SUBJECTIVITY AND SENTIMENT ANALYSIS OF AMHARIC COMMENTS ON SOCIAL MEDIA: THE CASE OF ETHIOPIA POLITICAL DISCOURSE

A Thesis Presented

by

GEBRIEL GIZATE MOLLA

to

The Faculty of Informatics

of

St. Mary's University

**In Partial Fulfillment of the Requirements
for the Degree of Master of Science
in**

Computer Science

June 2020

ACCEPTANCE

SUBJECTIVITY AND SENTIMENT ANALYSIS OF AMHARIC COMMENTS ON SOCIAL MEDIA: THE CASE OF ETHIOPIA POLITICAL DISCOURSE

By

GEBRIEL GIZATE MOLLA

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

____

Internal Examiner
{Full Name, Signature, and Date}

____

External Examiner
{Full Name, Signature, and Date}

____

Dean, Faculty of Informatics
{Full Name, Signature, and Date}

{Date of Defense}
June 2020

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

____
GEBRIEL GIZATE MOLLA

____
Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as an advisor.

____
Full Name of Advisor

____
Signature

Addis Ababa

Ethiopia
Exact Date of Defense

January 2020

ACKNOWLEDGMENT

Foremost, I would like to thank Dr. GETAHUN SEMEON, my advisor, for his encouragement, patience, brilliant guidance, constructive suggestions, and advice to finalize this thesis work. I would like to thank my workmates that helped through this work especially on preprocessing steps and supported me in one way or another. I also express my gratitude to Dr. Milion Meshesha on his support to encouraging me on the research work and gave his valuable comments. Finally, I would like to thank My parents and friends especially my father GIZATE MOLLA DESTA receive my deepest love for being the strength in me.

TABLE OF CONTENTS

LIST OF ACRONYMS

ANN	Artificial Neural Networks
API	Application Processing Interface
BOW	Bag of Word
BRM	Brand Reputation Management
EBC	Ethiopia Broadcasting Corporation
ESAT	Ethiopian Satellite Television
FB	Facebook
IR	Information Retrieve
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
SA	Sentiment Analysis
SVM	Super vector Machine
TF/IDF	TermFrequency/ Inverse Direction Frequency
TV	Television
VOC	Voice of Customers
VOM	Voice of Market

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The rise of social media and the rapid development of mobile communication technologies have dramatically changed the way to express the feeling, attitude, lifestyle, opinions, events, situations (like based on political, economic, and social point of view), passion, etc. People often express these behaviors through social media in the form of short texts. Concerning events and situations on behalf of the political sector, social media can be an enabler for participation and democracy among citizens of participating groups. This study is conducted to perform and gain insight into political sentiment using machine learning technique from the comment given in Amharic language on the topic of the idea which is posted through social media so as to understand the subjectivity of the comment. The study try to demonstrate the analysis of the sentiments and it's a subjectivity to identify their trends, numerical implications. The data that is considered in this research is based on two data sets that are self-build, The first one is a data set for the subjectivity of the comment and the other one is for sentiment analysis getting after subjective sentence or comments that are extracted but the main content of data set is the same. Finally, the results show the trends in social media users comment sentiments and subjectivity correlation to the real-world events associated with the respective keywords and provide a clear picture of the influence of real-world events on social media user's comment and it's sentiments of polarity.

Keywords: Sentiment analysis, Subjectivity, Social media, Natural language processing, Political discourse

CHAPTER ONE

INTRODUCTION

1.1. BACKGROUND

In the field of computer science, Natural Language Processing (NLP) is the major and recent research field of study. Computers nowadays having a greater improvement over the capability of NLP processing to increase their ability to understand, interpret, and communicate using human language. There has been a lot of work done and being done to incorporate these features of communication with computers. And from their application of features, sentiment analysis and subjectivity classification has been studided by many researchers .

Sentiment analysis(SA) is a field dedicated to extracting subjective emotions and feelings from the text.[5]. One common use of sentiment analysis is to figure out if a text expresses negative or positive feelings. It involves several research fields; such as natural language processing, computational linguistics, and text analysis.[6]. It refers to the extraction of subjective information from raw data, often in text form, and aims to assign a predefined sentiment class to online texts as negative, positive, or neutral. SA plays a substantial role in several domains such as financing, marketing, politics, and social. One of the main approaches used to solve the SA problem is the supervised machine learning (ML) approach.[20]. In this approach, texts are represented by feature vectors that are used to train ML classifiers, such as naïve Bayes (NB) and support vector machines (SVMs), to infer a combination of particular features yielding a certain sentiment class.

Subjectivity detection is, the task of determining if a piece of text contains opinions or not (i.e. subjective expression or objective).[5]. It is not so much about determining the polarity of the text itself. People always try to express sentiments about products, brands, ideas, situations, events, or services through social media.[5,6,21]. These messages contain an opinion about a specific subject. The sentiment of this opinion can be classified into different categories, such as of positive, neutral, and negative. In this interesting research area, this study tries to works on sentiment analysis of political comments in Ethiopia social media especially Facebook and their subjectivity related to their post.[26]. The first task is since the main problem is how we use the social media comments and the sentiment embedded in user comments relates to the topic of the post and to which percent sentiment polarity it has and how political comment polarization makes people vulnerable to disinformation or unrelated comments or in turn how does the increasing prevalence of dis-information lead to greater political polarization. As we look at different social and political researches especially on the use of social media, participations are based on without knowing ideas of the post. And more users commenting with bias, instead of dealing with someone with the ideology of politics they communicate using bad words, or we can say instead of understanding what the post is telling about simply based on someone's comment commenting the idea. [30].

Particularly after the coming of the Internet and the social media platforms, it is argued that media have become an integral part of interactions among different sections of people and individuals. This, in some ways, indicates that the change in the communication media seems to change the political narration in the global as well as in our country's context. The challenge, however, is that unlike the communication habit within the conventional media sphere, on social media, according to TanaseTasente, "communication is routed by the online opinion leaders" and their zealous followers rather than by professionally-oriented communication experts. For this reason, Facebook and the other social media platforms usually mediatize political and other related issues that are pertinent to the subjective interests of the general user and of some influential online writers (or opinion leaders).As a result of which, it is possible to argue that our perception towards the political context in our country appears to be (consciously or unconsciously) substantially shaped by the unrefined and subjectively framed messages from the social media than by the relatively 'objective' messages mediated through the traditional media. This kind of political communication also causes continuing (ethnic and political) controversies in our country. [1].

To answer the above mention problems it is proposed to work on first to collect the comments from different Ethiopia news and political Facebook pages using Facebook API, Python and R languages and then prepare this collected data for feature extraction, to detect subjectivity of comments to the related post, the sentiment of the comment. And finally based on this insight and alarm, messages can be forwarded for the government to works on social media users especially on youths who are ready for joining social media. Additionally, policymakers can look at what the sentiment of the people on particular ideas and react to the existing situation by forwarding relevant policy.

1.2. STATEMENT OF THE PROBLEM

The benefits of communication technologies in education, informatics, politics, economics, and entertainment, and other sectors are vast and enormous. Nevertheless when these communication systems are misused the rapid flow of false information through social media or another communication system could destroy individuals, society, and nation and business institutes at the global and national levels. [3]. The social networking or media awareness of American or European societies is much higher than Africa societies, in particular Ethiopian society. Current misuse on the social network by a few Ethiopian individuals and groups together imposes dangers on our society and country.

At present multiple hate speeches, videos, and images are continuously posted, shared, and circulating among Ethiopians on social media. This misinformation has a potential impact and could spark intolerance and hate among members of society one against others. Certain ethnic group hate and intolerance will gradually lead to ethnic or political violence. So to solve this problem it is important to have integration with quantitative and qualitative research sentiment analysis which facilitates deep rich insight into biased comments, hate speeches, trends of usage of social media, unsolicited opinions, thus facilitating a more meaningful understanding of how we use social media on Ethiopian political discourse run on Facebook. We should openly challenge the hate preachers, emphasize our rights as citizens irrespective of our ethnicity, foster an honest and open debate on all possible socio-political events, and engage in genuine reconciliation.

There are research works done in Amharic sentiment analysis [18]. But to the extent of these study reviews, taking sentiment Analysis and subjectivity classification of Amharic social media comments has never been studied. Therefore, these studies analyze the sentiment of Amharic political comment on social media and subjective of the comments to the post-it belongs.

1.3. OBJECTIVE OF THE STUDY

1.3.1. GENERAL OBJECTIVE

The main objective of this study is to identify and analyze sentiment and subjectivity of social media comments given on political news or post in Ethiopia using a machine learning approach and lexicon approach and to identifiy the impact of social media comments on political discourse and trend of comments.

1.3.2. SPECIFIC OBJECTIVES

Towards achieving the general objective of the study that deals with sentiment classification and subjectivity classification, the following specific objectives are formulated:

- To build both domain-specific and general-purpose corpus of Amharic language opinion terms where these terms are tagged as positive and negative.
- To develop the necessary algorithms to realize the proposed model in developing an Amharic sentiment analysis and the subjectivity classifier model.
- To Build a corpus from preprocessed data for sentiment and subjectivity analysis of the text of Amharic language.
- To Develop subjectivity classifiers using collected texts for training.
- To Develop a prototype to demonstrate that the model designed is valid.
- To evaluate the performance of the prototype designed in this study.

1.4. SIGNIFICANCE OF THE STUDY

At most basic, sentiment analysis is a social media analytics tool that involves checking how many negative and positive keywords are present in a chunk of conversation. If there are more positive keywords than negative, it is considered positive content. If there are more negative keywords, it is called negative content. But there's a lot more to it than that and its real worth is found in the details. In this study of sentiment analysis and classification in-depth analysis involves finding opinions of political comments given on social media content and extracting the sentiment they contain with the post they belong to. An opinion is made up of a target, in these study political comments on social media, and subjectivity on the topic of the comment. And the expected result will defend and analyzed view of political comments and their real-world effect on the degree of sentiment, what contribution they will have on usage of social media, and finally linkage of a post and their comments in our country usage of social media.

We can respond quickly to a crisis, bad experiences, and problems with political comments and posts given on social media. It provides a way to manage our online reputation by taking appropriate action with speed. One very practical use in our study area, People pay attention to comments and posts and buy more of the attention that has the most participation going on and Using sentiment analysis, we will be alerted to negative sentiments as they happen, allowing you to respond quickly. We can influence the conversation which is positive sentiment by biasing when we know which way it is trending the comment to make it positive.

Politicians can gain the trust and loyalty of their followers by reacting to their comments promptly, which tells the user and follower that they care about them. It helps to measure the success of a specific or campaign soon after it appears for political parties

It acts as an important resource for political and ideological research. Knowing that people like the idea of politicians help to figure out how to strengthen the ideology of the party that the party follows. Politicians can get a good insight into how social media followers stacks up against that of a competitor, in the opinion of consumers. This helps policymakers to select what way to run up the idea.

We can check our country’s political news pages, politicians, and social media user’s virtual popularity (like likes, shares, and followers). And it will have a speedy, accurate, and affordable output for policymakers, politicians, and researchers.

1.5. SCOPE AND LIMITATIONS OF THE STUDY

The scope of this study focuses on solving the problem mentioned in the problem statement that occurs in current Ethiopian social media comments given to political comments and their political news or posts.

