Agricultural Domain-Specific Jargon Words Identification in Amharic Text

Melaku Lake(✉) and Tesfa Tegegne

ICT4D Research Center, Bahir Dar Institute of Technology,
Bahir Dar University, Bahir Dar, Ethiopia

Abstract. Domain-specific jargon words are lists of words used in formal communication of a particular domain with domain experts and non-domain experts; however, it is difficult to understand by non-experts and society. Experts of an organization use jargon words in scientific and science communication to keep the protocol of the communication within a domain. The domain-specific Amharic jargon words negatively impact people out of the domain experts to understand the main theme of the disseminated content in science communication. We followed a design science research approach to conduct our study. We prepared a knowledge base with a list of domain-specific Amharic Jargon Words and the meaning of the word. Machine learning classifier algorithms are employed for model development with Support Vector Machine, Artificial Neural Network, and Naïve Bayes with TFIDF feature selection that returns a classification accuracy of 96.2%, 95.2%, and 94.7% respectively. The knowledge-based system best performs when a smaller number of test sentences are entered into the system. For the input of 20, 40, 60, and 80 test sentences, an accuracy of 88.2%, 86.7%, 85.4%, and 83.1% is observed. So that with the hybrid of machine learning and knowledge-based, identification of domain-specific Amharic jargon words is performed. Therefore, we observed promised result with the hybrid of machine learning and knowledge base for the identification of jargon words in jargony text.

Keywords: Natural language processing · Domain-specific jargon words · Science communication · Knowledge base · Machine learning

1 Introduction

The research in the area of NLP addresses problems in line with language modeling, morphological processing, syntactic processing, and semantic processing [1]. These days, NLP strives to obtain effective communication and accurate knowledge like human beings with increased use of human language in computational language processing to obtain human-like language processing [2]. We use scientific and science communication for communicating science [3]. Scientific communication is written by experts for experts and simple to understand for receivers. In science communication, experts of a domain in an organization communicate to the people outside the domain, and non-experts using domain-specific words to keep the protocol of communication. We attempted science communication to provide prominent information to non-domain

M. L. Berihun (Ed.): ICAST 2021, LNICST 411, pp. 412–423, 2022.
https://doi.org/10.1007/978-3-030-93709-6_27

experts. In science communication, jargon words are used intentionally or unintentionally. Usage of jargon words in formal communication makes the communication cumbersome as the meaning of the jargon words are unknown for the communicants and jargon words hamper the interaction [4].

Jargon words are defined by the Oxford English dictionary as "words or expressions that are used by a particular profession or group of people, and are difficult for others to understand". Jargon words are defined as a list of domain terminologies used by experts of an organization that is necessary for the communication of a particular field; however, it needs meaning for users of text out of the field [5]. Experts use the words to explore their ideas besides organizational related tasks and also the readers of the text have a common understanding of the words used in a text. Scientists that use domain terminologies in workshops, meetings are stressed to reach a non-expert reader with the target theme [4].

1.1 Motivation

Communication with a particular language requires the combination of words for communicants. The selection of words is the responsibility of writers for prominent communication between communicants. We attempted Amharic domain-specific jargon words to alienate the communication barrier between agricultural domain experts and non-expert readers. We considered agricultural domains that are highly vulnerable to the occurrence of Amharic jargon words. The agricultural domain in Ethiopia has huge customers because the domain is the dominant source of income for an estimated 85% of the people. Clear and precise communication between agricultural experts and the people in Ethiopia is fundamental to maximize yield and achieve food security [7].

1.2 Amharic Agricultural Jargon Words Justification

We surveyed non-experts such as farmers, non-domain experts, and agricultural domain experts to know the level of knowing and using agricultural jargon words in communication. We use judgmental sampling and random sampling techniques for selecting samples from a large size population to fill the prepared questionnaire. We prepared a questionnaire for selected respondents and we collected and analyzed the responses.

