SOCIAL NETWORK HATE SPEECH DETECTION FOR AMHARIC LANGUAGE

Zewdie Mossie[1] and Jenq-Haur Wang[2]

[1]Department of International Graduate Program in Electrical Engineering and Computer Science,
National Taipei University of Technology, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering,
National Taipei University of Technology, Taipei, Taiwan

ABSTRACT

The anonymity of social networks makes it attractive for hate speech to mask their criminal activities online posing a challenge to the world and in particular Ethiopia. With this ever-increasing volume of social media data, hate speech identification becomes a challenge in aggravating conflict between citizens of nations. The high rate of production, has become difficult to collect, store and analyze such big data using traditional detection methods. This paper proposed the application of apache spark in hate speech detection to reduce the challenges. Authors developed an apache spark based model to classify Amharic Facebook posts and comments into hate and not hate. Authors employed Random forest and Naïve Bayes for learning and Word2Vec and TF-IDF for feature selection. Tested by 10-fold cross-validation, the model based on word2vec embedding performed best with 79.83%accuracy. The proposed method achieve a promising result with unique feature of spark for big data.

KEYWORDS

Amharic Hate speech detection, Social networks and spark, Amharic posts and comments

1. INTRODUCTION

A major bottleneck for promoting use of computers and the Internet is that many languages lack the basic tools that would make it possible for people to access ICT in their own language. The status of language processing tools for European languages[2] states that only English, French and Spanish have sufficient basic tools. Thus the vast majority of the World's languages are still under-resourced in that they have few or no language processing tools and resources which particularly true for sub Saharan African languages. However, the evolution of the Internet and of social media texts, such as Twitter, YouTube and Facebook messages, has created many new opportunities for creating such tools, but also many new challenges [1].Amharic is one of the sub-Saharan countries Ethiopian's working language which is written left-to-right in its own unique script which lacks capitalization and in total has 275 characters mainly consonant-vowel pairs. It is the second largest Semitic language in the world after Arabic and spoken by about 40% of the population as a first or second language [3] but current population estimated to 102 million. In

Dhinaharan Nagamalai et al. (Eds) : NATL, CSEA, DMDBS, Fuzzy, ITCON, NSEC, COMIT - 2018
pp. 41–55, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80604

spite of its relatively large number of speakers, Amharic is still a language for which very few computational linguistic resources have been developed, and very little has been done in terms of making useful higher level Internet or computer based applications.

This paper focus only on hate speech detection from social media posts and comments. Recent advances in mobile computing and the Internet have resulted in an increase in use of social media to communicate, express opinions, interact with other, and to find and share information [4]. While social media provides an important avenue for communication to take place easily and efficiently, it also acts as a means of spreading hate speech online. Inherent characteristics of the Internet largely contribute to the misuse of social network to transmit and propagate hate speech.

Hate messages are prevalent and challenging in the Ethiopian online community as individuals spread hate messages hiding behind their screens. The government of Ethiopia oversee and monitor content in social network in a bid to govern hate speech through one time interruption of the internet service. Research conducted by Amnesty International and the Open Observatory of Network Interference (OONI) between June and October 2016 shows that access to WhatsApp and others was blocked, as well as at least 16 news outlets [6]. It is an open secret that the recent widespread hate speech and call for violence particularly targets persons of a particular group [5].In this regard no work is done before and the first for Amharic language even though the work of [36]done from the social science perspective. It is therefore, of critical importance to monitor and identify instances of hate speech, as soon as possible to prevent their spread and possible unfolding into acts of violence or hate crimes and destroys the lives of individuals, families, communities and the country.

The proposed method used Word2Vec and TF-IDF for feature selection and Naïve Bayes and Random forest machine learning algorithms known for hate speech detection performance. The rest of this paper is organized as follows. Section 2 reviews related work on hate speech detection. The method and data preprocessing steps are described in detail in Section 3. Architectural design and experimentations are illustrated and discussed in Section 4. Finally, conclusion and future work in Section 5.

2. RELATED WORK

2.1 Hate Speech on Social Media

Online spaces are often exploited and misused to spread content that can be degrading, abusive, or otherwise harmful to people. Hateful speech has become a major problem for every kind of online platform where user-generated content appears from the comment sections of news websites to real-time chat sessions. Legal and academic literature generally defines hate speech as speech or any form of expression that expresses hatred against a person or group of people because of a characteristic they share, or a group to which they belong [7]. But, there is no consensus definition because of prevailing social norms, context, and individual and collective interpretation. A recent study define hate speech as speech which either promotes acts of violence or creates an environment of prejudice that may eventually result in actual violent acts against a group of people[8]. In the case of Ethiopia the use of hateful words with an intention to bring about hatred against a group of people based on their ethnicity, political attitude, religion and socio -economic are prevailing [36].