The limitations of the study include:

- The privacy of the political news page will be a challenge for the study. Here we will conduct the challenge or these limitations with the help of different companies, advisors, etc.
- issues that require special attention when we are dealing with sentiment analysis work include:
  - Sarcasm: is one of the most difficult sentiments for automated tracking to interpret properly. Example: “It was awesome for the week that it worked.”
  - Relative sentiment: is not a classic negative, but can be a negative nonetheless. Example: “I bought an iPhone” is good for Apple, but not for Nokia.
  - Compound or multidimensional sentiment: contain positives and negatives in the same phrase. Example: “I love Mad Men, but hate the misleading episode trailers.”
  - Conditional sentiment: includes actions that may happen in the future. Example: the customer isn’t angry now but says he will be if the company doesn’t call him back.
  - Positive feelings can be unrelated to the core issue. For example, many comments about actors focus on their personal lives, not their acting skills.
  - Negative sentiment is not necessarily bad: This relates to the classic PR dilemma regarding negative publicity. Example: Sarah Palin’s appearance on the Today show generated many negative comments but still drove rating increases.

- Named Entity Recognition - What is the person talking about, e.g. is 300 Spartans a group of Greeks or a movie?
- Abbreviations, poor spelling, poor punctuation, poor grammar, …
- Speaker's emotional state: The speaker's emotional state may or may not have the same polarity as the opinion expressed by the speaker. [17].

1.6. ORGANIZATION OF THE REST OF THE THESIS

The thesis report is organized as follows:

Capter 1: Introduction: chapter one describes the highlited content of the paper work and it contains the proposal work of the thesis that we were planned to achieve this thesis work.

Chapter 2: Literature review and Related work: gives an overview of social media analytics, sentiment analysis, and the researchers concluded in this field. Finally presents gap analysis and summary of the chapter.

Chapter 3: Methodology: an overview of the research method and research design that the thesis work follows. It examines the different sentiment analysis works like preprocessing, methods of data analysis, and presents subjectivity and sentiment analysis challenges.

Chapter 4: Implementation: presents the experimental results we achieved through our works, datasets that prepare for sentiment and subjective lexicon, and finally discuss the experiment result.
• **Chapter 5: Conclusion and Future Work:** It concludes the thesis and mentions the possible direction for future work.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORKS

2.1. OVERVIEW

This section of the study tries to look at different works that have been done before in the research area; and key concepts are discussed concerning this work. In the first part, the study tries to look at what social media analytic mean, its purpose, subject of analysis and application, components and branches of Natural language processing with related to social media analysis. Secondly, the study also discuss; what are different classifications of sentiment analysis and subjectivity classification approaches? And what different methods, the algorithm that has been used in the same research articles and can be used for this study and finally related works and research gap analytics these studies try to identify will be discussed.

2.2. SOCIAL MIDEA ANALYTICS

Social media have been adopted by many businesses. More and more companies are using social media tools such as Facebook and Twitter to provide various services and interact with customers. As a result, a large amount of user-generated content is freely available on social media sites. To increase competitive advantage and effectively assess the competitive environment of businesses, companies need to monitor and analyze not only the customer-generated content on their social media sites but also the textual information on their competitors' social media sites, [39].

Social media analytics is the approach of gathering data from social media websites, blogs, and analyzing that data and evaluate using social media analytics tools to make decisions on the problem domain. The most common use of social media analytics is to mine user's_sentiment through comments and posts to spot trends and to see what's going on. The process goes beyond the usual monitoring or a basic analysis of “tweets”, “comments” or "likes" to develop an in-depth idea of the social consumer.

Analyzing publicly available content on various social media sites such as Facebook, YouTube, and Twitter, as well as social network sites such as Facebook, has become an increasingly popular method for studying socio-political issues. [23]. Those contents primarily available as comments and people express their opinions and sentiments on a given topic, news-story, or post while allowing social and political scientists to extend their analysis of political discourse to the social sphere.

There are lots of readymade tool for analytic purpose but more of them are workings on only the English language. Some of the best social media analytics tools currently available includes []; Rival IQ, Agorapulse, Zuum, SoTrender, Quaintly, which provides analytics purposes for all major networks, including Facebook, YouTube, Instagram, and Twitter. Social media is a good medium to understand consumer choices, intentions, and sentiments. Social media analytic have some purposeful target to study; from those targets table 2.1 presents some of the target and their descriptor which belongs to the science.

Social Media Analytics Targets	**Example of Descriptors**
Audience Size	Does the size of the audience matter?
Audience Profile	For example, add interest and see how many of your fans have that interest.
Traffic of usage	For example, if you are a media site you get paid for advertisements and more traffic means more money!
Content Analysis	We need to analyze the content to see what’s working/not working. Are videos, pictures, or text updates working best?
Community participation	If we’re not responsive to the community they’ll stop interacting so it’s important to measure this.
Competitor achievements	Compare the content to competitors by monitoring stats, audience profile, audience size growth, etc.
Sentiment Analysis	This is where we analyze positive, negative, or neutral mentions of your brand, product or service

Table 1 social media analytic targets and examples

In modern times, the political arena presents itself among numerous types of platforms. Society gains much of its information in regards to politics from social media outlets. Specifically, Facebook greatly impacts the way society interacts within the realm of politics.[22]. From those mentioned targets of social media analytic as described above these studies try to work on targets of sentiment analysis and subjectivity analysis of Amharic comments given on Facebook political posts.

2.3. NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is a field of computer science, artificial intelligence (also called machine learning), and linguistics concerned with the interactions between computers and humans through (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Natural language processing techniques play an important role to get accurate sentiment analysis. [6].

By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation and to analyze, understand and derive meaning from a human language such as Amharic**,** English, Spanish, Hindi, etc.

NLP algorithms are typically based on machine learning algorithms. [5,6,20]. Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analyzing a set of examples (i.e. a large corpus, like from a book to a collection of sentences), and making a static inference.

“One of the most compelling ways NLP offers valuable intelligence is by tracking sentiment — the tone of a written message (tweet, Facebook update, etc.) — and tag that text as positive, negative or neutral,” [36]. NLP involves two major branches that help us to develop NLP applications. [5,6,20,21,30]. One is computational, the Computer Science branch, and the other one is the Linguistics branch. (see figure 2.1)

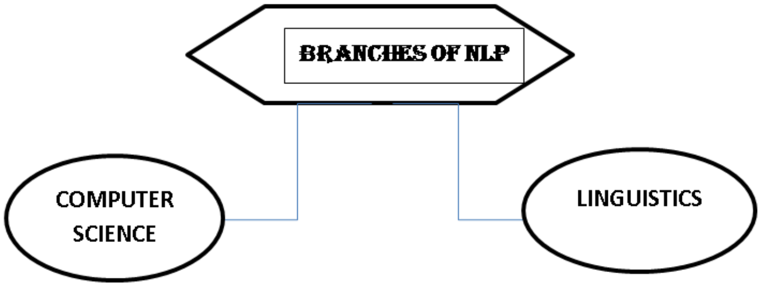


Figure 1 Branches of NLP

The Linguistics branch focuses on how the NL can be analyzed using various scientific techniques. So, the Linguistics branch does a scientific analysis of the form, meaning, and context. Here linguistics analysis can be implemented with the help of computer science techniques. We can use the analysis and feed elements of analysis in a machine-learning algorithm to build an NLP application. Therefore, by combining these two branches we can gain the fruits of science.

Knowing what customers are saying on social media about a concerned entity on some identified sector can help the sector continuing to offer good insight, product, service, or customer experience is the need of the business. NLP makes monitoring and responding to that feedback easily.

2.3.1. COMPONENTS OF NLP

There are five main Component of Natural Language processing, which include the following []:

i. MORPHOLOGICAL AND LEXICAL ANALYSIS

The process of finding morphemes of words is called morphological analysis. It is an important component of Spelling Correction, Machine Translation, Information Retrieval, Text Generation, and other natural language systems.[19]. Morphemes are minimal units of morphology, e.g. build –> building

Lexical analysis is a vocabulary that includes its words and expressions. It depicts analyzing, identifying, and description of the structure of words. It includes dividing a text into paragraphs, sentences and words. Individual words are analyzed into their components, and non-word tokens such as punctuations are separated from the words.

ii. SYNTACTIC ANALYSIS

The words are commonly accepted as being the smallest units of syntax. The syntax refers to the principles and rules that govern the sentence structure of any individual language. Syntax focus on the proper ordering of words which can affect its meaning. This involves the analysis of the words in a sentence by following the grammatical structure of the sentence. The words are transformed into the structure to show how’s the word is related to each other.

iii. SEMANTIC ANALYSIS

Semantic Analysis is a structure created by the syntactic analyzer which assigns meanings. It shows how the words are associated with each other. It focuses only on the literal meaning of words, phrases, and sentences. This only abstracts the dictionary meaning or the real meaning from the given context.

It involves the extraction of context-independent aspects of a sentence's meaning, including the semantic roles of entities mentioned in the sentence, and quantification information, such as cardinality, iteration, and dependency. The culture of society has an impact on semantic analysis [19].

For example, "Colorless green idea." here; this would be rejected by the Symantec analysis as colorless green doesn't make any sense.

iv. DISCOURSE INTEGRATION

It means a sense of the context. The meaning of any single sentence which depends upon the previous or next sentence. It also considers the meaning of the following sentence. For example, the word "that" in the sentence "He wanted that" depends upon the prior discourse context. It deals with the properties of the text as a whole that convey meaning by making connections between component sentences.

v. **PRAGMATIC ANALYSIS**

Pragmatic Analysis deals with the overall communicative and social content and its effect on interpretation. It means abstracting or deriving the meaningful use of language in situations. The pragmatic analysis helps users to discover this intended effect by applying a set of rules that characterize cooperative dialogues.

It is the study of how linguistic properties and contextual factors interact in the interpretation of utterances, enabling hearers to bridge the gap between sentence meaning and speaker’s meaning. [19].

For example, "close the window?" should be interpreted as a request instead of an order.

2.3.2. APPROACHES OF NLP

NLP requires a set of rules to represent knowledge about the linguistic structures and depending on how rules are acquired; approaches to NLP can be rule-based or statistical. A rule-based approach where rules are written manually and a Statistical approach where rules are acquired from large size corpora.

i. Rule-Based Approach

It needed linguistic expertise and it is slower since it is a manually designed sequence of words, or part-of-speech, or another way of representing words in a sentence, and matches these sequences with the text and it does not need frequency information. Regular expressions and context-free grammars are examples of rule-based approaches to NLP.

ii. Statistical Approach

The statistical approach aims to perform statistical inference for the field of NLP. Statistical inference consists of taking some data generated following some unknown probability distribution and making inferences. It needed not much linguistic expertise required and it is based on frequency information. It uses large text corpora to allow rapid, robust, and accurate handling of the ambiguities in human languages.

iii. Hybrid approach

Both rule-based and statistical approaches use mathematical foundations and have their pros and cons. Thus, rule-based and statistical approaches are usually combined to benefit from their synergy effect. This gives rise to hybrid approaches.

2.4. APPLICATIONS OF NLP FOR SENTIMENT ANALYSIS

Natural language processing has a lot of applications,including Information retrieval (IR), Query expansion, Natural language search, Automatic summarization, Natural language generation, Text simplification, Text-proofing, Topic segmentation and recognition, Reference resolution, Relationship extraction, Sentiment analysis, Automated essay scoring, Natural language understanding, Discourse analysis. It is not an interest of this study to clarify those mentioned above but from those listed applications of NLP, this study tries to discuss what the sentiment analysis application means, its features, computing steps to perform with relation to social media analytic since our study concerns on this application area.