We randomly selected 40 Amharic agricultural domain-specific jargon words around 9% of the total from the knowledge source. We prepared different close-ended questionnaire formats for the agrarian society (farmers), non-experts, and domain experts. We selected 3 samples from each group of respondents. The following figure depicts the analysis of collected data from the respondents. The rate depicted in the following figure is based on the agreement of all respondents. For example, 92.5% of the randomly selected jargon words are known by all of the regional bureau expert respondents; all regional bureau expert respondents use 70% of the randomly surveyed words for communication. So that we observed equal treatment for usual and rare usage of words on the respondents (Fig. 1).

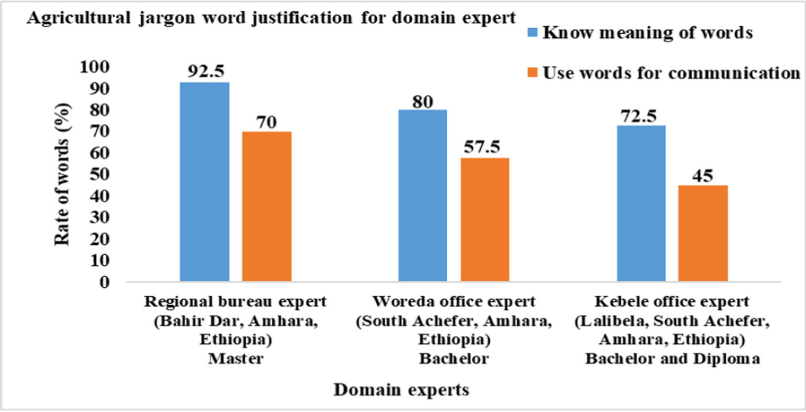

Fig. 1. Agricultural jargon word justification for domain-experts

Though knowing the meaning of jargon words is a challenge for agrarian society, the words are available in a text that confuses readers to understand the target theme. Based on the survey, we found the use of DSAJW in agricultural discourse is a weighty problem for domain experts, non-domain experts, and society.

Experts that use domain-specific jargon words on the letter, advertisement, reports, and social media are stressed to reach a non-expert reader [4]. The existence of jargon words in a text return wastage of time and money for the reader, increase the traffic of searching words. The proposed system considered literate farmers in which the ability of reading and writing is possible. The running industry-leading business of the government attempt to literate Ethiopian farmers such as the launching adult's education curriculum in Ethiopian general education.

Experts of the agricultural domain handle simple communication with non-experts and society. Readers of news, advertisement, business reports, working manuals, periodicals, magazines, social media are full of information besides the theme of the content. The developed DSAJWI system motivates experts to use domain terminology in organizational discourse to handle simple communication with customers. For the Amharic language, the usage of domain terminology used for the development and usage of Amharic language in a domain.

The following are the main contributions of our study.

- We assured the problem is weighty for experts and non-experts with the survey.
- We developed a system for the agricultural society for both experts and non-experts.
- We integrated the machine learning and knowledge-based component.

The rest of this paper is organized as follows. Section 2, presents reviews of related work made on domain-specific jargon words, Sect. 3, presents the proposed system, and a description of the function of phases in the proposed system. Section 4, discusses the experimentation and dataset preparation. Section 5 discusses on evaluation. Section 6 discusses on discussion of performance results. Finally, Sect. 7 presents the conclusion and recommendation of future works.

2 Related Work

The use of domain-specific jargon words in a text increases the frustration of the people during reading documents, emails; It returns a wastage of time and money to understand the accurate meaning of words. Removing jargon words is impossible because the words are formal languages of organizations; jargon with new concepts is invented in various domains at different times. Besides, experts are expected to minimize the use of more jargon in organizational discourse to increase the content to be understandable by the targeted user [8].

Physicians and patients require effective communication to come up with the best outcomes of the treatment and consultancy process. Translation of clinical jargon-to-layperson understandable language is essential to improve the communication between physician and patient in the process of treatment, and consultation. This clinical jargon translation is also used for physicians with the active involvement of patients to increase their decision-making ability concerning the patient's health conditions. The authors use unsupervised learning for unseen datasets using representation learning, bilingual dictionary induction, and statistical machine translation. The authors use unsupervised bilingual dictionary induction (BDI) to learn a mapping dictionary for the alignment of embedding spaces and return a precision of 82.7% at the subword level [9].