2.2 Social Media Definition of Hate Speech

- Hate speech is to incite violence or hate: The several definitions use slightly different terms to describe when hate speech occurs. The majority of the definitions point out that hate speech is to incite violence or hate towards a minority (Code of conduct, ILGA, YouTube and Twitter)

- Hate speech is to attack or diminish: Additionally, some other definitions state that hate speech is to use language that attacks or diminishes these groups (Facebook,YouTube, and Twitter).

After consulting those papers, authors use these dimensions of analysis to define what is hate speech in the scope of this paper.

2.3 Existing Techniques Used in Hate Speech Detection in Social Media

The study of hate speech detection has been growing only in the few last years. However, some studies have already been conducted in few languages. Papers focusing algorithms for hate speech detection, and also other studies focusing on related concepts, can give us insight about which features to use in this classification task. Therefore, authors allocate this specific section to describe the features already employed in previous works dividing into two categories: general features used in text mining and specific hate speech detection features.

Dictionaries and lexicons: The majority of the papers authors found try to adapt strategies already known in text mining to the specific problem of hate speech detection. The work categorize the features as the features commonly used in text mining which is dictionaries and lexicons. This approach consists in making a list of words that are searched and counted in the text. In the case of hate speech detection this has been conducted using content words such as insult and swear words, reaction words, and personal pronouns [24], number of disrespectful words in the text, with a dictionary that consists of words for English language including acronyms and abbreviations [26], label specific features which consisted in using frequently used forms of verbal abuse as well as widely used stereotypical words[27], Ortony lexicon was also used for negative affect detection (list of words denoting a negative connotation and can be useful because not every rude comment necessarily contains bad language and can be equally harmful) [11].

Bag-of-words(BOW): Another model similar to dictionaries is the use of bag-of-words [9,10, 11]. In this case, a corpus is created based on the words that are in the training data, instead of a pre-defined set of words, as in the dictionaries. The disadvantages of this kind of approaches is that the word sequence is ignored, and also it's syntactic and semantic content. Therefore,it can lead to misclassification if the words are used in different contexts. To overcome this limitation n-grams were implemented. N-grams are one of the most used techniques in hate speech automatic detection and related tasks [11, 12, 13, 14, 15]. In a study character ngram features proved to be more predictive than to kenn-gram features, for the specific problem of abusive language detection [16].

TF-IDF was also used in this kind of classification problems. It is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times that a

word appears in the document. However, it is distinct from a bag of words, or n-gram, because the frequency of the term is off-setted by the frequency of the word in the corpus, which compensates the fact that some words appear more frequently in general [17].

Part-of-speech (POS) approaches also make it possible to improve the importance of the context and detect the role of the word in the context of a sentence. These approaches consist in detecting the category of the word, for instance, personal pronoun (PRP), Verb non-third person singular present form (VBP), Adjectives (JJ), Determiners (DT), Verb base forms (VB). Part of speech has also been used in hate speech detection problem even though proved to cause confusion in the class's identification [14]. It was also used to detect sentences such as "send them home", "get them out" or "should be hung" [18].

Word Embedding: Deep learning techniques are recently being used in text classification and sentiment analysis with high accuracy [28]. One of the approaches of this is word embedding which allows finding both semantic and syntactic relation of words, which permits the capturing of more refined attributes and contextual cues that are inherent in human language. Therefore, Word2Vec [19], an unsupervised word embedding-based approach to detect semantic and syntactic word relations was used. Word2Vec is a two-layer neural network that operates on a set of texts to initially establish a vocabulary based on the words included in such set more times than a user-defined threshold to eliminate noise. According to [19] 50-300 dimensions can model hundreds of millions of words with high accuracy. Possible methods to build the actual model are CBOW (i.e., Continuous bag of words), which uses context to predict a target word, and Skip-gram, which uses a word to predict a target context. Skip-gram works well with small amounts of training data and handles rare words or phrases well, while CBOW shows better accuracy for frequent words and is faster to train. Word embedding combined with Convolutional Neural Networks (CNN) show better performance [20, 28]. Authors [26] use a paragraph2vec approach to classify language on user comments as abusive or clean and also to predict the central word in the message. Alternatively, other authors propose comment embedding to solve this problem [27]. FastText is also being used [28] in a problem that sentences must be classified and not words. **Sentiment Analysis** bearing in mind that hate speech has a negative polarity, authors have been putting the sentiment as a feature for hate speech detection [15, 23, 24, 25, 31,].