Sentiment analysis has been first introduced by Liu,[42]. It deals with identifying and classifying opinions or sentiments expressed in the source text. It has a broad application ranging from e-commerce, marketing, to politics and research. The science behind sentiment analysis is based on algorithms using natural language processing to categorize pieces of writing as positive, neutral, or negative. The algorithm is designed to identify positive and negative words, such as “fantastic”, “beautiful”, “disappointing”, “terrible”, etc.

NLP helps companies to analyze a large number of reviews on a product. It also allows their customers to give a review of the particular product. Sentiment analysis is a method for measuring opinions of individuals or groups, such as the intent of comments, the brand’s audience, or an individual opinion in communication. Based on the scoring of measuring mechanism, sentiment analysis monitors conversations and evaluates language and voice inflections to quantify attitudes, opinions, and emotions related to their topic like a business, product or service, political issues, etc.

Itis the automated process of understanding an opinion about a given subject from written or spoken language. It uses text mining and natural language processing procedures so that a computer can take in the expression of emotions. Moreover, it helps bring out the sentiment and emotional expressions from unstructured text and providing the best method to classify a given sentiment analysis.[41].

The sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes, and categories within a sentence or phrase. As we know Social media empowers different sectors to read what people are saying about them and to join the conversation. To be able to make good use of comments, likes, reviews, ratings, recommendations, and other forms of online expressions, sectors need to apply them to their thought, ideas, and products, identify new opportunities and manage their needs. So to deal with this social media analysis has to be done on the sentiment of what peoples saying in the sectors.

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc.; is positive, negative, or neutral.

2.5. HOW DOES SENTIMENT ANALYSIS WORK?

As these studies try to describe above sentiment analysis or opinion mining is extracting opinions, emotions, and sentiments in text. It allows us to track attitudes and feelings on the web. People write blog posts, comments, reviews, and tweets about all sorts of different topics. So to do this it operates on different fields of study like linguistics, NLP, artificial intelligence where it uses the information given by the NLP and uses a lot of algorithms to determine whether something is negative or positive, and machine learning.

So this opinion is quintuple, in which an object made up of 5 different things for the machine, **(O, f, s, h, t)**

Where **O** is the opinion (or sentiment) target, “**f”** is features of an object o, **s** is the sentiment about the target, “**h”** is the opinion holder, and “**t”** is the time when the opinion was expressed. And these five elements have to be identified by the machine.[5].

This sentiment or opinion analysis can be performed at three levels of investigation; those are at a document level, sentence level, and Entity and Aspect level.

Document-level sentiment analysis is to classify a document or whether a whole opinion document expresses a positive or negative sentiment [20]. It aims to classify an opinion document as expressing a positive or negative opinion (or sentiment), which are called sentiment polarities. The task is referred to as document-level analysis because it considers each document as a whole and does not study entities or aspects inside the document or determine sentiments expressed about them. It assumes that the opinion document d expresses opinions on a single entity e and contains opinions from a single opinion holder h. And it obtains the sentiment of a complete document or paragraph.

Sentence level sentiment analysis where these studies try to conduct is determined whether each sentence expressed a positive, negative, or neutral opinion and it obtains the sentiment of a single sentence.

And the third on aspect level sentiment analysis is since both the document level and the sentence level analyses do not discover what exactly people liked and did not like. The aspect level performs a finer-grained analysis.[5]. The aspect level was earlier called the feature level (feature-based opinion mining and summarization).

With relation to social media analytic, sentiment analysis means automated extraction of expressions of positive or negative attitudes from a given blog, posts, tweets, and comments that have received on the site. For example on a given Facebook post what someone’s comments indicate whether is it positive or negative.

2.6. APPLICATION OF SENTIMENT ANALYSIS

Sentiment analysis is being used for different applications and can be used for several others in the future. From those applications area the common one that is currently used and many researchers try to study are; Online Commerce, Voice of Customers (VOC) and Voice of Market (VOM), Brand Reputation Management (BRM), Recommendation Systems, Policy-Making. [30,53].

For sentiment analysis one can use either objective or subjective sentences. An objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs [5].

I. SUBJECTIVITY CLASSIFICATION

Subjective analysis or classification is one branch of NLP application and it is classifying a sentence as subjective or objective, known as subjectivity classification. The basic concept here is that the sentiment classification includes both subjective and objective information. Subjective information indicates the opinions of opinion holders, while objective texts show some objective facts. Sentiment classification is one main task of opinion mining. The approaches of sentiment classification can roughly fall into two basic categories. The methods in the first category rely on language resources. The language resources include sentiment lexicons and natural language corpus libraries. Researchers usually use some natural language processing techniques combined with language resources to improve the accuracy of the sentiment classification. The methods in the second category these studies try to employ are machine learning and lexicon-based to do sentiment classification.

Machine learning-based sentiment classification methods contain supervised and semi-supervised methods that need some training instances to learn to get the final sentiment classifiers[30].

II. OBJECTIVITY CLASSIFICATION

As we try to discuss on literature review part of this work textual information can be broadly categorized into two main types: subjective and objective or facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties. The concept of opinion is very broad. In this work, we only focus on opinion expressions that convey people's positive or negative sentiments specially on political discourse of Ethiopian social media users. Which means that, on the objective point of view we are not try to find the fact where someone's happness or sadness feeling source, we just try to deal with the subjective of those feelings.Much of the existing research on textual information processing has been focused on mining and retrieval of factual information, e.g., information retrieval, Web search, text classification, text clustering and many other text mining and natural languageprocessing tasks. Yet,opinions are so important that whenever we need to make a decision we want to hear others' opinions. This is not only true for individuals but also true for organizations.

2.7. CLASSIFICATIONS OF SENTIMENT ANALYSIS APPROACHES

There are different approaches used for sentiment analysis. Here under we discuss some of the widely used approaches.

2.7.1. MACHINE LEARNING

Machine Learning is a system that can learn from example through self-improvement and without being explicitly coded by the programmer. [5,6,20]. It combines data with statistical tools to predict an output. This output is then used by different concerned sectors to makes actionable insights. Machine learning has different tasks are to provide a recommendation, fraud detection, predictive maintenance, portfolio optimization, automatize task, and so on.

Machine learning uses two types of techniques [5]: supervised learning and unsupervised learning, While supervised learning, trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, finds hidden patterns or intrinsic structures in input data.

Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes. As the algorithms ingest training data, it is then possible to produce more precise models based on that data. A machine learning model is the output generated when you train your machine learning algorithm with data. After training, when you provide a model with an input, you will be given an output. For example, a predictive algorithm will create a predictive model. Then, when you provide the predictive model with data, you will receive a prediction based on the data that trained the model. [27].

ML is one branch of AI where, it is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages; it has lots of applications. [6].

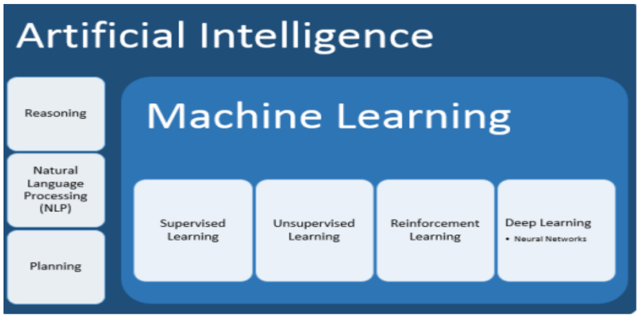


Figure 2 The overall category of AI that includes machine learning and natural language processing. [27]

To do this, Machine learning applications use different techniques. These techniques are required to improve the accuracy of predictive models. Depending on the nature of the business problem being addressed, there are different approaches based on the type and volume of the data. The machine learning approach involves text classification techniques. This approach treats the sentiment classification problem as a topic-based text classification problem. [5].

The features of a machine learning-based approach for sentiment classification are [5,6,20]:

- Term presence and their frequency: that includes Uni-grams or n-grams and their presence or frequency.
- Part of speech information: used for disambiguating sense which is used to guide feature selection.
- Negations: has the potential of reversing sentiments opinion words/phrases: that expresses positive or negative sentiments..

2.7.1.1. SUPERVISED LEARNING

Supervised learning is a process of learning algorithm from the training dataset. []. It is the learning of the model wherewith input variable (x) and an output variable (Y) and an algorithm to map the input to the output;That is, Y = f(X).

In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on patterns and associating the patterns to the unlabeled new data. Here in supervised learning, all data is labeled and the algorithms learn to predict the output from the input data.

The supervised machine learning algorithms are those algorithms that need external assistance. The input dataset is divided into train and test datasets. The training dataset has an output variable that needs to be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification. [24].

Input data or training data has a pre-determined label e.g. True/False, Positive/Negative, Spam/Not Spam, etc. A function or a classifier is built and trained to predict the label of test data. The classifier is properly tuned (parameter values are adjusted)to achieve a suitable level of accuracy.[28].

Supervised Learning can be divided into 2 categories i.e. Classification & Regression.[5,20]. The first one is Classification predicts the category the data belongs to. e.g.: Spam Detection, Churn Prediction, Sentiment Analysis, and Regression predict a numerical value based on previously observed data. E.g. House Price Prediction, Stock Price Prediction, Height-Weight Prediction. From those categories, since this study tries to perform sentiment analysis we will go through the first categories of supervised learning machine learning algorithms.

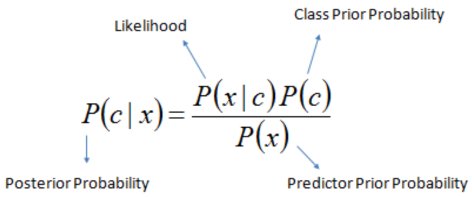
As we indicate above supervised learning categories; the classification one performs through the linear classifier and probabilistic classifier to determine the categorical class labels of new instances based on past observations. These two categories of supervised learning ML they have their branches of algorithm and from those algorithms, these study will discuss navies Bayes (NB) from the probabilistic classifier, which will be used for classification purpose in this study.

NAIVES BAYES (NB)

Navies Bayes supervised learning machine learning classifiers. It is a probabilistic classifier that makes classifications using the Maximum a Posteriori decision rule in a Bayesian theorem. Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection. [28].The posterior probability can be calculated by first, constructing a frequency table for each attribute against the target. Then, transforming the frequency tables to likelihood tables and finally uses the NaiveBayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of the prediction.

The Navies Bayes (NB) classifier is derived from Bayes rule which tells us how to flip the conditional reason about effects to causes. It uses the Bayes theorem but assumes that the instances are independent of each other which is an unrealistic assumption in practical world naïve Bayes classifier works well in complex real-world situations.[5,6,20].

The Naïve Bayes classifier algorithm can be trained very efficiently in supervised learning for example a political party which intends to promote a new policy the company can collect the historical data for its members, including current policies support, insight about the new one and the previous policy, and information on whether a member loves the proposed policy. Using Naïve Bayes classifier the company can predict how likely a member is to respond positively to a policy offering. With this information, the political party can gain insights from its member to gain in advance on other competitors.


Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Here in Navies Bayes text classification defined as [5,6]:

Given an Input: a document d and a fixed set of classes C = {c1, c2… cJ} the output will be: a predicted class *c* ∈*C*. Navies Bayes relies on a very simple representation of document using a bag of words assumption where it assumes position doesn't matter and conditional independence where it assumes the feature probabilities *P* (*xi*|*cj*) are independent given the class *c*.