Web-based treatment and patient consultation today have increased [10]. In a web-based application, physicians use many medical jargon words for treatment and consultation; this results in the patient's frustration and confusion. The use of medical words in the digital world using different platforms on the internet is increasing. Because of the confusion and frustration of patients, the authors generate new Consumer Healthcare Vocabulary (CHV) using predefined lexical source or ontology for the medical jargon in the online consultation process to increase the understanding of patients. The authors use word embedding with GloVe Iterative Feedback (GloVeIF) and basic GloVe. The GloVeIF outperforms by 8.7% F-measure from the basic GloVE [6].

Dark Jargon words are benign-looking words that have hidden meaning to the user and require clean words. The authors use the word distribution model with Kullback-Leibler Divergence (KL), and cross-context lexical analysis (CCLA) methodology to detect the presence of jargon words in a text and mapped to the word meanings. Binary mapping of dark words to clean words is investigated using dark corpus and clean corpus. The word distribution of KL methodology outperforms around 90% of MRR from CCLA for all words and simulated dark words, however, the CCLA performs better for all words of 97.4% and performs worse for simulated dark words. So that KL outperforms the CCLA for the target dark jargon words detection and identification to provide meaning [11].

Medical words are challenging to understand by ordinary people (by non-medical people). Biological concepts require induction of meaning to be understandable to non-experts using predefined ontologies by domain expert annotators. The authors use dictionary-based Variable-step Window Identification Algorithm (VWIA) for biomedical concepts. Datasets are collected by crawling the URL of the necessary

website. After the necessary preprocessing techniques are performed the developed system returns biomedical concepts based on the constructed dictionary with an F-measure of 95%. However, this work is intended for biomedical concept classification for further analysis, there is no meaning of concepts [12].

To the best of our knowledge, there are no prior works in domain-specific Amharic jargon word identification (DSAJWI) using texts in a particular domain. So that we are motivated to do our work on domain-specific Amharic jargon word identification.

3 Proposed System

The three main components in the proposed domain-specific Amharic jargon words identification (DSAJWI) system are preprocessing component, model development component, and knowledge base component. The following Fig. 2 describes the main phases and necessary steps in the DSAJWI system.