2.4 Algorithms Commonly Used For Hate Speech Detection

Consulting different sources on algorithms of hate speech detection are the focus of this section, because authors aim to work in this specific topic. In the majority of the works the used language is English. However, there were some researched works done for languages Dutch [21]and Italian [22]to author's knowledge. The most common approach found in the work of [15] as a literature review consists in building a machine learning model for hate speech classification. It is found that the most common algorithms used are SVM, Random Forests, Decision Trees, logistic regression, Naïve Bayes and Deep learning respectively on the use of frequency by authors. The data classification is based on general hate speech, racism, sexism, religion, anti-Semitism, nationality, politics and socio-economics status respectively on the categorization use of frequency. Authors propose Random Forest and Naïve Bayes for their good performance.

3. PROPOSED METHODOLOGY AND DATA COLLECTION

Aiming at classifying the hate level across Facebook for Amharic language users, authors have built a corpus of comments retrieved from Facebook public pages of Ethiopian newspapers, individual politicians, activist, TV and radio broadcast and groups. These pages typically posts discussions spanning across a variety of political and religious topics. By doing so, authors could capture both casual conversations and politically hated posts and comments. Authors have employed a versatile Facebook crawler, which exploits the Graph API to retrieve the content of the comments from Facebook posts using Facepager. Facebook is selected to collect data from social media for the following reasons. Facebook is the most important platform for reaching out to online audiences, and especially the youth. Comparative studies have shown how in countries with limited Internet penetration, like Ethiopia, Facebook has become almost a synonym for the Internet, a platform through which users access information, services, and participate in online communications.

3.1 Data Preparation and Annotation

Authors then preprocessed the posts and comments according to the following rules:

- Only kept comments that were in Amharic and all punctuations were removed by passing to the apache spark map function
- All null values are also removed with isNull attribute of apache spark DataFrame
- Checked to assure that no repetitions with the same text by passing to the map function using distinct attribute available on apache RDD and DataFrame
- Removed the HTML and different symbols in the same way using apache spark since authors focus only on texts
- All elongations were removed to the same fixed size character based on the nature of Amharic language and finally Trim text as final step

After all the above preprocessing authors consider the following three bases for future annotation:

(1) **Discourse analysis:-**places the text in its wider political, ethnicity, socio-economic and religious context in order to understand the currents of thought which illustrate and rationalize why it is to be considered hateful or not.

(2) **Content analysis:-**analyses the text deemed not hate and hate in order to pick out the key semantic components and targets of the speech. This can then be coded and quantitative techniques applied to draw wider patterns and trends.

(3) **Automated techniques:** - a relatively novel method of tracking hate speech that can be usefully employed to mine high volumes of text from different sources to search for keywords which are highly indicators of hate speech in an efficient manner which authors followed in labeling the collected data. After the initial cleaning authors got 25,890 posts and comments

available, however authors sampled to be 10, 000 due to the limitation in resources for the annotation task.

3.2 Annotation Instructions

Despite the differences between the previous studies that analyzed in the related work, the majority of the described works present instructions for the annotation task. Some authors point out that having vague annotation guidelines [8, 30]is a problem for hate speech detection due to the complexity of the task. In this work, authors prepared a complete set of annotation instructions in chart in order to better standardize the annotation procedure and to make clear all hate and not hate speech related category concepts. A set of instructions and examples that contain the indicators of the category was defined in figure 1. These are based on the definitions, rules and examples that presented already in the related work. The annotators were given the instructions as guidelines in the classification of the messages