For example: Suppose we are building a classifier that says whether a text is positive or Negative. Let us look at Navies Bayes's example for text classification. The training set consists of 10 Positive Reviews and 10 negative reviews and considered word counts are as follows: ሌብነትሰራነውይከበርምብለዋል።

Positive: ሌብነት =4, ሰራ =6, ይከበርም =5, ብለዋል =3

Negative: ሌብነት =5, ሰራ =20, ይከበርም =5, ብለዋል =4

Given the test set as "ሌብነትሰራነውይከበርምብለዋል።" Let us find the sentiment for the given test set

Given the training, set consists of the following information Positive reviews =10 and Negative reviews=10. Total no of Reviews=positive reviews+ negative reviews=20

The prior probability for the positive reviews is P (positive) =10/20=0.5

The prior probability for the negative reviews is P (negative) =10/20=0.5

Conditional probability is the probability that a random variable will take on a particular value given that the outcome for another random variable is known.

The conditional probability for the word 'ሌብነት' in a positive review is P(ሌብነት/positive) =4/10 =0.4, for the word 'ሰራ' in the positive review is P(ሰራ/positive)=6/10=0.6, for the word 'ይከበርም' in the positive review is P(ይከበርም/positive) =5/10 =0.5, and for the word 'ብለዋል' in a positive review is P(ብለዋል/positive)=3/10 =0.3

The conditional probability for the word ‘ሌብነት’ in the negative review is P(ሌብነት/positive)=5/10= 0.5, for the word ‘ሰራ’ in negative reviews P(ሰራ/positive) =20/10 =2, for the word ‘ይከበርም’ in the negative review is P(ይከበርም/positive)=5/10 =0.5, and for the word ‘ብለዋል’ in the negative review is P(ብለዋል/positive)=4/10 =0.4

The posterior probability is the product of prior probability and conditional probabilities.

Since, Posterior probability= prior probability *conditional probability

The posterior probability for the positive review is P(positive)=**0.4*0.6*0.5*0.3 = 0.036** and for the negative review is P(negative)=0.5*2*0.5*0.4= 0.2

The posterior probability for the negative reviews is greater than the posterior probability of the positive review P(negative)>P(positive) or 0.2 > 0.036. The given test set “ሌብነትሰራነውይከበርምብለዋል።” is predicted by Naïve Bayes as a Negative Sentiment. From the above example, we consider the preprocessing step is as done like stop word “ነው” and “።” are removed.

2.7.1.2. UNSUPERVISED LEARNING

Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses. The main purpose of unsupervised learning is to group data having similar characteristics into different clustering. Clustering is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include gene sequence analysis, market research, and object recognition. Common algorithms for performing clustering include k-means and k-means, hierarchical clustering, Gaussian mixture models, hidden Markov models, self-organizing maps, fuzzy c-means clustering, and subtractive clustering.

The unsupervised learning algorithm learns a few features from the data. When new data is introduced, it uses the previously learned features to recognize the class of the data. It is mainly used for clustering and feature reduction.[24].

2.7.2. LEXICON-BASED APPROACH

Lexicon Based is one branch of sentiment classification techniques and it works on an assumption that the collective polarity of a document or sentence is the sum of polarities of the individual words or phrases. These approaches for sentiment classification are based on the insight that the polarity of a piece of text can be obtained on the ground of the polarity of the words which compose it. There are three techniques to construct a sentiment lexicon: manual construction, corpus-based methods, and dictionary-based methods.[5,6,20,21,51].

I. MANUAL CONSTRUCTION

It is the process of creating sentiment words and phrases constructing manually. This phrase and sentiment are basic linguistic units to express feelings and to construct such terms manually are time-consuming and difficult.

II. CORPUS-BASED METHOD

The Corpus-based approach helps to solve the problem of finding opinion words with context-specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus.[34].

Corpus approach uses the statistical co-occurrence information in a large collection of documents and is based on the assumption that the sentiment words that have the same polarity occur together in the corpus.[25].

III. DICTIONARY BASED

The dictionary-based approach utilizes synonyms and semantic relations to determine the positive and negative polarity of words. These methods produce sentiment lexicon using a dictionary.

In these methods, at first, a set of seed words with known positive and negative orientation is collected manually and then using bootstrapping algorithms to find their synonyms and antonyms in the dictionary, the newly found words in each iteration are added to the positive and negative lists until no more new words can be found.[25].

2.7.3. HYBRID APPROACH

This approaches is the combination of both the machine learning and the lexicon-based approaches has the potential to improve the sentiment classification performance. There are some advantages and limitations in using these different approaches depending on the purpose of the analysis.[18].

It employs the lexicon-based approach for sentiment scoring followed by training a classifier to assign polarity to the entities in the newly find reviews. And most of the currently available research works have dealt with the problem of sentiment and subjectivity classification using this approach.

2.8. RELATED WORKS AND GAP ANALYSIS

A lot of work has been done in the field of sentiment analysis. Different techniques are used to classify the text according to the polarity of the entity that the study concerned. Most of these techniques can be classified under different categories as this study try to discuss above in the classification of sentiment analysis. Therefore, in this part different literature or works that have been done about social media analytics and politics and related works has been done on sentiment analysis will be discussed by dividing into two parts. Firstly we look at different works related to this work with relation to social media and politics and secondly works about sentiment analysis and subjectivity classification with a variety of approaches.

2.8.1. RELATED WORKS ON SOCIAL MEDIA AND POLITICS

In this part of related work, the study presents some works done so far based on social media and politics. Among them, we have chosen the most relevant ones which are related to our works that are done in different languages.

The research work presented on “Young adults’ political engagement using Facebook” [40] explores how young adults understand and experience political engagement on Facebook. And to explore this valuable active research area they used to work on using in-depth interviews on 10 different Facebook users. Their findings suggest that these groups of political participants choosing political engagement through a non-traditional way which is social network websites. And finally, they try to indicate an insight on how political participation on Facebook is greatly affected by the social nature of the site and gives a direction for any study of politics and Facebook must consider how the nature of social networking sites affects interaction in relevant fields.

Different news media handle the same news story in different ways. Serious newspapers and television news programs more often used the responsibility and conflict frames in the presentation of news. The research work presented in “The Role of Social Media in Political Campaigns” by [32] elaborates the news operator connection on social media posts and examine its relationship with public sentiment, by studying how emotion is produced and consumed via news exposure on social media. The instrument used was based on a dictionary-based approach, which consists of a simple word count of the frequency of keywords in a text from a predefined dictionary. Then they analyze a large amount of news feeds from news media’s social media outlet (i.e. Facebook), and they try to analyze the linguistic patterns of those news feeds and corresponding comments, to identify sentimental elements embedded in news reporting and commenting that are related to particular political campaigns.

The research work presented in “Politics and Young Adults: The Effects of Facebook on Candidate Evaluation” by [37] tries to explore the relationship between young voter's social media use to evaluate political candidates. To do this they try to make work based on a larger experimental lab study on how social media and traditional media (meaning non-social media material) might interact when subjects. To conduct this they use a semi-structured interview with two politicians participating in the election and users of Facebook. They contribute to the literature on how voters use information communication technologies (ICTs) to inform the vote decision process.

The other related work this study tries to look is that the work is done by [38]; “The role of online social networking in the 2008 democratic presidential primary campaigns”, which try to examine whether social networking sites enhanced political engagement, expanded the campaigns’ reach to traditionally ignored audience groups, and increased the interactivity between the campaigns and voters. They try to conduct searching for resources in two parts as campaign staff and the online Facebook user through interviews and online survey questions posted online. From their work, they get an insight where some meet their first hypotheses on the behalf of campaign staff on SNN their candidate’s profile has led to a high level of involvement in terms of discussion and support and SNN to reach out to young people or to demographic that is connected online and effective as a communication tool they able to post contents to get a lot of feedback. But in contrast on the behalf of online users, their level of political activity or awareness had not changed after browsing or supporting candidates on Facebook. Even if the way they followed to achieve their goal is good but for the second hypotheses which are given by the online user, it might not fully reasonable since the way they perceive response of online user is not to get an awareness instead to give only support and to give assurance for the availability of the campaigns.

2.8.2. RELATED WORKS ON SENTIMENT AND SUBJECTIVITY CLASSIFICATION

2.8.2.1. SUBJECTIVITY ANALYSIS

The problem of determining whether a sentence is subjective or objective is called subjectivity classification[26]. Alternatively, the detection of subjectivity and polarity can be jointly addressed by treating the problem as a 3-class classification problem with classes: neutral, positive, and negative. Different works are done on subjectivity classification and those works have been done on a single language and multilingual through by creating a dictionary of a translator. But those works that are closely related to this research work that helps to achieve its target will present.

The research work presented in [26] tries to works on presenting subjectivity classifiers from a standardly prepared dataset for subjectivity classification. They performed subjectivity classification of sentences using basic features such as the presence of a pronoun, an adjective, a modal, etc. in the sentence through world Press articles using unannotated data. In addition to this, they advance the state of the art in objective sentence classification by learning extraction patterns associated with objectivity and creating objective classifiers that achieve substantially higher recall. They used high precision rule-based classifiers for generating initial training data and then used semi-supervised learning to iteratively learn subjectivity patterns and augment the training data. The rule-based subjective classifier classifies a sentence as subjective if it contains two or more strong subjective clues (otherwise, it does not label the sentence).In contrast, the rule-based objective classifier looks for the absence of clues: it classifies a sentence as objective if there are no strong subjective clues in the current sentence, there is at most one strong subjective clue in the previous and next sentence combined, and at most 2 weak subjective clues in the current, previous, and next sentence combined (otherwise, it does not label the sentence).

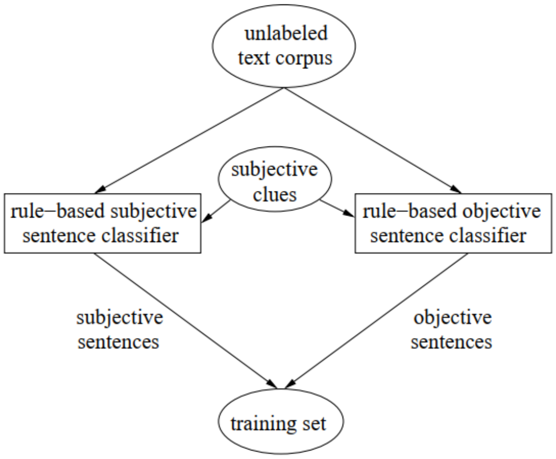


Figure 3 Initial training data creation (source from their paper)

The second related work this research work tries to look at is “Recognizing Subjectivity: A Case Study of Manual Tagging” by [35], they work on subjectivity recognition through manual tagging on a sentence- level categorization in which tagging instructions are developed and used by four judges to classify clauses from the Wall Street Journal as either subjective or objective. By analyzing those arguments given by judges it make a final classification.