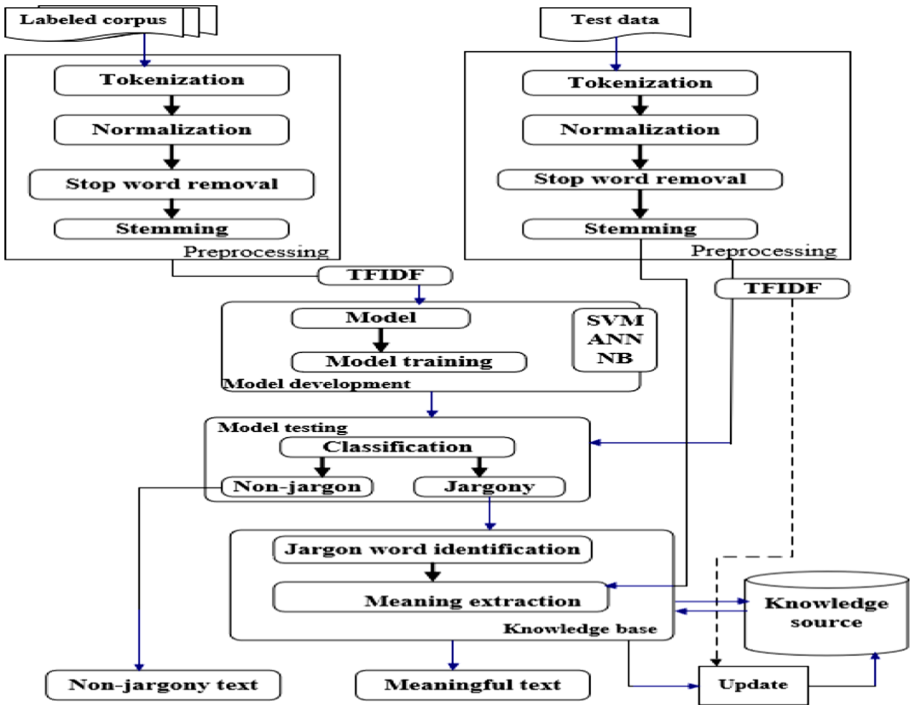

Fig. 2. Proposed system architecture for domain-specific Amharic jargon words identification system

3.1 Preprocessing

Preprocessing components make the input sentence suitable for further analysis with different preprocessing techniques that include tokenization, normalization, stop word removal, and stemming [13]. Tokenization is the first step in the preprocessing technique that can be performed right next to the input sentence to segment an input sentence into a list of tokens [13]. The target DSAJW can be generated from the list of tokens. Normalization is the process of making words having similar pronunciations to have similar representation [13]. Variant forms of Amharic characters have different meanings in a language. So that further work is required to represent variant forms of Amharic characters as per their meaning. The Amharic language use various forms of morphologically generated stop words and stop words are removed to work with content-bearing words that include DSAJW [14].

Amharic is a morphologically complex language and removing affixes of words and generating morphemes from inflectional Amharic words inline to bound morphemes is a challenging task in Amharic morphological analysis. We collected a list of prefixes and suffixes from Amharic language experts, and we removed these affixes to get the stem of the Amharic word.

TFIDF: We used a powerful feature engineering technique Term Frequency Inverse Document Frequency (TFIDF) to identify the important and precisely rare words in the text data [15]. For our work, we used the techniques to convert the strings of a text into numbers so that the developed SVM, ANN, and NB machine learning models consume the input data in numerical formats. The TFIDF feature selection technique is used for scoring words in machine learning models for Amharic language processing.

3.2 Knowledge Base

Amharic Jargon Machine Readable Dictionary (AJMRD) is a knowledge-based lexical resource used to store Amharic agricultural jargon words and the words meaning collected from various agricultural sources to employ for computational linguistics. The meaning of Amharic agricultural jargon words is sourced from agricultural domain experts. Agricultural erudite reviewed the constructed knowledge source on the behalf of the meaning of words obtained from domain experts. The AJMRD helps users to extract the meaning of an exact jargon word with binary lexical mapping, and an over-stemmed jargon word with the help of a close match to the stored words. Because there is no prior AJMRD developed for any of the reasons, we developed interactive AJMRD.

The meaning of collected jargon words in the text are stored in the knowledge source. However, jargon word is invented for different reasons besides the organization's business. The newly invented jargon words by agricultural domain experts and the jargon words that are not included in our knowledge source require meaning for users of text. So that the knowledge source becomes updated as new words occur in the input text and domain experts are required to provide meaning.

The Amharic jargon identification phase is the first phase in the knowledge base component of DSAJWI. The input of the jargon identification phase is a list of tokens passed from the classification phase of the machine learning component. So that the

existence of each list of tokens in the input sentence is checked from the knowledge source to extract the meaning of words. Amharic jargon identification phase is used to identify a particular jargon word from the input text hence, AJMRD is the main lexical knowledge source for our identification.

The meaning extraction phase of DSAJWI extracts the meaning of the identified jargon word from the knowledge source. Meaning extraction is performed from the knowledge source when a word is identified as a jargon word. So that for the occurrence of DSAJW in a text of domain, the meaning of the word is extracted. Therefore, Amharic text containing DSAJW with prominent meaningful text is returned to the user.

Over-stemming: though stemming is a challenge for the meaning extraction from the knowledge source, we handle the problem with entering the over-stemmed word.