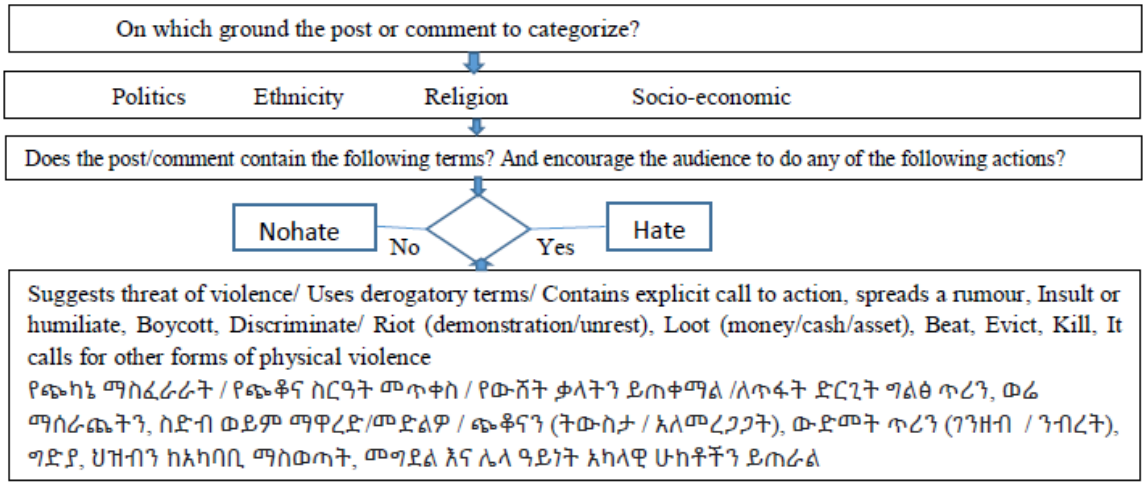

Figure 1: Hierarchical Structure of Dataset Annotation

In the work of [32] the training data was hand-coded and manually annotated and admits the potential for hard-to-trace bias in the hate speech categorization. The study concerned the detection of racism using a Naive Bayes classifier. The work established the definitional challenge of hate speech by showing annotators could agree only 33% of the time on texts purported to contain hate speech. Another considered the problem of detecting anti-Semitic comments in Yahoo news groups using support vector machines [29].

In this work first, authors consider a definition of hateful speech that could be practically useful to platform operators of social media and previous work definitions. Second, develop a general method of hierarchical annotation method shown in figure 1 for selected annotators of 3 PHD,2 MSC students and 1 assistant professor from Amharic Language studies. The annotators were instructed to use the chart originates from the figure 1.In addition to the annotation rules the Kapa decision agreement based on the Cohen's kappaκ statistic which is an estimate of the population coefficient between $0 \leq \kappa \leq 1$ [32] is also used. This work show how the values are interpreted? What does a specific kappaκ value mean?

Table 1: kappaκ values

Nominal	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa vlaue	0.0	0.20	0.40	0.60	0.80	10.

Kappaκ agreement < 0 less than chance agreement, 0.01–0.20 Slight agreement, 0.21- 0.40 Fair agreement, 0.41-0.60 Moderate agreement, 0.61-0.80 Substantial agreement, 0.81-0.99, almost perfect agreement. Not everyone would agree which one is "good" agreement but as commonly cited scale is kappa value of 0.57 is in the "moderate" agreement range for better agreement. Remember that perfect agreement would equate to a kappa of 1, and chance agreement would equate to 0. Given that the majority of comments has been annotated by more than one annotator, authors have also computed the kappaκ inter-annotator agreement metric [33], which measures the level of agreement of different annotators on a task. In this case, considering 1,821 comments that received annotations from all the 6 annotators and obtain = 0.64 when discriminating over two classes and the work of [26] using number of disrespectful words in the text, with a dictionary that extracted from the annotated dataset and identified by the language experts. Then the dataset becomes larger than before which is 6, 120 to be used for this work.

Table 2: samples of kappaκ inter-annotator agreement result

Language	Amharic Comment Text and its English translation	class
Amharic	መፍትሄውጎረቤትህያለውጉትግሬመግደልነው	Hate
English	The solution is to kill the neighboring Tigrian	
Amharic	ኦሮሞየአማራጠላትነው	Hate
English	Oromo is an enemy of Amhara	
Amharic	አንድአማራለሁሉምአማራ	Nohate
English	One Amhara to All Amhara	

4. ARCHITECTURAL DESIGN AND EXPERIMENTATION

An Apache Spark Standalone cluster was used for data preparation and developing models for machine learning classification which is suitable for big data processing like Facebook data. Spark ML pipeline is used in providing a set of tokenization mechanisms. In addition, Spark offers modules for feature selection and machine learning MLLIB library. Python programming language was used for both preparation of dataset and machine learning with RDD and DataFrame file format used as the back end for storing lazy operations which is ideal for large data. Spark designed to efficiently store RDD data while providing powerful MAP, Reduce and filter transformation operations and take actions for further process as shown in the figure 2.

The model was trained using 4,882 posts and used to correctly classify Facebook data according to the two classes mentioned above prepared based on the requirements of Naïve Bayes and Random forest algorithms. These classifier were selected based on previous work result in related work for English and other languages.

