilitates a decision on article for delection. Before a final decision is taken to delete such an article, several members are given opportunity to provide thier reactions accordingly. Views and their justification are compiled and studied to establish strenghths and weaknesses raised towards either in support or otherwise. If a big number of members are generating reactions, it is cumbersome to evaluate, reach and take a final decision. In (Mao et al., 2015) text-to-text similarity measure and sentence-level sentiment analysis algorithm is applied to facilitate this process on a suggestion concerning an article for Deletion in Wikipedia.

Available corpa such tweets, google, blogs provide opportunities for comparative studies to identify which corpus is more approapriate than the others in tasks such as sentiment analysis, recommender system for specific situation. In (Feng et al., 2013) such an opportunity is exploited. Comparative study was carried to identify which corpus is appropriate for sentiment similary of text from three corpra; google, tweets, and

blogs. In relation to these studies, our research is towards blending similarity and polarity measure in context of classifying if text is on food insecurity or otherwise, are which has been neglicated.

Methodology

The study was on generating a classifier to identify if an article is on food insecurity from a news feed using polarity and similarity as handcrafted features. The study adopted text classification process described by Kadhim(2019). This is a generic strategy and it maps closely with classification processes described by other researchers (for example Korde, 2012; Patra & Singh, 2013; Tilve & Jain, 2017). The strategy adopted is a six step process as describe in most relevant and these include; (1) data collection, (2) text pre-processing, (3) feature extraction, (4) dimenstion reduction, (5) classification techniques, and (6) perfomance evaluation. Below is a description on how these steps were customized or applied to suit the study objective.

Data collection

This stage is concerned with acquisition of data. The study used articles retrieved from Monitor site, a Ugandan news media. The articles wereretrieved based on food as a search keyword. Food was used as a keyword because an article containing this keyword can be on food insecurity or not. Other keywords such as crops, harvest and rains could equally be used as search keyword. Much as these other words can play similar role, the study chose food as a representative. These articles were cleaned to remove formatting programming tags such as html tags and java scripting codes. Each article was annoted as 1 if it was on food insecurity or 0 otherwise.

Pre-processing, feature extaction, dimension reduction

These next three steps (text pre-processing, feature extraction, dimension reduction) are for generation of sieved features ready for classifier manipulation.

Text pre-processing stage deals with preparation of text data ready for data manipulation. Tasks at this stage include tokenization, removal of stop words, stemming and vector space model. Tokenization deals with determining entities to be used as features. Several schemes have proposed such as n-gram approach. With n-gram, n represents number of words to represent as single feature. In n-gram strategy n can vary from 1 to more words though it is not advised n to exceed 5 due to computational overhead. Removal of stop words deals with elimination of words that do not contain essential signal in text computation and includes words such as articles and prepositions. Stemming reduces words to the root form. For example the words stop, stops, stopping and stopped can be reduced to their common root of stop. This reduces word redundancy which in turn reduces computational overhead. Vector space model deals with conversional of words into numerical format ready for data manipulation.

At feature extraction stage, the researcher determines and can sieve the most relevant features using some statistical techniques. Dimensional reduction can be applied in the case of high dimensional data-points and this can require using some special techniques.

These three stages are for conversion of text data into numerical formats for algorithmic manipulation.

These stages however represent a traditional approach for generating standard features. To generate the handcrafted features proposed, the study took the following strategy instead of going through these three stages.

Generating handcrafted features

Polarity measure was proposed as a potential feature in the study. Polarity measure gives negativity or positivity of a statement. This is experimented on classifying if an article is on food insecurity or not. An article on food insecurity has tendencies of negative polarity. We are using this idea of negative tendencies as a clue for article is on food insecurity. The challenge is that some can have negative tendencies yet they are not

on food insecurity. This raises a problem of confusion to the classifier if it is dependent on polarity measures. This bipolar tendencies for articles not on food insecurity introduces some negative effect on classifier learning. The research however remained optimistic amidst this opposition. Several options are available to generate polarity measures and these include; (1) developing a polarity using appropriate corpus, (2) using a polarity measure from previous researchers, and (3) using appropriate libraries to generate text polarities. The study opted to use polarity measures generated using textblob library.

The second feature used was similarity measure. Now this is a description on how this feature was generated. Suppose sample a text T is selected for a particular class of interest. Any other text belonging to the class of interest, is expected to posses high similarity with text T. Conversely the similarity is low if the text does not belong to this class of interest. This idea was extended to this task of identifying if a text is on food insecurity or not. A sample text was created from a group of phrases/sentences on food insecurity randomly selected from news article secured. Similarity measure was computed between the sample text and each article. Text similarity computed was based on cosine measure metric with the text converted into numerical representations based on TF-IDF.

