ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

Department of Computer Science and Engineering

Engineering Research and Development Methodology (CSE-4221)

a Research Proposal on

HATE SPEECH DETECTION ON ETHIOPIAN SOCIAL MEDIA TEXT USING SENTIMENT ANALYSIS

Participants:

1. **AREFAT HYEREDIN** **A/UR5082/09**
2. **GIRUM GETACHEW** **A/UR4045/09**
3. **BINIYAM GOSSAYE** **A/UR4861/09**
4. **GIZEALEW ENDESHAW** **A/UR3981/09**
5. **ABERHAM BEKELE** **A/UR4499/09**

Submitted to: Mr. Anteneh Alemu **Due Date: December 13, 2019**

Hate Speech Detection on Ethiopian Social Media Text using Sentiment Analysis

1. INTRODUCTION

In recent years, social networking has grown and become prevalent with every people, it is simpler for people to interact and share with each other. However, every problem has different sides. It also has some negative issues, hate speech is a crucial topic in the domain of social media. With the freedom of speech on social networks and anonymity on the world wide web, people are free to comment on hate, insults, fake news, and disinformation. Hate speech can have an adverse effect on human behavior and livelihood and it also affects societal queues. We can't manually censor, delete or control all these comments, posts, news feeds, reviews and updates; which would be time-consuming, tedious and boring. This spurs research to build to build an automated system that detects, tags and eliminate such content. With said system, we would be able to detect and monitor the spread of hate speech and disinformation on social media and thus reduce the dangerous consequences of societal conflicts, displacement and civil war. With Ethiopian indigenous languages such as Amharic and Oromiffa, we can use sentiment analysis application on social media with other natural language processing methods for the detection.

2. PROBLEM STATEMENT

A major obstacle 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 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. Ethiopian languages are among the languages 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 proposal is aimed to address the need for such high level computational linguistic model to combat hate speech from social media posts and comments. Recent advances in mobile computing and the Internet have resulted in an increase in use of social networks to communicate, express opinions, interact with other. While social media provides an important avenue for communication and sharing, it also acts as a means of spreading hate speech online. Inherent characteristics of the Internet largely contribute to the misuse of social networks 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, Facebook and others was blocked, as well as at least 16 news outlets. It is an open secret that the recent widespread hate speech and call for violence particularly targets persons of a particular group. 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.

A recent study defines 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. 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.

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 sentiment analysis in hate speech detection to reduce the challenges.

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text. The entities can be products, services, organizations, individuals, events, issues, or topics.

3. OBJECTIVE OF THE STUDY

The general and specific objectives of the study are given below:

General Objective: the general objective of this research work is to design and develop a sentiment analysis or opinion mining model to detect hate speech content from social media textual contents.

Specific Objective: the specific objectives of this research work are:

- Analysis of the general structure of Amharic and Oromiffa statements related to opinions and sentiments such as identifying, negative, positive and neutral statements.
- Analyze the semantics between opinion expressions across languages of Amharic and Oromiffa and their intensity.
- Design a model for sentiment mining to classify social media texts
- Develop necessary algorithms to realize the proposed model in developing a sentiment mining model to detect hate speech
- Building domain specific and general-purpose lexicon of Amharic and Oromiffa language's opinion terms where these terms are tagged accordingly.
- Develop a prototype to demonstrate that the model design is valid.
- Design an auto-tagging corpus for offensive hate speech content using third party APIs provided by the social network.
- Evaluate the model designed on texts from social media

4. LITERATURE REVIEW

Hate speech detection related researches are done for different language opinionated social media textual content such as English, Chinese and French using different techniques and approaches are reviewed. Different authors used different techniques such as machine learning, ontology-based approaches, rule-based approaches, sentiment analysis, lexicon-based approaches, feature-selection, cross-validation and others. In addition to the techniques, the employed approaches, goals, motivation, domain, target language, dataset source, procedures, experimental results, performance, and challenges are the main points given focus when going through the different works. This proposal introduces sentiment analysis techniques to detect hate speech from Ethiopian indigenous languages particularly Amharic and Oromiffa; which makes it ideal since the contents are mostly of sentimental value or opinion and also the speeches are analyzed from social media where sentiment analysis is widely practical upon social media analysis.

After consulting those papers, we generalized that these dimensions of analysis model detection of hate speech into general-feature and specific feature approaches. Approaches from the literature reviews of related works are generalized as follows:

Dictionaries and lexicons: The majority of the papers found try to adapt strategies already known in text mining to the specific problem of hate speech detection. The work categorizes 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, number of disrespectful words in the text, with a dictionary that consists of words for English language including acronyms and abbreviations, label specific features which consisted in using frequently used forms of verbal abuse as well as widely used stereotypical words.