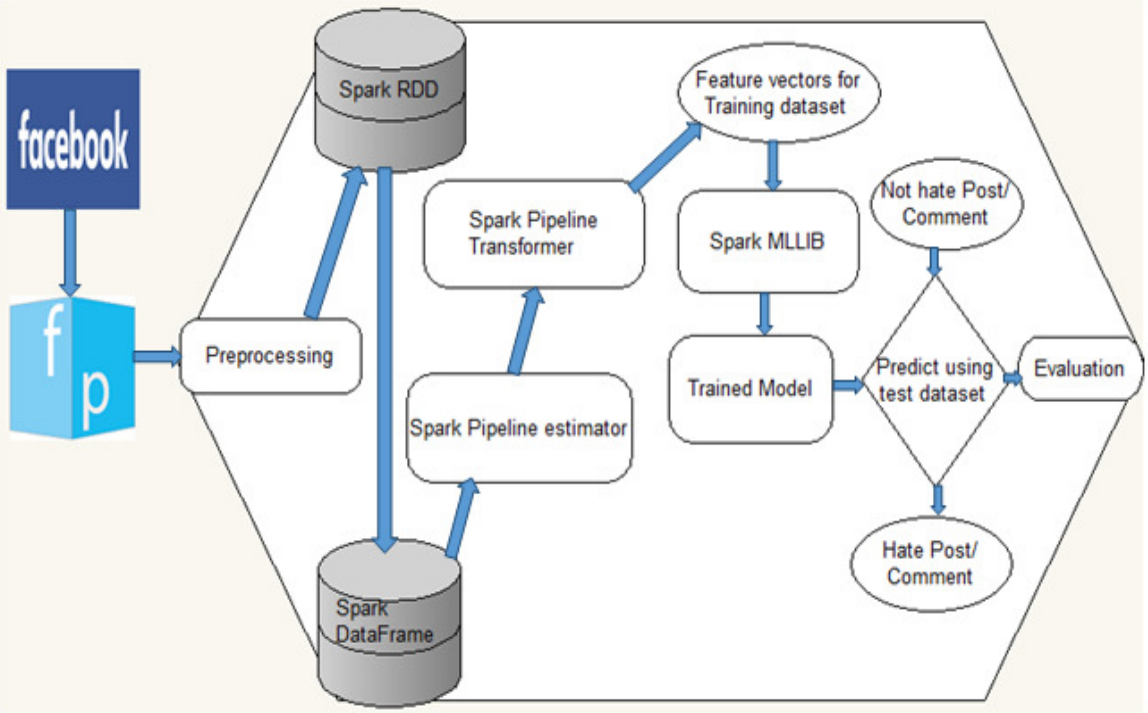

Figure 2: Architectural Design of Amharic hate speech detection

Table 2: Dataset information for the paper (new)

Training Dataset		
Nohate	Hate	Total
2,629	2,253	**4,882**
Test Dataset		
667	571	**1,238**
3,296	**2,824**	**6,120**

4.1 Feature Selection

This involved selecting a subset of relevant features that would help in identifying hate and no hate posts and can be used in the modeling of the classification problems. Authors use Word2Vec as used in [11, 12, 13, 14] for such work and text classification [34]. TF-IDF [16] also used in text classification by different authors for feature selection in other tools. But authors propose to use both of them for Apache Spark feature selection and transformation API. The main feature of interest for this work is comments and posts sentiment of users towards hate speech in social media. The classification is supervised learning task because the objective is to use machine learning to automatically classify comments/posts into categories based on previously labelled comments and posts [11]. Author's contribution is preparation of new dataset, using tf-idf and word2vec as feature extraction, first in its kind, for the Amharic language hate speech detection proficient to big data on spark.

4.2 Model Design and Classification

To develop the model, Spark ML API (spark.ml) which provides ML pipelines (workflow) for creating, tuning, and evaluating of machine learning model was utilized. In Spark ML, a pipeline is defined as a sequence of stages, and each stage is either a Transformer or an Estimator. These stages are run in order, and the input DataFrame with spars vectors were transformed as it passes through each stage.

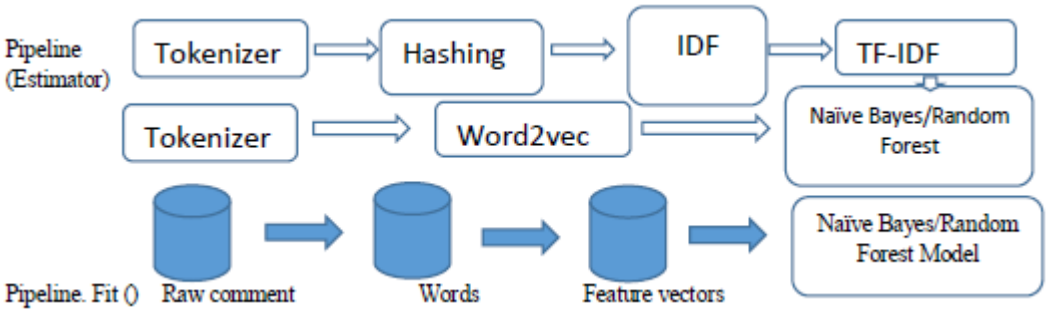

Figure 3: Spark ML pipeline for training (Adopted from Apache spark)

Annotated data were given to the pipeline to get features as feature vectors. The study split the dataset into two datasets, 80% (4882, comments) as training dataset and 20% (1238, comments) as testing dataset using the spark DataFrame random split function with the seed of 100. The training dataset was used to train model, and test dataset was used to evaluate the model performance.