Classification techniques and evaluation

Now the two formed features (polarity and similarity measures) from news articles were subjected to N-BAYES and KNN classifiers for supervised training. AUC_measure was used as a metric to assess/evaluate the usefulness of these features. Scatter plot was also applied to give assessment but on visual aspect.

Results

Performance measures

The AUC_measures computed for the two classifiers are as shown in table 1. The measures were computed against a random classifier performance of 0.91. It can be observed that performance for the classifiers was above the random performance classifier.

Table 1: showing AUC_ measures for classifiers.

Classifier	Classification accuracy
NAÏVE BAYES	0.927
KNN	0.931

Graphical visualization

Figure 1 shows a 2-dimension visualization how polarity and similarity measures can aid in distinguishing if an article is on food insecurity or otherwise. Grey labels is for articles on food insecurity and dark labels are for articles on otherwise. It can observed that to some extent polarity and similarity measure can provide discrimination if an article is one food insecurity or not.

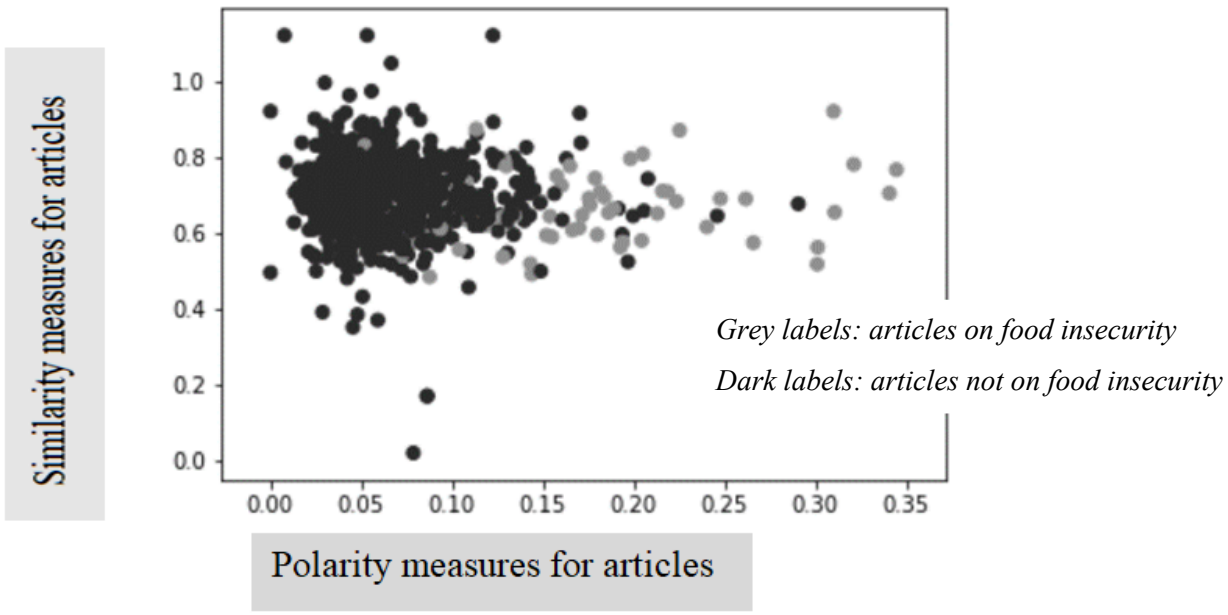


Figure 4.1: scatter diagram showing for articles on food insecurity and otherwise.

Discussion

Findings show that classifier AUC-measures (shown in table1) with proposed feature are better than random classifier performance of 0.91. The visualization graph also shows that the proposed features can provide discrimination between classes. This is an empirical evidence that the traditional engineered features (polarity and similarity measures) are relevant in this context. The study relates to Ranwez et al.,(2013) who showed use of ontology to generate unique features to enhance performance. In similar way Elagamy et al., (2018) used linguistic knowledge of word negation to improve performance. In general sense, the study relates to other investigations which attempted to introduce external knowledge into feature engineering to enhance performance. The difference is on topical area and in our study we have not integrated the engineered features with automatic featuring engineering. To best of our knowledge were have not come across a research which has used our innovation to generate the potential engineered features in this context. A major limitation of this study is the low AUC-measures; the measures are slightly better than random classifier. This effect is vividly observable of the figure where there is some overlap between two regions. To promote better performance future work considers; (1) Integrating these engineered feature with automatic feature extraction, and (2) introducing ensemble learning in classifier training.