Bag-of-words (BOW): Another model similar to dictionaries is the use of bag-of-words. 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. In study character n-gram features proved to be more predictive than to kenn-gram features, for the specific problem of abusive language detection.

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.

Word Embedding: Deep learning techniques are recently being used in text classification and sentiment analysis with high accuracy. 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.

This section reviewed different research's perspective to tackle hate speech using sentiment mining for different languages. The review showed that machine learning, ontology based and lexicon- based are the commonly used approaches to deal with sentiment mining. The works reviewed indicated that the approaches except the machine learning rely on tagged list of positive or negative sentiment terms to identify the polarity of terms. The machine learning technique is based on the concept of training the machine to learn to classify opinionated texts into predefined categories of positive, negative or neutral. Ontology is employed particularly to extract feature of an object for the purpose of refining feature level sentiment analysis. The lexicon based approaches are based on the concept of counting the sentiment terms available in the opinionated texts. It can be summarized that sentiment analysis provides various methodologies and also ideal for this particular research case.

5. METHODOLOGY

The primary research method for this study is literature review and conceptual modeling. Constraint identification and classification through a structured approach is productive way to test out different algorithms that are offered by Sentiment Analysis. This study will first review various types of text analysis to determine and distinguish the structure of hate speech in the model.

A survey of different demography will be commenced in order to gain a perspective of hate speech from the public with questionnaires, and also a sample textual data collection from different social media like Facebook, Telegram from public pages, individual politicians, activists, News pages and various group pages; posts and comments from different timelines will be retrieved to build the corpus. 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. A versatile Facebook crawler, which exploits the Graph API to retrieve the content of the comments from Facebook posts using Facepager would be employed. 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.

A comprehensive model of sentiment analysis using one of the following algorithms: TI-IDF, Naïve-Bayesian, Random Forest, Logistic Regression and also string labeling algorithms like Conditional Random Field(CRF), Model Hidden Markov (HMM), Word2Vec, Bidirectional Long Short-Term Memory (Bi-LSTM) or Entropy. However, we have to choose the features

manually to bring the model with high accuracy. The model would undergo multiple tests for multiple text samples to classify them accordingly, and tag them as efficiently as possible.

6. SIGNIFICANCE OF THE STUDY

With the advent information age, people are puzzled with vast amount of information from the Internet. Social media is a highly active platform which is challenged with different issues like hate speech and disinformation. Today, social media in Ethiopia is full of turmoil where hate is a concern at every turn and scroll. People are openly defaming, demeaning, or devaluing others which is inhumane. Hate speech detection implemented at feature level would create a monitored environment that no hate speech is tolerated and exercised; creating a safe and tolerable web environment where everyone respects and interacts peacefully. Hate speeches that are tagged would be investigated and people exercising it would become accountable for their actions which may create catastrophic results. In the long run, hate-free speech would build societal unity and peace for all people. It is also meant to create awareness about the seriousness of the action and its dangerous outcome.

7. DURATION AND PLAN OF ACTION

This study upon approval of the proposal will be conducted between December 13, 2019 and January 3rd,2020. Due to the limited time for the research, all tasks and objectives may not be fulfilled.

	Dec 13 -18	Dec 19 – 23	Dec 24 – 27	Dec 30 – Jan 3
Deep Literature Review				
Survey				
Data Collection				
Design & Modelling				
Development				
Testing				

8. BUDGET / COST

There is no official budget analysis for this research. Yet, here is an estimate cost:

Printing and Paper – Research and Survey papers = ETB 200

Professional NLTK Spark tools – Licensing fee (USD 20) = ETB 600

Internet connection – 10 hours = ETB 150

Miscellaneous cost – others = ETB 250

Total Cost = ETB 1200

9. REFERENCES

[1] L. Bing, “Sentiment Analysis: *Mining Opinions, Sentiments, and Emotions*.,” University of Illinois at Chicago, pp 1-4, Cambridge University Press, 2*015.*

[2] S. Gebremeskel, “Sentiment mining model for opinionated Amharic text,” Addis Ababa University, Dept of Computer Science, November 2010.

[3] H. Thi-Thuy Do, H. Duc Huynh, K. Van Nguyen, N. Lu-Thuy Nguyen and A. Gia-Tuan Nguyen, “Hate Speech Detection on Vietnamese Social Media Text using the Bidirectional-LSTM Model,” University of Information Technology, VNU-HCM, November, 2019.

[4] S. Rosenthal, A. Ritter, “SemEval-2014 Task 9: Sentiement Analysis in Twitter,” Columbia University, Johns Hopkins University, December, 2019.

[5] Z. Mossie, J. Wang, “Social Network Hate Speech Detection for Amharic Language,” National Taipei University of Technolgy, Taipei, Taiwan, 2018.