The research work presented in [33], try to discussed and evaluated methods to develop subjectivity analysis tools for selected languages by applying machine translation on the available subjectivity analysis tools and resources for the English language. They investigate methods to automatically generate resources for subjectivity analysis for a new target language by leveraging on the resources and tools available for English. Specifically, through experiments with a cross-lingual projection of subjectivity, they follow two classification techniques to generate and evaluate subjectivity and sentiment analysis tools rely on manually or semi-automatically constructed lexicons in which corpus-based and lexicon-based. On the lexicon approach through by translate subjective lexicon in which by translating an existing English lexicon using a bilingual dictionary and rule-based subjectivity classifier using a Subjectivity Lexicon and the other one is through the corpus-based approach. This approach builds a subjectivity-annotated corpus for the target language through projection and then trains a statistical classifier on the resulting corpus. In which, generate a subjectivity-annotated corpus in a target language by projecting annotations from an automatically annotated English corpus. Finally, they try to validate their work through measuring the extent to which subjectivity is preserved across languages in each of the two resources and validated the two automatically generated subjectivity resources by using them to build a tool for subjectivity analysis in the target language. They try to indicate that a similar way can be used to derive tools for subjectivity analysis in other languages. But the thing they miss is that the projection of annotations across parallel texts can be successfully used to build a corpus annotated for subjectivity in the target language. However, parallel texts are not always available for a given language pair. Therefore, instead of relying on manually translated parallel corpora, we think that machine translation to produce a corpus in the new language will be a good approach.

2.8.2.2. SENTIMENT ANALYSIS

There is a large amount of work on Sentiment Analysis, and good comprehensive overviews are already available[5],[20]. So this study will review the most representative and closest to this work. Since sentiment analysis is a recent research area branch of NLP and linguistic and a lot of work has been done, but instead of dealing with the actual problem that the world faces most of them are based on the creating of the method or model that used for the classifier. This study proposed a two-step classification method. It first classified sentences as its subjectivity related to the post they belong and then classifies the subjective sentence as positive, negative, or neutral. To determine whether a sentence expresses a positive or negative sentiment, two main approaches are commonly used: the lexicon-based approach and the machine learning-based approach.

The study first trains using a machine learning-based approach where typically trains sentiment classifiers using features such as unigrams or bigrams; by applying different techniques using some form of supervised learning, such as Naive Bayes, and Support Vector Machines. These methods need manual labeling of training examples for each application domain. There are also some approaches that different researchers utilize, both the opinion words/lexicon and the learning approach.

The research work presented on “Sentiment mining model for opinionated Amharic texts” by [16] tries to classify a given opinionated document or text into predefined classes since polarity classification is concerned with categorizing a given opinionated document into predefined categories based on the weights obtained from the weight assign and polarity propagation process. By assign an initial value to the sentiment terms and propagates the initial polarity values, lexica of properly tagged Amharic sentiment terms are used to detect sentiment terms. To accomplish the task they follow the approaches of building Amharic sentiment lexicon.

The work done by [26] used a subjectivity lexicon to identify training data for supervised learning for subjectivity classification. Then the lexicon-based approach applied to determine the sentiment or polarity of opinion via some function of opinion words in the sentence [29],[31] is other works that use a lexicon-based approach for sentiment classification.

2.9. SUMMARY OF REVIEW WORK

The reason why it is essential to make sentiment analysis on political discourse through social media analytic is that it is beyond sentiment analysis. Which means that it is not only classifying a document or a sentence positive or negative instead it plays a significant role in discovering a new insight within a domain or tracking point of view that are necessary for analyzing opinions through the political domain in blogs, newspaper, and articles where a third person narrates his/her views.

The study discusses in detail about social media analytic, applications, components, and branches of NLP, various approaches to Sentiment Analysis and subjectivity classification, mainly Machine Learning and lexicon approach. The study tries to look at the application of ML techniques like Naïve Bayes and Support Vector Machines and lexicon approaches used for SA and provides a detailed view of these techniques.

We also look at specific topics on sentiment and subjectivity classification that help us to gain knowledge for this study and how to make a sentiment analysis work, their methods and approaches, potential challenges of social media analysis, Sentiment Analysis, and subjectivity classification.

Here through this work what we have seen and tries to recommend is that most of the research works focus not only in developing classifiers for a single language or multilingual but also they should have to deal with currently available world problem which means they should have to work on cooperatively with social science, health sector, political science and soon research problem areas.

CHAPTER THREE

METHODOLOGY

3.1. OVERVIEW

In this chapter, the study discusses the research design, the research method used, area of a study conducted, the period of the data collected for the study and why, the sample of the population, techniques used, the instrument for data collection, how the corpus was prepared, method of data analysis.

3.2. RESEARCH QUESTIONS

In this research we discuss the following questions and try to give answer to each one of them.

RQ1. Are the sentiments associated with the comments an indicator of the correlation between the content of post and real world events in politics?

RQ2. How does the sentiment of a particular posts trend over a commenters?

RQ3. How does the sentiment embedded in user comments relate to the topic of the post and to which percent sentiment polarity it has?

RQ4. How does political polarization make people vulnerable to disinformation or unrelated comments or in turn how does the increase prevalence of dis-information lead to greater political polarization?

To answer this research questions, since, as we try to describe above social media is the new knowledge chain for all age groups and become a platform to express sentiments in the form of opinions and reviews on almost everything- movies, brands, product, social, political and economic activities, here I ask myself "The next big question here is; how can I get data for sentiment analysis and how can I actually analyze the sentiment data?"

3.3. RESEARCH METHODS

Broadly, research done based on methods is classified into two, qualitative research and quantitative research.[]. The main focus of qualitative research is to understand, explain, explore, discover, and clarify situations, feelings, perceptions, attitudes, values, beliefs, and experiences of a group of people. The study designs mainly entail the selection of sentence or comment polarity and subjectivity from the information gain from the posts, through techniques available, that was explored and gathered. The parameters of the scope of a study, and information gathering methods and processes, are often flexible and evolving; On the other hand, in quantitative research, the measurement and classification requirements of the information that is gathered demand that study designs are more structured, rigid, fixed and predetermined in their use to ensure accuracy in measurement and classification. [43].

Since this study was conducted to investgate a relationship between the collected data (which is Amharic comments given on political discourse on Ethiopian political, social media pages) and current social media use on the political discourse on the case of Ethiopia, the finally observations is established based on mathematical calculations. It seems to be a qualitative research design but it also has statistical conclusions to collect actionable insights which means that quantitative research was conducted. Therefore, the research conducted applied a combination of both methods of research design and followed mixed research methods. This provides a holistic approach combining and analyzing the statistical data with deeper contextualized insights and also enables Triangulation, or verification, of the data from two or more sources.

3.4. RESEARCH DESIGN

Research design is a framework of methods and techniques chosen to combine various components of research in a reasonably logical manner to efficiently handle the research problem. It provides insights about "how" to conduct research using a particular methodology.

Therefore, the general research design or the framework, method, or technique this study followed is that first data collection, classification of subjectivity followed by an analysis of the polarity of the comment, and finally the analysis of the data generated by subjectivity and sentiment analysis work.

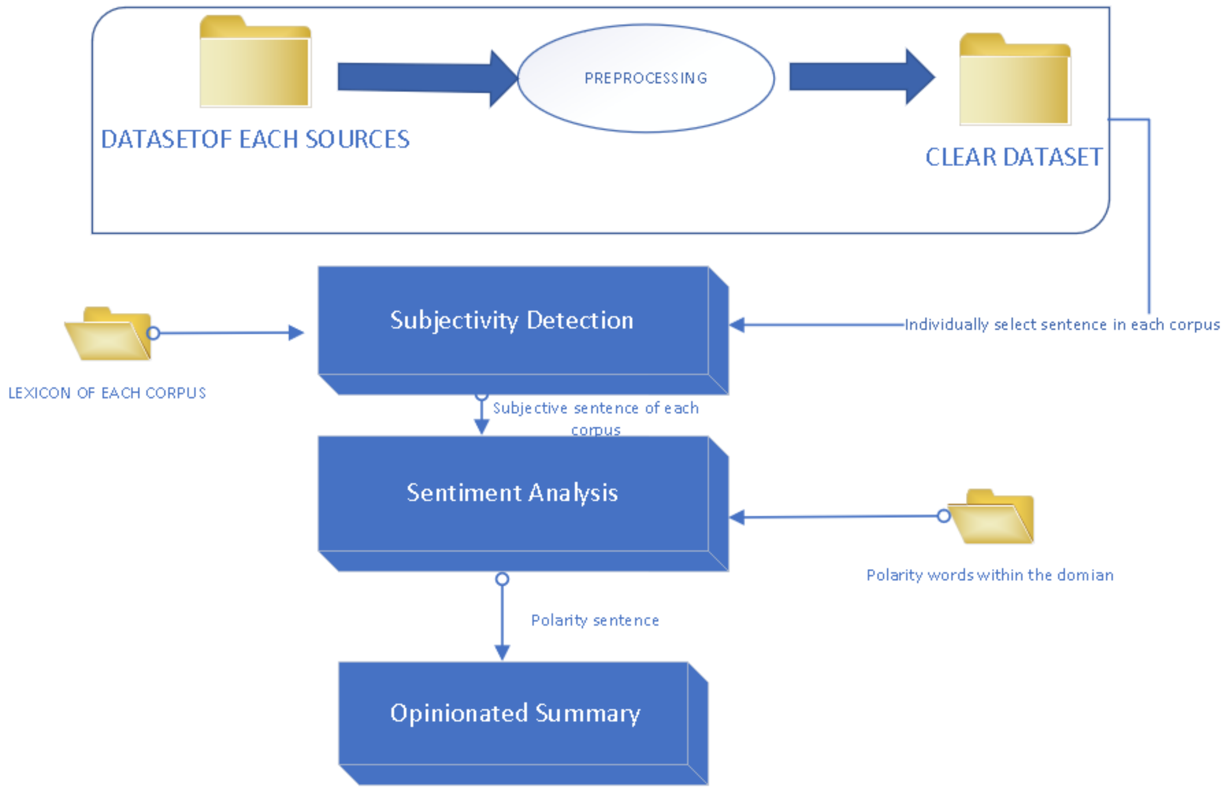


Figure 4 General approach/method conducted

3.4.1. METHODS OF DATA COLLECTION

This study followed primary data collection since the data was not previously collected and published for a specific purpose. They are critically analyzed to find the trends of our political discourse participation through social media, trends of comments with the post-it belongs.

SA works are mostly related to sentiment classification, as the polarity is a natural abstraction of collective opinion. Most studies classify comments or documents into two categories: positive versus negative. This is also referred to as a binary polarity classification. [52]. The study tries to follow sentiment analysis through a binary classification. This means that as early work describes, Sentiment analysis consists of two kinds of classifications: the first one is called the binary classification where opinion lies on one of two categories (e.g., positive or negative), while the second one is called multiclass classification where opinion lies on one of the multiple categories (e.g., strongly agree, agree, fair, disagree and strongly disagree) and subjectivity classification where comments or sentences are analyzed to identify the document that holds an opinion.

About tools and methods used to identify and collect data, different approaches have been applied to this area of research. It may be a ready-made dataset, manually or automatically fetching data. But manually collecting the data is not realistic and definitely, the most time consuming and inefficient. Therefore, to make it more practical the first task of the study was identifying the source of the data and collecting the data through different tools that have been developed and readymade tools; those are Facebook API, Python code, and websites to export comments. We can easily scrap any public Facebook Page or Group posts and comments to Excel spreadsheets using the Facebook scraper tool (ready-made Python Scripts). The thing we have to do was that first, we should have to register to Facebook account then register to Facebook Graph API Access Token to access the contents of the public pages. Finally, through the Graph API access token, we scrapped Facebook page posts with URL queries and we download Facebook page posts and comments to Excel. The first part of the data was collected on public and individuals news pages that used social media as an alternative tool to reach their followers and people of Ethiopian. Since the third part of media categories is religious it does not belongs to our political sentiment analysis concept and it automatically discarded from our work.