4 Experiment

4.1 Experiment Setup

We use python version 3.82 programming language for our implementation because Python is the former programming language in the current computing environment and it supports many open-source libraries. We use anaconda distribution and Jupyter notebook editor to work with our experiment. We imported various python libraries that are compatible with our experiment. Domain-specific Amharic jargon word identification has been done with a hybrid approach using the labeled trained corpus and the knowledge source. So that we used machine learning techniques to develop a model with labeled trained corpus and also, we used a knowledge base for meaning extraction.

4.2 Dataset Preparation

Table 1. Dataset prepared for machine learning and knowledge base

Dataset	Machine learning			Knowledge base	
	Training	Testing	Total	Testing	AJMRD
Sentences	832	208	1040	80	–
Jargon word	–	–	–	59	358

The experiments are done with SVM, ANN, and NB machine learning classifiers with the TFIDF feature selection technique. The performance of the ML classifiers is compared with precision, recall, f1-score, and accuracy for the two-way classification (Table 1).

We collected sentences manually from agricultural reports, training manuals, working guidelines, advertisements of product and service delivery processes that contain Amharic agricultural jargon words. Labeled trained corpus was prepared from

sentences with and without Amharic agricultural jargon words. We collected 1.04k dataset that comprises jargon and non-jargon words. In this study, the 80/20 split ratio is used for training and testing sets/phases.

We collected 358 domain-specific Amharic agricultural jargon words from different agricultural sources with the help of agricultural domain expert curators to prepare the AJMRD. We randomly prepared a total of 80 test sentences of different lengths for different experiments to test the knowledge base performance.

5 Evaluation

Performance evaluation is required for our developed system to know the effectiveness and efficiency of the system. The performance of the proposed system is evaluated with machine learning to classify the input text as jargony or non-jargony and also, evaluated with the knowledge-based system. The evaluation of the knowledge-based component is based on the capability of the system to extract the meaning of identified jargon words from the predefined explanatory lexical knowledge source. We used precision, recall, f1- score, and accuracy to evaluate the machine learning and knowledge base component of our proposed system [16]. We used precision, recall, f1-score, and accuracy that calculated the correctness and completeness of the test set to evaluate the performance.

5.1 Machine Learning Evaluation

The evaluation of our proposed system on the machine learning component is committed with the comparison of models developed from the machine learning algorithms. We developed machine learning models to select the most likely model for the classification of the input text. So that we selected SVM, ANN, and NB machine learning algorithms to compare the classification result and select the outperformed model. The performance result of three supervised ML models is compared for the same labeled input corpus for the agricultural domain. The same algorithm for feature vector representation with TFIDF vectorizer was employed. The algorithms are compared with precision, recall, f1-score, and accuracy (Table 2).

Table 2. Precision, recall, f1- score result of SVM, ANN, and NB

Performance metrics	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Naïve Bayes (NB)
Precision	96	95	94
Recall	96	95	94
F1-score	96	95	95
Accuracy	96.2	95.2	94.7

We observed that SVM outperforms the other model. Because of the performance result of the models, SVM is selected to predict the input test data for the knowledge base.

5.2 Knowledge Base Evaluation

The performance evaluation on the knowledge-based component measures the capability of the knowledge-based system to extract the meaning of jargon words for the input text with the AJMRD.

We randomly used a total of 80 test sentences of different lengths for different experiments to test the performance of the knowledge base system with 20, 40, 60, and 80 test sentences. The different lengths of sentences are used to evaluate the performance on the behalf of the number of input test sentences. We performed different experiments to measure the knowledge base performance. For the occurrence of 17 Amharic agricultural jargon words in 20 test sentences, a knowledge base extracts the meaning of jargon words with a performance of 88.2% accuracy. We prepared 40 sentences containing 30 agricultural jargon words, and the meaning of words are extracted with the performance 86.7%. The following figure depicts the number of input test sentences and the performance (Fig. 3).