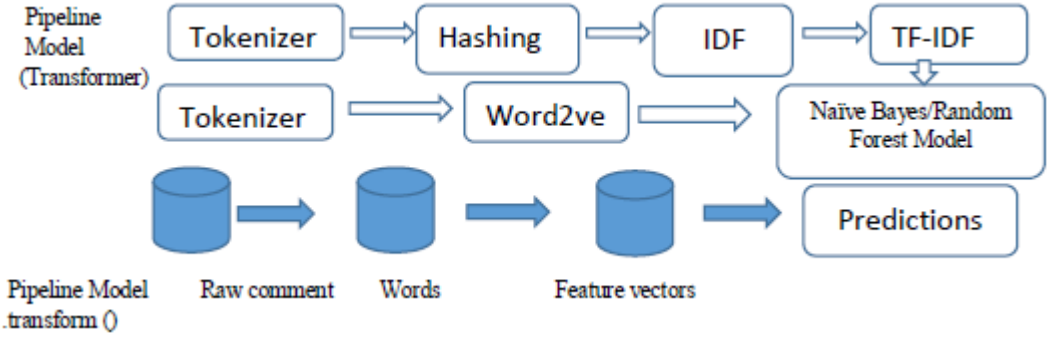

Figure 4: Spark ML pipeline for testing (Adopted from Apache spark)

4.3 Model Evaluation

For accuracy of the model, authors used cross validation using Spark evaluation tool namely Binary class Classification Evaluator within the spark ML. To evaluate the performance of the model classification in terms of quality or predictive effectiveness, different metrics appropriate for the work accuracy, ROC score and Area under curve F-measure (F1-score) were used as shown in table 3.

Table 3: Classification Performance result

Classifier Algorithm	Feature Model	Evaluation Metrics Result		
		Accuracy	ROC score	Area under PR
Naive Bayes	TF-IDF	0.73021	0.8053	0.7993
	Word2Vec	**0.7983**	**0.8305**	**0.8534**
Random Forest	TF-IDF	0.6355	0.6844	0.6966
	Word2Vec	0.6534	0. 7097	0.7307

4.4 Results and Analysis

Authors evaluated classification model by using the 10-fold cross-validation method, achieving an average result as presented in table 3. It was evident that the Naïve Bayes classifier with word2Vec feature model outperform to classify hate and Nohate speech 0.7983, 0.8305 and 0.8534 accuracy, ROC score and area under Precision and Recall respectively with Facebook social network for Amharic language posts and comments. The Naïve Bayes also achieve better result for TF-IDF feature model with 0.73021, .08053 and 0.7993 for accuracy, ROC score and area under precision and recall respectively. The Random Forest with word2vec feature is better than TF-IDF with the result 0.6534, 0.7097 and 0.7307 accuracy, ROC score and area under precision and recall respectively. This is followed by TF-IDF with 0.6355, 0.6844 and 0.6996 respectively. Even though may not be appropriate to compare the result with different experimental setups authors got the state of the art result found in other languages with unique feature of scalability for big data.

The following two charts shows sample of the hate speech classification performance using ROC score area.