Conclusion and Future Work

The study has demonstrated usefulness of polarity and similarity measures in the context of classifying if a news article is on food insecurity or not. Promising findings obtained have provided a proof of concept nevertheless classification performances were not excitingly high. Future work considers blending this strategy with automatic standard features to promote performance in the context of deep learning and ensemble learning.

References

Al-Ash, H. S., Putri, M. F., Mursanto, P., & Bustamam, A. (2019). Ensemble Learning Approach on Indonesian Fake News Classification. *3rd International Conference on Informatics and Computational Sciences (ICICoS)*, 4. https://doi.org/10.1109/ICICoS48119.2019.8982409

Bahiigwa, G. B. a. (1999). Household Food Security in Uganda: an Empirical Analysis. In *Economic Policy Research Center*.

Bhardwaj, A., Narayan, Y., & Dutta, M. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. *4thInternational Conference on Eco-Friendly Computing and Communication Systems*, *70*, 85–91. https://doi.org/10.1016/j.procs.2015.10.043

Brahimi, B., Touahria, M., & Tari, A. (2016). Data and text mining techniques for classifying Arabic tweet polarity. *Journal of Digital Information Management*, *14*(1), 15–25. Retrieved from https://pdfs.semanticscholar.org/524d/40624482a7136a0745d9ad5d3b4d49dbc2ca.pdf?_ga=2.109524613.1250956934.1581511305-1292638669.1565597119 [Accessed on 12 Feb 2020]

Clover, J. (2003). *FEATURE FOOD SECURITY IN SUB-SAHARAN AFRICA*.

Elagamy, M. N., Stanier, C., & Sharp, B. (2018). Stock market random forest-text mining system mining critical indicators of stock market movements. *2nd International Conference on Natural Language and Speech Processing, ICNLSP 2018*, 1–8. https://doi.org/10.1109/ICNLSP.2018.8374370

FAO. (2020). Crop Prospects and Food Situation. In *Crop Prospects and Food Situation #1, March 2020*. https://doi.org/10.4060/ca8032en

Feng, S., Zhang, L., Li, B., Wang, D., Yu, G., & Wong, K. F. (2013). Is twitter a better corpus for measuring sentiment similarity? *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (October), 897–902.

FEWSNET-Uganda. (2009). *Uganda Food Security Outlook*.

Gomaa, Wh, & Fahmy, A. (2012). Short Answer Grading Using String Similarity And Corpus-Based Similarity. *Int. Journal of Advanced Computer Science and Applications*, *3*(11), 115–121. https://doi.org/10.14569/IJACSA.2012.031119

Gomaa, WH, & Fahmy, A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, *68*(13), 13–18. https://doi.org/10.5120/11638-7118

Gupta, S., Singh, R., & Singla, V. (2020). Emoticon and text sarcasm detection in sentiment analysis. *First International Conference on Sustainable Technologies for Computational Intelligence, Advances in Intelligent*, 1–10. https://doi.org/10.1007/978-981-15-0029-9_1

Hajeer, S. I. (2012). Comparison on the Effectiveness of Different Statistical Similarity Measures. *International Journal of Computer Applications*, *53*(8), 14–19. https://doi.org/10.5120/8440-2224

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & De Jong, F. (2011). Polarity analysis of texts using discourse structure. *Proceedings of the 20th ACM International Conference on Information and Knowledge*

Management. https://doi.org/10.1145/2063576.2063730

Hernández, A. R., Lorenzo, M. M. G., Simón-Cuevas, A., Arco, L., & Serrano-Guerrero, J. (2019). A semantic approach for topic-based polarity detection: a case study in the Spanish language. *Procedia Computer Science*, *162*(2019), 849–856. https://doi.org/10.1016/j.procs.2019.12.059

IPC. (2017). *Uganda - Current Acute Food Insecurity Situation*.

Jayakodi, K., Bandara, M., Perera, I., & Meedeniya, D. (2016). WordNet and cosine similarity based classifier of exam questions using bloom's taxonomy. *International Journal of Emerging Technologies in Learning*, *11*(4), 142–149. https://doi.org/10.3991/ijet.v11i04.5654

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273–292. https://doi.org/10.1007/s10462-018-09677-1

Khedr, A. E., Salama, S. E., & Yaseen, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *International Journal of Intelligent Systems and Applications*, *2017*(7), 22–30. https://doi.org/10.5815/ijisa.2017.07.03

Korde, V. (2012). Text Classification and Classifiers: A Survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85–99. https://doi.org/10.5121/ijaia.2012.3208