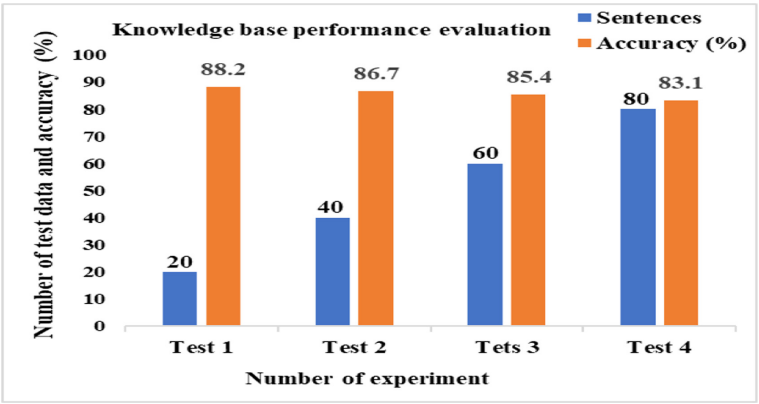

Fig. 3. Performance result of the knowledge base for test 1, test 2, test 3, and test 4

The result of the knowledge base shows as the number of domain-specific Amharic jargon words in the input test sentences and the number of test sentences increases, we observed a decrease rate of performance. Over-stemming on some of the jargon words is committed because of the nature of the word and binary lexical mapping between the over-stemmed jargon word in the input text and the jargon word in the knowledge source is impossible.

6 Discussion

Experimental results described that the hybrid system for DSAJWI is affected by the labeled trained corpus to classify a text as jargony or non-jargony text. The machine learning component of our proposed system minimizes the workload of the knowledge-based system by discarding non-jargon text without entering the knowledge-based system. The developed machine learning model identified the input text as jargony solely entered to the knowledge-based system for further analysis and extraction of meaning. We observed machine learning models predict the input text with 96.2%, 95.2%, and 94.7% accuracy using SVM, ANN, and NB respectively. For the prepared labeled trained data, SVM model outperformed the other developed models because SVM works well for binary classification [17]. We selected SVM for the model testing phase of our proposed system to classify unseen test data as jargony and non-jargony. The knowledge-based component of our proposed system best performs when fewer input sentences are entered into the system. We observed accuracy of 88.2%, 86.7%, 85.4%, and 83.1% for the input of 20, 40, 60, and 80 test sentences respectively.

Therefore, the proposed hybrid system works well for the identification of jargon and non-jargon text. The machine learning model decreases the workload of the knowledge base by discarding non-jargon text from entering the knowledge base system. Texts classified as jargony text are entered into the knowledge-based system. So that for every occurrence of a jargon word in a jargony text, the meaning of the word is extracted from the knowledge source.

7 Conclusion and Future Work

The study focused on the identification of jargon words in a text and provide meaning of words for agriculture domains. We performed operations using both the machine learning and knowledge-based approach. Evaluation of the developed machine learning models are performed and we selected the outperformed model. We have developed models with SVM, ANN, and NB. First, we evaluated the machine learning model and we achieved an accuracy of 96.2%, 95.2%, and 94.7% using SVM, ANN, and NB respectively. We selected the outperformed SVM model.

We observed the best performance of the knowledge-based system for the input of the small number of test sentences. For the input of 20, 40, 60, and 80 test sentences, an accuracy of 88.2%, 86.7%, 85.4%, and 83.1% is observed. So that for a few sentences entered into the knowledge-based system, the best performance of the system is observed. Therefore, we observed the best performance of our proposed system with the knowledge base to extract the meaning of jargon words from the predefined explanatory lexical knowledge source for jargon text with less amount input test data. Therefore, we have achieved a promised result for domain-specific Amharic jargon word identification. In our current study, we only consider Amharic jargon word identification in the agricultural domain with a hybrid of machine learning and knowledge-based. For future work, we are intended to consider other domains in which Amharic is the working language of a domain by increasing features for our proposed system. Additionally, we will compare our result with other techniques inline to

provide the meaning of jargon words in a text of domain. The limitation of the proposed system is that the knowledge base is created manually to retrieve meaning. This problem can be alleviated by using other techniques. We recommend for the future, automatic generation of the meaning of domain-specific jargon words in various domains and select best hyperparameter value combination to benefit customers.