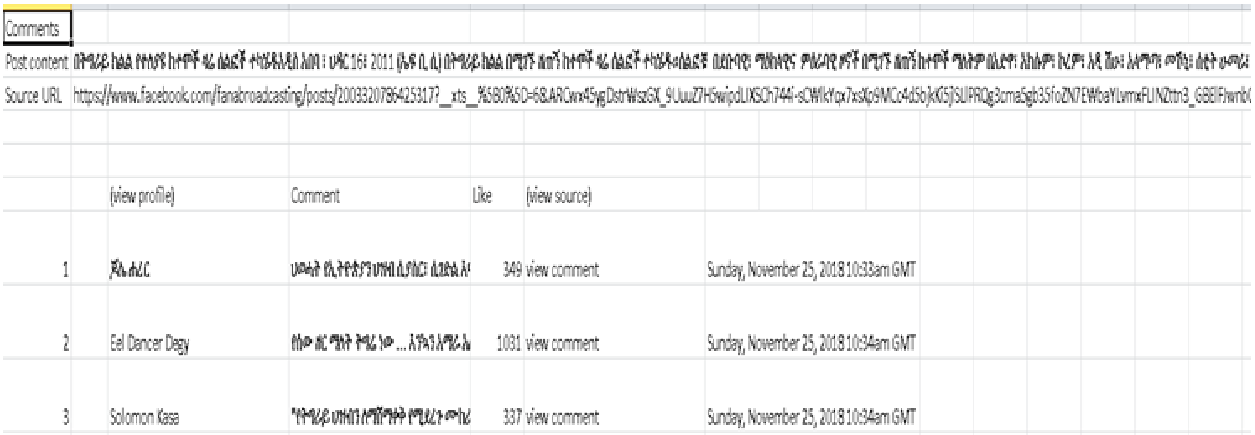
And from the governmental or public TV and radio which use social media as an alternative tool, we selected EBCZENA because the reality that almost all of the federal and regional governmental media try to release the same contents of news especially political news and the news language of transmission is Amharic. And finally, from the privately-owned company of media we select FANABROADCASTING AND ESAT, the reason that we only select from 26 private companies is that they have more followers than the rest of the members and secondly they use the Amharic language as a medium of transmission. Most of the other categories are entertainment groups as we can see the below table.

The second part of the data was collected from the most influential politician and activists that have a lot of followers regardless of their political views. But here we try to balance by taking the most influential people on three actively participant ethnic groups of the country, YONI MAGNA, JAWA MOHAMMED, DANIEL BERHANIE from Amhara, Oromia, and Tigray respectively. And additionally their follower and public graph they have in Ethiopian politics'.

Media	**Stations which have social media tool(FB)**	**Category of General News and uses FB as an alternative tool**	**Language of transmissions mostly**
Government-Owned Television Stations	ETV News , ETV Entertainment ,ETV Language ,Addis TV ,Amhara TV ,Debub TV ,Harar TV ,OBN ,Tigray TV - 11	General, news,	Amharic – 7, Afaan Oromo – 1, Tigrigna – 1, Harari – 1, Somali - 1
Privately-owned TV stations	Kana TV ,LTV Ethiopia ,EBS TV ,ENN TV , JTV Ethiopia , Nahoo TV , OBS TV , Walta TV , Fana TV , Aleph TV, Afrihealth TV , ARTS , DW TV , Bisrat TV , Asham TV , AHADU TV , Asrat HD , Arki TV , ESAT , OMN , Balageru TV , ABN TV , Ethio lijoch , TV 9 Ethio , SMN-ETH, TMH TV - 26	General Entertainment, Culture, Health News, General News, Children's' channel	Amharic - 21 Afaan Oromo – 2 Tigrigna – 2 Simian – 1
Religious company	EOTC TV, Christ Army TV, Bethel TV, Vision TV, Holy TV, CJ TV, Jesus TV, El Shaddai Television Network (ETN HD) Presence TV, Elhori TV, Elohi TV, MO'A TV HD, Arara TV, GMM TV Ethiopia, Africa TV 1, Zawya TV,	Ethiopian Orthodox, Protestant, Islam	Amharic – 28 Afaan Oromo – 5

	Nuuralhudaa TV, Nesiha TV HD, Hamilton TV, As-Sunnah TV, Anointing TV, Aleph TV, Christ Mission, Holy Sprit TV, Marcil TV, Evangelical TV HD, Glory TV ETH, Fover TV, WW TV, 7 Spirit TV, Rehobot TV HD, As-Sunnah TV - 33		

Table 2 Data collection source



Comments

Post content: በትግራይ ክልል የተለያዩ ከተሞች ዛሬ ሰልፎች ተካሄዱ [illegible] ህዳር 16፣ 2011 (ኤፍ ቢ ሲ) በትግራይ ክልል በሚገኙ ዞኖች ከተሞች ዛሬ ሰልፎች ተካሄዱ [illegible]

Source URL: https://www.facebook.com/fanabroadcasting/posts/2003320786425317?__xts__%5B0%5D=68.ARCwx45ygDstrWszGX_9UuuZ7H5wipdLIXSCh744i-sCWlkYqx7xsXp9MCc4d5bjkKG5jlSLlPRQg3cma5gb35foZN7EWbaYLvmxFLlNZttn3_GBElFJwnb(

	(view profile)	Comment	Like	(view source)	
1	[illegible]	[illegible]	349	view comment	Sunday, November 25, 2018 10:33am GMT
2	Eel Dancer Degy	[illegible] ... [illegible]	1031	view comment	Sunday, November 25, 2018 10:34am GMT
3	Solomon Kasa	[illegible]	337	view comment	Sunday, November 25, 2018 10:34am GMT

Figure 5 Sample data collected

POSTS	Unprocessed comment collected in all selected data source and their posts
30	15000

Table 3 size collected data

3.4.2. SAMPLE SIZE

In the work of [44], they were interested in the sentiment of political communication of Austrian parties and media and assembled a corpus of party press releases, minutes of parliamentary debates and media reports on election campaigns from the years 1995–2013 where the texts are available in a machine-readable format. The corpus initially consists of about 470,000 sentences. Pre-filtering with seed words cuts its size to about 215,000 sentences with negative sentiment. From that corpus, they randomly select 13,000 sentences for crowd coding. Pre-filtering with seed words is not required, but it reduces the coding costs.

In the case of our work, the data was collected in the time of the current Ethiopian government political transition period which is form June first, 2018 up to Nov 2018. The reason why the study selected this period is that firstly it was a large political movement that happened in the country, secondly, many social media participants provide their opinion freely without any frustration of the government and the ruling party. The data was collected specifically on domain-based and manually by selecting political news hotly spreads throughout all news agencies within the country and which has been a lot of comments given on.

The data was collected about 30 posts with each of them have 500 comments from that selected giant public and private social media news but in the period of data collection, the study excluded the post that has fewer comments and non-political posts from those data sources. The study selected those posts that have political news and content which has been a hot issue on Facebook and different social media with more than 500 comments or participants, and those are power full in terms of spreading on different news agency pages or websites.

In the works of [52], he used different sampling approaches; Blogosphere or weblog with 23 politics posts, Twitter Collection, and Movie Review Dataset. For example, in his work for the third sampling approach he used the Cornell movie-review dataset, which was created by Pang used as a source of the data for but only the Sentiment polarity dataset is used. Where The dataset includes 1000 positive and 1000 negative review articles labeled at the document level.

3.4.3. PREPROCESSING AND PREPARATION OF CORPUS

I. CORPUS

For the subjectivity detection and sentiment analysis part, the study followed not preparing one corpus for all of the analysis instead it prepared the dataset for each post with belonging comments for subjectivity detection then from those selected subjectivity sentences the study made sentiment or polarity analysis on those comments or sentences.

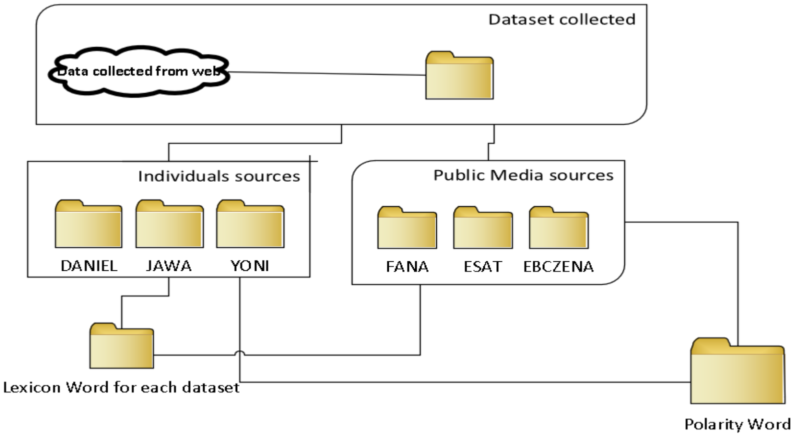


Figure 6 Corpus preparation step

II. TEXT PREPARATION

Text preparation involves cleaning the extracted data before sentiment analysis and subjectivity detection is performed. It involves in identifying and eliminating non-textual content from the textual dataset, and any information that can have a privacy issue like source URL, commenter’s name, commenter’s location, comment date, numbers of likes, comments that is not relevant to the area of study, comments given on different language like English, removing numbers, stripping whitespace, removing punctuation and stop words, was removed from the textual dataset.

III. TOKENIZATION

Each post or comment is stored in a string. To enable further processing, this string must be split into individual words. Tokenization is the process that splits a string into one or more words, phrases, symbols, or other meaningful elements called tokens. Larger chunks of text can be tokenized into sentences, sentences can be tokenized into words, etc. Tokenization is also referred to as text segmentation or lexical analysis. Sometimes segmentation is used to refer to the breakdown of a large chunk of text into pieces larger than words (e.g. paragraphs or sentences), while tokenization is reserved for the breakdown process which results exclusively in words.

Since sentence/comments are used to express the opinion of social media user and they need have to correctly structure in the form that the commenter need of expressing his/her idea of what the post content has. But this is not realistic and they don’t care about the sentence structure instead their only concern will be the case they commented on. For example, they don’t concern about sentence ending or full stops, commas and soon and after preprocessing this error we also need to break down sentences to get meaning full message from the comments given.