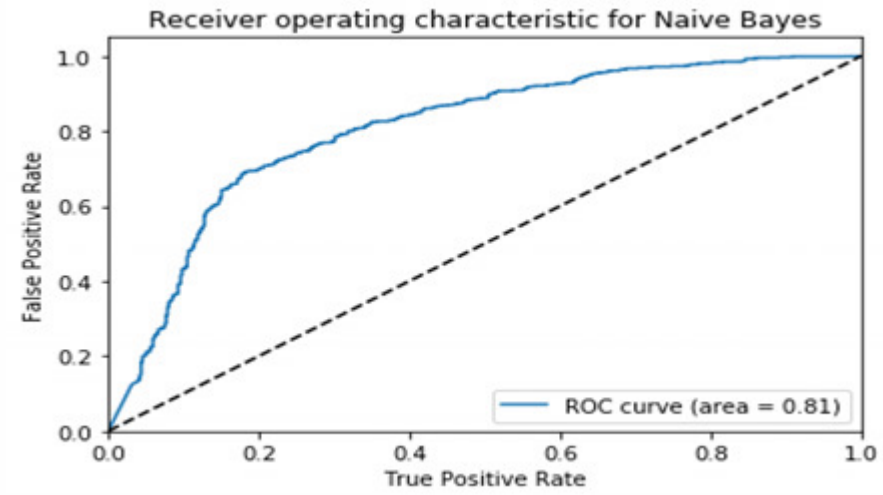

Figure 5: ROC for Naïve Bayes with TF-IDF

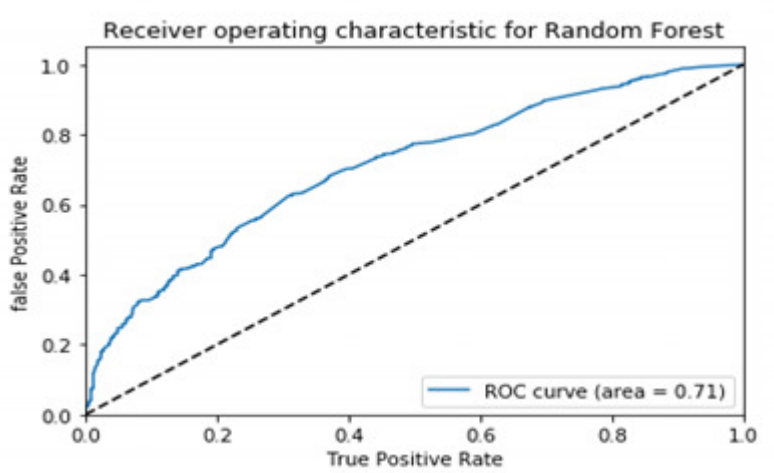

Figure 6: ROC for Random Forest with Word2Vec

5. CONCLUSIONS AND FUTURE WORK

The study developed a model for Amharic text hate speech detection that analyzes posts and comments to identify hate speech using spark machine learning techniques. To conduct the experiments, thousands of Amharic post and comments on suspected social network pages of organizations and individual person's public pages are crawled as dataset. First prepressed according to the requirement of the language and human annotators selected to label the comment in to hate or not hate. Here after, features are pipelined to word2vec neural network tool and TF-IDF in apache spark environment so that feature vectors are obtained.

The classification algorithms were implemented in Apache Spark local cluster using the Apache Spark's Machine Learning library. The model developed using Naïve Bayes and Random forest utilizing a dataset of 6,120 Amharic posts and comments out of this 4,882 to train the model and 1,238 for testing after passing different steps as stated in the experiment section. The model was tested to classify whether the post and comments are hate or not and able to detect and classify in an accuracy of 79.83 % and 65.34% for Naïve Bayes with word2vec feature vector and Random Forest with TF-IDF feature modeling approach respectively. The workshow that word2vec feature model is better in maintaining the semantics of the posts and comments as proved in other works. The result are promising for such work in social network big data which can be extended to compute large volumes data since the work used the distributed platform of apache spark.

Even if the results are promising for hate detection, our research is far from perfect. A lot of work ahead of us to work on technical improvements that can be made for the language interms of: (1) expand the dataset that would reduce the risk of overfitting and improve the statistical significance of the results (2)analyzing the different aspect of the category of hate, either hate with politics, ethnicity, religion and socio-economy (3)utilize the information provided by Facebook so that, classification can be improved by expanding the feature space with profile information, list of followers and geolocation etc. (4) crawl other sources to improve the feature space for such under resourced language for computational purpose by adding synonyms from other sources such as Twitter, forums and other homepages.

Finally, the proposed methods could be applied in different domains where the posts about the anticipation to get service and buy product by the review of the service after serving or buying it for this particular language showing the sentiment of costumers as positive or negative can be explored.

REFERENCES

[1] Bjorn Gamback and Utpal Kumar Sikdar, Named Entity Recognition for Amharic Using Deep Learning. IST-Africa 2017 Conference Proceedings Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, ISBN: 978-1-905824-57-1, 2017

[2] META-NET White Paper Series, Retrieved from Multilingual Europe Technology Alliance: http://www.meta-net.eu/whitepapers/overview , (2018, January Tuesday)

[3] Grover Hudson, Linguistic analysis of the 1994 Ethiopian census, Northeast African Studies, 6(3):89–107, 1999

[4] Raphael Cohen-Almagor. Internet History, International Journal of Techno ethics, Vol. 2, No. 2, pp. 45-64, 2011

[5] Waseem & Hovy, Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of NAACL-HLT, 2016, pages 88–93

[6] White Paper series, Retrieved from Amnesty International, Social media and Internet: https://www.amnesty.org/en/latest/news/2016/12/ethiopia-government-blocking-of-websites-during-protests-widespread-systematic-and-illegal/, 2016, 2018.

[7] Saleem, Haji Mohammad, Kelly P. Dillon, Susan Benesch, and Derek Ruths, A web of hate: Tackling hateful speech in online social spaces. ArXiv preprint arXiv: 1709.10159, 2017.

[8] Fortuna, Paula Cristina Teixeira, Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes, 2017.

[9] Kwok, Irene, and Yuzhou Wang, Locate the Hate: Detecting Tweets against Blacks, In AAAI. 2013

[10] Del Vigna12, Fabio, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi, Hate me, hate me not: Hate speech detection on Facebook, 2017.

[11] Silva, Leandro Araújo, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber, Analyzing the Targets of Hate in Online Social Media, In ICWSM, pp. 687-690, 2016.

[12] Waseem, Zeerak, and Dirk Hovy, Hateful symbols or hateful people? Predictive features for hate speech detection on twitter, In Proceedings of the NAACL student research workshop, pp. 88-93, 2016.