Acknowledgment. The routine tasks of this paper are surely granted by the great contribution of agricultural domain experts, erudite, and agrarian society in Ethiopia.

References

1. Sparck Jones, K.: Natural language processing: a historical review. In: Zampolli, A., Calzolari, N., Palmer, M. (eds.) Current Issues in Computational Linguistics: In Honour of Don Walker, pp. 3–16. Springer Netherlands, Dordrecht (1994). https://doi.org/10.1007/978-0-585-35958-8_1
2. Kevitt, P.M., Partridge, D., Wilks, Y.: Approaches to natural language discourse processing. Artif. Intell. Rev. **6**(4), 333–364 (1992). https://doi.org/10.1007/BF00123689
3. Burns, T.W., O'Connor, D.J., Stocklmayer, S.M.: Science communication: a contemporary definition. Public Underst. Sci. **12**(2), 183–202 (2003). https://doi.org/10.1177/09636625030122004
4. Rakedzon, T., Segev, E., Chapnik, N., Yosef, R., Baram-Tsabari, A.: Automatic jargon identifier for scientists engaging with the public and science communication educators. PLoS One **12**(8), 1–13 (2017). https://doi.org/10.1371/journal.pone.0181742
5. Helmreich, S., Llevadias Jané, J., Farwell, D.: Identifying jargon in texts. Identif. Jarg. Texts **35**(35), 425–432 (2005)
6. Ibrahim, M., Gauch, S., Salman, O., Alqahatani, M.: Enriching consumer health vocabulary using enhanced glove word embedding. In: CEUR Workshop Proc., vol. 2619 (2020)
7. Demeke, M., Ferede, T.: Agricultural Development in Ethiopia : Are There Alternatives to Food Aid? (2014)
8. Willoughby, S.D., Johnson, K., Sterman, L.: Quantifying scientific jargon. Public Understand. Sci. **29**(6), 634–643 (2020). https://doi.org/10.1177/0963662520937436
9. Weng, W.H., Chung, Y.A., Szolovits, P.: Unsupervised clinical language translation. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 3121–3131 (2019). https://doi.org/10.1145/3292500.3330710
10. Cyr, A.: Social media: don't discount the benefits! Oncol. Times **34**(8), 1–3 (2012). https://doi.org/10.1097/01.COT.0000414683.49317.3b
11. Seyler, D., Liu, W., Wang, X., Zhai, C.: Towards Dark Jargon Interpretation in Underground Forums, pp. 1–8 (2020). Available at: http://arxiv.org/abs/2011.03011
12. Gong, L., Yang, R., Liu, Q., Dong, Z., Chen, H., Yang, G.: A dictionary-based approach for identifying biomedical concepts. Int. J. Pattern Recognit. Artif. Intell. **31**(9), 1–12 (2017). https://doi.org/10.1142/S021800141757004X
13. Hermawan, R.: Natural language processing with python, vol. 1, no. 1 (2011)
14. El-Khair, I.A.: Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study (2006, 2017). Available at: http://arxiv.org/abs/1702.01925
15. Jing, L.P., Huang, H.K., Shi, H.B.: Improved feature selection approach TFIDF in text mining. In: Proc. 2002 Int. Conf. Mach. Learn. Cybern., vol. 2, pp. 944–946 (2002). https://doi.org/10.1109/icmlc.2002.1174522

16. Dalianis, H.: Evaluation metrics and evaluation. In: Dalianis, H. (ed.) Clinical Text Mining, pp. 45–53. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-78503-5_6
17. Holts, A., Riquelme, C., Alfaro, R.: Automated text binary classification using machine learning approach. In: Proc. Int. Conf. Chil. Comput. Sci. Soc. SCCC, pp. 212–217 (2010). https://doi.org/10.1109/SCCC.2010.30