In our case, as we try to analyze and classify each comment given against the post we need to clear understanding of the comment. And each comment may not be only one sentence they maybe three or more, even at text paragraphs level. So to have a clear sentence we need a tokenizer(sentence) to breakdown a text paragraph into sentences. For the case of detecting polarity also we need word tokenizer to get single contents of the word. For example, the comment given below is the one taken from the data from our corpora.

```
from nltk.tokenize import sent_tokenize

comments = "የደቡብ ክልል ፖሊስ ኮሚሽን በሰኔ ወር በሃዋሳ ተከስቶ በነበረው የዘር ማጥፋት ወንጀል ተሳታፊ የነበሩ ተጠርጣሪ ወንጀለኞችን
ጉዳያቸው በዕርቅ ስላለቀ ለፍ/ቤት ማቅረብ አልችልም ስል ምላሽ ሰጠ \
 የፌደራል ከፍተኛ ፍርድ ቤት 13ኛ ተዘዋዋሪ ወንጀል ችሎት  ጀምሮ በሀዋሳ ከተማ እያስቻለ እንደሚገኝ ይታወቃል. ይህ የወንጀል ችሎት በዋናነት
ከዚህ ቀደም በሰኔ ወር በሀዋሳ ተከስቶ የነበረውን \
 የዘር ማጥፋት ወንጀል የተከሰሱ ግለሰቦችን ክስ ይመለከታል. ክስ የተመሰረተባቸው ከፍተኛ የሥራ ሀላፊዎችን ጨምሮ ሌሎች ተጠርጣሪዎችን
እንዲያቀርብ የተጠየቀው የክልሉ ፖሊስ ኮሚሽን ሀላፊነቱን \
  አለመወጣቱን ከሥፍራው ለወላይታ ሚዲያ ኔትዎርክ ተገልጿል. የክልሉ ፓሊስ ኮሚሽን ከዚህ ቀደም ክስ ቀርቦባቸው ያልተያዙና በዋስ ወጥተው
የጠፉ ተከሳሾችን ያላቀረበበት ምክንያት አስገራሚ እንደነበር \
 ምንጮቻችን ገልጸዋል. ለፍ/ቤቱ ተጠርጣሪ ወንጀለኞች ያለቀረበበት ምክንያት ተጠይቆ ምላሽ የሰጠው ፖሊስ ጉዳያቸው በዕርቅ ሂደት ላይ በመሆኑ
ሌላ ረጅም ተለዋጭ ቀጠሮ እንዲሰጥ የሚጠይቅ ነበር. \
 የክልሉ ፓሊስ ኮሚሽን ክትትልና አቀራረብ ኃላፊ ኢ/ር አብርሃም ይሄንን ተቀባይነት የሌለው ጉዳዩ በዕርቅ እያለቀ ነው የምል አይነት እንድምታ ያለው
መልስ ለፍ/ቤቱ ያቀረቡ ቢሆንም የፍ/ቤቱ ዳኞች \
 ጉዳዩን ውድቅ ማድረጋቸውን በሥፍራው ከሚገኙ ምንጮች አረጋግጠናል. የክልሉ ፖሊስ ለፍ/ቤቱ የሰጠው በዕርቅ ሂደት ላይ ነው ያለው ምላሽ
በክልሉ በሚገኙ ከፍተኛ ባለሥልጣናት የተመከረበትና \
 የተደገፈ መሆኑን ለፍ/ቤቱ ጨምሮ ቢገልፅም ፍ/ቤቱ ግን ይህ ከባድ የዘር ማጥፋት ወንጀል በመሆኑ በዕርቅ ሂደት ማለቅ አይችልም፤ ይህን ትዕዛዝ
የሰጡ ግለሰቦችም ሀላፊነት የጎደላቸው ናቸው ስል ውድቅ አድርጉዋል. \
 በመሆኑም በጉዳዩ ላይ ብይን ለመስጠት በ22/03/2011 ተለዋጭ ቀጠሮ ሰጥቷል. የክልሉ ፖሊስም ሆነ ትዕዛዝ የሰጡ የክልሉ ከፍተኛ ባለሥልጣናት
የያዙት ስልት ፍትህ ሂደቱን የማዘግየትና ጉዳዩ ውሳኔ \
 ሳያገኝ እንዲቀር የማድረግ እንደሆነ የወላይታ ሚዲያ ኔትዎርክ መረጃ ይገልጻል. በተጨማሪም በዚሁ ወንጀል ተከሰው ማረሚያ ቤት ከነበሩ ተከሳሾች
አንዱ ሆን ተብሎ እንዲያመልጥ ተደርጎ በስህተት ከማረሚያ \
 ወጥቷል የሚል ምላሽ ለችሎቱ በክልሉ ፖሊስ ኮሚሽን እንደቀረበ ለማወቅ ተችሏል. ጉዳዩ ምንም አይነት ትኩረት እንዳልተሰጠው የሚያሳየው
```

```
በቀድሞ ማረሚያ ቤት ለ6 ሰዎች ሞት ቀጥተኛ ተጠያቂ የነበሩ \
 ዮሐንስ ዮንቆራ የተሻለ ሹመት ክልሉ ላይ ተሰጥቷዋቸው እየሰሩ መሆናቸው ግርምት እንደፈጠረባቸው ያነጋገርናቸው ሰዎች ገልጸዋል. ይህ በእንዲህ
እንዳለ መቅረብ ከነበረባቸው ተከሳሾች 25ቱ ያልቀረቡ \
 ሲሆን ቀጣይ በተያዘው ቀጠሮ በ21/03 2011 እና 24/03/2011 የክልሉ ፖሊስ ተከሳሾችን እንዲያቀርብ ትዕዛዝ ተላልፏል. በመሰረቱ ከዚህ በፊት
የተፈጸመው እርቅ ከፍትህ ሂደቱ ጋር እንደማይገናኝ \
 የተገለፀ ቢሆንም በዕርቁ ምክንያት ወንጀለኞችን ማቅረብ አለመቻል ሂደቱን ማጠልሸት እንደሆነ ያነጋገርናቸው አስተያየት ሰጪዎች ገልጸዋል::
አሁንም በዕርቅ ስም ሆን ተብሎ የፍትህ ሂደቱን የማጨናገፍ ስራ \
 በአስቸኳይ መቆም ይገባዋል."

tokenized_comment=sent_tokenize(comments)

print(tokenized_comment)
```

OUTPUT:

```
['የደቡብ ክልል ፖሊስ ኮሚሽን በሰኔ ወር በሃዋሳ ተከስቶ በነበረው የዘር ማጥፋት ወንጀል ተሳታፊ የነበሩ ተጠርጣሪ ወንጀለኞችን ጉዳያቸው በዕርቅ
ስላለቀ ለፍ/ቤት ማቅረብ አልችልም ስል ምላሽ ሰጠ  የፌደራል ከፍተኛ ፍርድ ቤት 13ኛ ተዘዋዋሪ ወንጀል ችሎት  ጀምሮ በሀዋሳ ከተማ እያስቻለ
እንደሚገኝ ይታወቃል.', 'ይህ የወንጀል ችሎት በዋናነት ከዚህ ቀደም በሰኔ ወር በሀዋሳ ተከስቶ የነበረውን  የዘር ማጥፋት ወንጀል የተከሰሱ ግለሰቦችን
ክስ ይመለከታል.', 'ክስ የተመሰረተባቸው ከፍተኛ የሥራ ሀላፊዎችን ጨምሮ ሌሎች ተጠርጣሪዎችን እንዲያቀርብ የተጠየቀው የክልሉ ፖሊስ ኮሚሽን
ሀላፊነቱን   አለመወጣቱን ከሥፍራው ለወላይታ ሚዲያ ኔትዎርክ ተገልጿል.', 'የክልሉ ፓሊስ ኮሚሽን ከዚህ ቀደም ክስ ቀርቦባቸው ያልተያዙና በዋስ
ወጥተው የጠፉ ተከሳሾችን ያላቀረበበት ምክንያት አስገራሚ እንደነበር  ምንጮቻችን ገልጸዋል.', 'ለፍ/ቤቱ ተጠርጣሪ ወንጀለኞች ያለቀረበበት
ምክንያት ተጠይቆ ምላሽ የሰጠው ፖሊስ ጉዳያቸው በዕርቅ ሂደት ላይ በመሆኑ ሌላ ረጅም ተለዋጭ ቀጠሮ እንዲሰጥ የሚጠይቅ ነበር.', 'የክልሉ ፓሊስ
ኮሚሽን ክትትልና አቀራረብ ኃላፊ ኢ/ር አብርሃም ይሄንን ተቀባይነት የሌለው ጉዳዩ በዕርቅ እያለቀ ነው የምል አይነት እንድምታ ያለው መልስ
ለፍ/ቤቱ ያቀረቡ ቢሆንም የፍ/ቤቱ ዳኞች  ጉዳዩን ውድቅ ማድረጋቸውን በሥፍራው ከሚገኙ ምንጮች አረጋግጠናል.', 'የክልሉ ፖሊስ ለፍ/ቤቱ
የሰጠው በዕርቅ ሂደት ላይ ነው ያለው ምላሽ በክልሉ በሚገኙ ከፍተኛ ባለሥልጣናት የተመከረበትና  የተደገፈ መሆኑን ለፍ/ቤቱ ጨምሮ ቢገልፅም
ፍ/ቤቱ ግን ይህ ከባድ የዘር ማጥፋት ወንጀል በመሆኑ በዕርቅ ሂደት ማለቅ አይችልም፤ ይህን ትዕዛዝ የሰጡ ግለሰቦችም ሀላፊነት የጎደላቸው ናቸው ስል
ውድቅ አድርጉዋል.', 'በመሆኑም በጉዳዩ ላይ ብይን ለመስጠት በ22/03/2011 ተለዋጭ ቀጠሮ ሰጥቷል.', 'የክልሉ ፖሊስም ሆነ ትዕዛዝ የሰጡ
የክልሉ ከፍተኛ ባለሥልጣናት የያዙት ስልት ፍትህ ሂደቱን የማዘግየትና ጉዳዩ ውሳኔ  ሳያገኝ እንዲቀር የማድረግ እንደሆነ የወላይታ ሚዲያ ኔትዎርክ
መረጃ ይገልጻል.', 'በተጨማሪም በዚሁ ወንጀል ተከሰው ማረሚያ ቤት ከነበሩ ተከሳሾች አንዱ ሆን ተብሎ እንዲያመልጥ ተደርጎ በስህተት ከማረሚያ
ወጥቷል የሚል ምላሽ ለችሎቱ በክልሉ ፖሊስ ኮሚሽን እንደቀረበ ለማወቅ ተችሏል.', 'ጉዳዩ ምንም አይነት ትኩረት እንዳልተሰጠው የሚያሳየው
በቀድሞ ማረሚያ ቤት ለ6 ሰዎች ሞት ቀጥተኛ ተጠያቂ የነበሩ  ዮሐንስ ዮንቆራ የተሻለ ሹመት ክልሉ ላይ ተሰጥቷዋቸው እየሰሩ መሆናቸው ግርምት
እንደፈጠረባቸው ያነጋገርናቸው ሰዎች ገልጸዋል.', 'ይህ በእንዲህ እንዳለ መቅረብ ከነበረባቸው ተከሳሾች 25ቱ ያልቀረቡ  ሲሆን ቀጣይ በተያዘው
ቀጠሮ በ21/03 2011 እና 24/03/2011 የክልሉ ፖሊስ ተከሳሾችን እንዲያቀርብ ትዕዛዝ ተላልፏል.', 'በመሰረቱ ከዚህ በፊት የተፈጸመው እርቅ
ከፍትህ ሂደቱ ጋር እንደማይገናኝ  የተገለፀ ቢሆንም በዕርቁ ምክንያት ወንጀለኞችን ማቅረብ አለመቻል ሂደቱን ማጠልሸት እንደሆነ ያነጋገርናቸው
አስተያየት ሰጪዎች ገልጸዋል ', 'አሁንም በዕርቅ ስም ሆን ተብሎ የፍትህ ሂደቱን የማጨናገፍ ስራ  በአስቸኳይ መቆም ይገባዋል.']
```

And for word tokenizer,

```
from nltk.tokenize import word_tokenize

commnets  = "ማንኛውም ሰውደመወዝ እየተከፈለው በአሠሪው መሪነት በቀጥታም ሆነ በተዘዋዋሪ መንገድ ለተወሰነ ወይም ላልተወሰነ ጊዜ
ወይም የተወሰነ ሥራ ለአሰሪው ለመሥሪት ቢስማማ በሁለቱ መካከል የሥራ ውል ይመሠረታል"


amh_tokenize = word_tokenize(commnets)

print(amh_tokenize)
```

OUTPUT:

```
['ማንኛውም', 'ሰውደመወዝ', 'እየተከፈለው', 'በአሠሪው', 'መሪነት', 'በቀጥታም', 'ሆነ', 'በተዘዋዋሪ', 'መንገድ', 'ለተወሰነ', 'ወይም', 'ላልተወሰነ',
'ጊዜ', 'ወይም', 'የተወሰነ', 'ሥራ', 'ለአሰሪው', 'ለመሥሪት', 'ቢስማማ', 'በሁለቱ', 'መካከል', 'የሥራ', 'ውል', 'ይመሠረታል']
```

IV. NORMALIZATION

Since the Amharic language has some characters that represent the same sound like - ሀ፥ሐ፥ኀ፤ ሰ፥ሠ፤ አ፥ዐ and others, these need to be excluded and one character representation for the same sound is necessary. It also involves in normalizing the content of the data through different scenarios like short forms of the word using forward-slash (/) and period (.) for example ፕ / ር written as ፕሮፌሰር and አ.አ written as አዲስ አበባ.