[13] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang, Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, pages 145-153. International World Wide Web Conferences Steering Committee, 2016.

[14] Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber, Automated hate speech detection and the problem of offensive language, arXiv preprint arXiv: 1703.04009, 2017.

[15] Yashar Mehdad and Joel Tetreault, Do characters abuse more than words? In Proceedings of the SIGdial 2016 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 299–303, 2016.

[16] Keith Cortis and Siegfried Handschuh, Analysis of cyberbullying tweets in trending world events. In Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business, page 7. ACM, 2015.

[17] Agarwal, Swati, and Ashish Sureka, Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website, arXiv preprint arXiv: 1701.04931, 2017.

[18] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv: 1301.3781, 2013.

[19] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, Distributed representations of words and phrases and their compositionality, In Advances in neural information processing systems, pp. 3111-3119, 2013.

[20] BjornRoss, MichaelRist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki, Measuring the reliability of hate speech annotations: The case of the European refugee crisis, 2017.

[21] Stephan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans, A dictionary-based approach to racism detection in Dutch social media, 2016.

[22] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi, Hate me, hate me not: Hate speech detection on Facebook. In Proceedings of the First Italian Conference on Cybersecurity, pages 86–95, 2017.

[23] Liu, Shuhua, and Thomas Forss, Combining n-gram based similarity analysis with sentiment analysis in web content classification, In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-Volume 1, pp. 530-537. SCITEPRESS-Science and Technology Publications, Lda, 2014.

[24] Liu, Shuhua, and Thomas Forss, New classification models for detecting Hate and Violence web content, In Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on, vol. 1, pp. 487-495. IEEE, 2015.

[25] Maloba, Wilson Jeffrey, Use of regular expressions for multi-lingual detection of hate speech in Kenya, PhD diss., iLabAfrica, 2014.

[26] Njagi Dennis Gitari, Zhang Zuping, Hanyurwim fura Damien, and Jun Long. A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering, 10(4):215–230, 2015.

[27] Nemanja Djuric, JingZhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati, Hate speech detection with comment embed dings. In Proceedings of the 24th International Conference on World Wide Web, pages 29–ACM2, 2015.

[28] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma, Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[27] Maral Dadvar, Franciska de Jong, Roeland Ordelman, and Dolf Trieschnigg, Improved cyberbullying detection using gender information. In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop, pages 23–25, University of Ghent, 2012.

[28] Shuhan Yuan, Xintao Wu, and Yang Xiang, A two phase deep learning model for identifying discrimination from tweets, In International Conference on Extending Database Technology, pages 696–697, 2016.

[29] Kwok and Wang, Detecting Tweets against Blacks. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013

[30] William Warner and Julia Hirschberg, Detecting Hate Speech on the World Wide Web. Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pages 19–26, Association for Computational Linguistics, Montreal, Canada, 2012

[31] Anna Schmidt and Michael Wiegand, A survey on hate speech detection using natural language processing. Social NLP 2017, page 1, 2017.

[32] Anthony J. Viera, Understanding Inter observer Agreement: The Kappa Statistic, From the Robert Wood Johnson Clinical Scholars Program, University of North Carolina, 2005

[33] Kilem L Gwet. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.

[34] Joseph Lilleberg et al, Support Vector Machines and Word2vec for Text Classification with Semantic Features. Proc. 20151IEE 14th Inl'l Coni. On Cognitive Informatics & Cognitive Computing IEdsJ, 2015.

[35] United Nations Educational, Scientific and Cultural Organization, Countering Online Hate Speech. Published in 2015 by the United Nations Educational, Scientific and Cultural Organization 7, place de Fontenot, 75352 Paris 07 SP, France

[36] Iginio Gagliardone, Alisha Patel and Matti Pohjonen, Mapping and Analyzing Hate Speech Online:Opportunities and Challenges for Ethiopia, 2014

AppendixA

Experiment setup

Apache environment setup	
Apache Spark-2.2.0	Ubuntu 16.4 virtual machine Standalone clustering mode 1 master node 2 worker node
Feature selection Algorithms	
Name of Algorithm	Parameter set up
Word2Vec	choice of training model 0: Skip gram model dimension of vectors=3 Window size=5 Minimum count=0
TF-IDF	Default
Machine learning Algorithms	
Name of Algorithm	Parameter set up
Naïve Bayes	Model type =multinomial Smoothing =1
Name of Algorithm	Parameter set up
Random Forest	NumTree=200 MaxDepth=3 Seed=2
10-fold cross-validation	
Name of Algorithm	Parameter set up
Naïve Bayes	Smoothing =1 Numfolds=10
Random Forest	NumTree=[50,100,200] MaxDepth=[3,4,5] Numfolds=10