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms : A Survey. *Information*, *10*(4), 1–68. https://doi.org/10.3390/info10040150

Krishnamoorthy, S. (2017). *Sentiment Analysis of Financial News Articles using Performance Indicators*. https://doi.org/10.1007/s10115-017-1134-1

Mandelbaum, A., & Shalev, A. (2016). *Word Embeddings and Their Use In Sentence Classification Tasks*. Retrieved from https://arxiv.org/pdf/1610.08229.pdf [Accessed on 26th Nov 2018]

Mao, W., Xiao, L., & Mercer, R. (2015). The Use of Text Similarity and Sentiment Analysis to Examine Rationales in the Large-Scale Online Deliberations. *Proceedings Ofthe 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 147–153. https://doi.org/10.3115/v1/w14-2624

Mellor, J. W., & Gavian, S. (1987). Famine : Causes , Prevention , and Relief. *Science*, *235*, 539–545.

Pang, B., Lee, L., Rd, H., & Jose, S. (2002). Thumbs up ? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (July), 79–86. Retrieved from https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf [Accessed on 12 Feb 2020]

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.1109/ICIIECS.2017.8276047

Patra, A., & Singh, D. (2013). A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms. *International Journal of Computer Applications*, *75*(7), 14–18. https://doi.org/10.5120/13122-0472

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic Cosine Similarity. *The 7th International Student*

Conference on Advanced Science and Technology ICAST. Retrieved from https://www.researchgate.net/profile/Faisal_Rahutomo/publication/262525676_Semantic_Cosine_Similarity/links/0a85e537ee3b675c1e000000/Semantic-Cosine-Similarity.pdf [Accessed on 4th Jan 2020]

Ranwez, S., Duthil, B., Montmain, J., & Augereau, P. (2013). How Ontology Based Information Retrieval Systems May Benefit from Lexical How ontology based information retrieval systems may benefit from lexical text analysis. *New Trends of Research in Ontologies and Lexical Resources*, (February), 209–231. https://doi.org/10.1007/978-3-642-31782-8

Rıfkı Aydın, C., Güngör, T., & Erkan, A. (2020). *Generating Word and Document Embeddings for Sentiment Analysis*. Retrieved from https://arxiv.org/pdf/2001.01269.pdf [Accessed on 2nd Feb 2020]

Surjandari, I., Naffisah, M. S., & Prawiradinata, M. I. (2015). Text Mining of Twitter Data for Public Sentiment Analysis of Staple Foods Price Changes. *Journal of Industrial and Intelligent Information*, *3*(3), 253–257. https://doi.org/10.12720/jiii.3.3.253-257

Tidke, B., Mehta, R., Rana, D., & Jangir, H. (2020). Topic sensitive user clustering using sentiment score and similarity measures: Big data and social network. *International Journal of Web-Based Learning and Teaching Technologies*, *15*(2), 34–45. https://doi.org/10.4018/IJWLTT.2020040103

Tilve, A. K. S., & Jain, S. N. (2017). A Survey on Machine Learning Techniques for Text Classification. *International Journal of Engineering Sciences & Research Technology*, *6*(2), 513–520. Retrieved from http://www.ijesrt.com/issues pdf file/Archive-2017/February-217/72.pdf [Accessed on 14th Oct. 2020]

Umana-Aponte, M. (2011). *Long-term effects of a nutritional shock: the 1980 famine of Karamoja, Uganda*. Retrieved from http://www.bristol.ac.uk/cmpo/publications/papers/2011/wp258.pdf%5Cnhttp://search.ebscohost.com/login.aspx?direct=true&db=ecn&AN=1233542&site=ehost-live&scope=site

USAID. (2017). *Climate Risk Screening for Food Security, Karamoja Region, Uganda*. Retrieved from https://www.usaid.gov/sites/default/files/documents/1866/170130_Karamoja_Food_Security_Climate_Screening.pdf [Accessed on 30th May 2019]

Vijaymeena, M. K., & Kavitha, K. (2016). A SURVEY ON SIMILARITY MEASURES IN TEXT MINING. *Machine Learning and Applications: An International Journal(MLAIJ)*, *3*(1), 19–28. Retrieved from http://aircconline.com/mlaij/V3N1/3116mlaij03.pdf[Accessed on 29th March 2017]

Zhu, X., Klabjan, D., & Bless, P. (2017). Semantic Document Distance Measures and Unsupervised Document Revision Detection. *Proceedings Ofthe The 8th International Joint Conference on Natural Language Processing*, 947–956. Retrieved from http://arxiv.org/abs/1709.01256 [Accessed on 30th Nov 2018]