3.4.4. METHODS OF DATA ANALYSIS

After the corpus was prepared concerning each post and comments the next step was analyzing the subjectivity and sentiment polarity of the data in each dataset.

I. SUBJECTIVITY DETECTION

To examine the sentence as subjective or objective the study classified the preprocessed comments and post. Each sentence is examined for subjectivity by comparing it to the content of the posts it belongs to. The analysis made on the subjective sentence quantity within each post and trends of the commenter regarding knowing or understanding the post they intended to comment on. Finally, only sentences or comments with subjective expressions were kept in the dataset for sentiment classification. The subjectivity classifier was trained using the Navies Bayes algorithm as other studies like the one conducted by [12] Used similar techniques.

II. SENTEMENT CLASSIFICATION

Sentiment or polarity classification was done by classifying each subjective sentence gain on subjectivity detection step in the textual dataset into positive, negative, and neutral sentences or comments sentiment classification groups. From the sentences that were collected or selected as subjective, the study was prepared a corpus of polarity or lexicon words.

In many research works conducted before recommended that valence shifter is appropriate in the works of sentiment analysis. There are two different aspects of valence shifting that are used to improve sentiment analysis system development. The First one is negations in which that can switch the sentiment of positive or negative terms in a sentence and the Second one is intensifiers, terms that can change the degree to which a word is positive or negative. From those listed valance shifters the study applied negation and opinion words within the corpus to classify. In this research, word intensifiers were not the concern of the study since they try to express the degree of the expressed comments or sentiments of the comments concerning the degree it imposes.

Polarity words are terms that can express opinions towards an object such as ‘ጥሩ’ (good) that expresses a positive opinion and ‘መጥፎ’ (bad) that expresses negative opinion towards an object. These terms are properly tagged in the lexicon with ‘+’ for positive opinion terms and ‘-’ for negative opinion terms. Then, if a term is found in the lexicon and if it’s corresponding value is ‘+’, this opinion term is positive. Similarly, if a term is found in the lexicon and if it’s corresponding value is ‘-‘, then this opinion term is negative.

CHAPTER FOUR

IMPLEMENTATION

4.1. OVERVIEW

This chapter presents clearly and clean preparations of categorized corpus developed for each of the data collected, the lexicon subjective word developed for each data sets from the comments, the data set of the subjective sentence, and the lexicon word dictionary developed for polarity. The dataset is then divided into training and test data to apply the proposed algorithm.

4.2. SUBJECTIVITY DETECTION

Work in polarity classification assumes that the incoming documents or sentences need to be opinionated. For many applications, we may need to decide whether a given document contains subjective information or not, or identify which portions of the document are subjective.

While there are several subjectivity lexicons available for research purposes in different languages it is not available for the Amharic language. We describe the process of constructing subjectivity lexicon(s) for recognizing sentence subjectivity from each content posted. We build subjectivity lexicons from the text content posted in each categorized corpus. We manually selected and added some Amharic words which have content related to the domain-based text.

In the literature several studies addressed sentence-level or sub-sentence-level subjectivity detection in a different domain, subjectivity classification is mostly known as a supervised learning task. Previous works [54][55] use machine learning techniques such as Naïve Bayes and Support Vector Machines with different types of features. In[54], the presence of pronouns; adjectives; and verbs were used as features to classify subjective sentences. In [55], a subjectivity detector was used to detect subjective sentences based on minimum cuts in the sentence graph. The technique first built a sentence graph using local labeling consistencies that produce an association score between two sentences. Sentences with similar association scores are more likely to belong to the same subjective or objective classes. More recently, [56], used nouns abstraction for in-domain and cross-domain subjectivity classification.

While the above approaches have performed moderately, however, it is difficult to determine effectively the comments belongingness with the post and we proposed lexicon/ dictionary of words based on the post and detecting the subjectivity of the comment/ sentence by applying the presence or absence of the word in the post within the comments or sentence. Therefore, by checking whether a list of words/dictionaries of the post is available within the sentences /comments to give we classify the comments subjectivity. So the lexicon of words was developed for each category of actors. To detect subjective sentence there are different approaches as we discussed in the literature review. And from those methods that different researchers applied the most common one is developing a lexicon of the subjective dataset from the general dataset.

To develop subjectivity lexicon different researchers used manual annotation and extended this data set of the lexicon by using seeding automatically. For example in the work of [57], the methods applied were automatically expanding manually prepared lexicon dictionary-based, distributional in the domain and out-of-domain information, as well as using Amazon Mechanical Turk to help “clean up” the expansions. They firstly manually annotated 400 words and the final result of Seed Lexicon contains 749 words, 406 positives, and 343 negatives.

So from the above work and some other similar works, we considered the problem of getting high subjective and sentiment classification accuracy without a labeled dataset(which means that the dataset generated automatically) from the target domain. That means that generating dataset using only automatically. To avoid the problem of not getting high subjective and sentiment classification accuracy without a labeled dataset in previous works, this study applied automatic extraction method for subjective lexicon and enriched it through manually feeding word family For that purpose as we discussed above, we present and evaluate the extraction method based on both automatically and enriched it through manually feeding the word family.

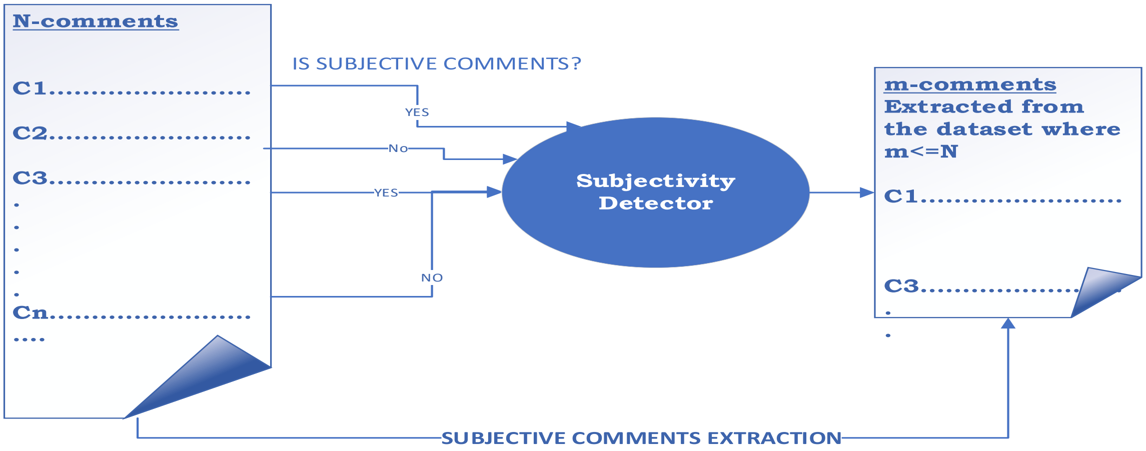

N-comments
C1........................
C2........................
C3........................
.
.
.
.
.
Cn........................
....
IS SUBJECTIVE COMMENTS?
YES
No
YES
NO
Subjectivity Detector
m-comments Extracted from the dataset where m<=N
C1........................
C3........................
.
.
SUBJECTIVE COMMENTS EXTRACTION

Figure 7 Subjective comment extraction

To extract subjective comments from the give list of comments which were fetched from the belonging post with automatic extraction and manually enriched them is handled in this work as depicted in the figure. Let us clarify the steps applied.

Step 1: The first step is getting preprocessed posts and comments from each categorized corpus where we do have N- comments.

Step 2: The subjective detector read post from the required categories (FANA, EBC, ESAT,..etc)

(This step is done for each category without the intervention of other corpora.)

Step 3: Create a list of words from the readied post using tokenization and n-gram generation.

Step 4: Write this generated list of words to Lexicon data of the corresponding categories of comments and posts.

Step 5: Then after finished automatically generating the list of words, in this research, unlike other work that felt the list of words automatically from the given input of sentence. Our work is based on a manual feeding list of words based on their Amharic morphological family of each word tokenized word taken from the post. The final lexicon of the subjective word is sum automatically developed list of words and manually enhanced list of words.

Step 6: After developing a lexicon of the subjective words from the post we need to compare each comment against the lexicon. To do this we need to read each comment individually and generate a list of words and their family with the same step we used to develop a lexicon of subjective words to the post.

Step 7: comparing the list of words generated on step 6 with that of step 5 with corresponding to the belonging lexicon data of the categories to determine whether the comment contains the subjective word that the post content has. And finally, if the comments are detected as subjective to the post, the comments list of words is passed to check it's polarity against the polarity data set.

```
# coding=utf-8
# Program to Check if given words appear together in a list of comments

def check(sentence, words):
    res = list(map(lambda x: any(map(lambda y: y in x.split(),
                                 words)), sentence))
    return [sentence[i] for i in range(0, len(res)) if res[i]]
```

So in this section, we described the subjectivity lexicon creation technique. The technique is further improved by using the subjective word extended approach, in our case through manual adding of words. Since our goal is to automatically identify likely subjective sentences using a set of automatically and manually annotations in each categorized corpus.

4.2.1. DATASET AND SUBJECTIVE LEXICON DICTIONARY

DATASET

In this study, the dataset consists of seven categories, Three datasets are from public media and 4 datasets are collected from well-known individuals based on their large number of followers on social media which are recognized by the world social media meter list.

For Daniel post(D_p) = {P_1, P_2,….., P_5} have 5 posts in total that have domain-based content of their post, and in each post, they have their collection of comments.

P_1 = {C01.txt, Co2.txt,……….,C101.txt,C102.txt} and for post 1 of Daniel, we do have 102 comments.

P_2 = {C01.txt, Co2.txt,……….,C128.txt,C129.txt} and for post 2 of Daniel, we do have 129 comments.

……………………..

P_5 = {C01.txt, Co2.txt,………., C123.txt, C124.txt} and for post 5 of Daniel, we do have 124 comments. So totally for Daniel's post a total of 5 posts and 700 comments used.

Given the above number of posts and comments under Daniel as an example, the following table is presented to summarize the number of posts and the sum of all comments made by followers under each category. And from the table- 5 up to the table- 10 the subjective dataset (subjective list of words taken from automatically and manually).

Categorized Dataset's	Posts	Comments
Daniel Berhane Posts	5	700
Jawar Mohammed Posts	2	120
Yoni Magana Posts	4	267
Fana broadcasting posts	7	1225
ESAT posts	7	812
EBC posts	4	512
Total	30	3636

Table 4 Categorized corpus/dataset with their total comments extracted

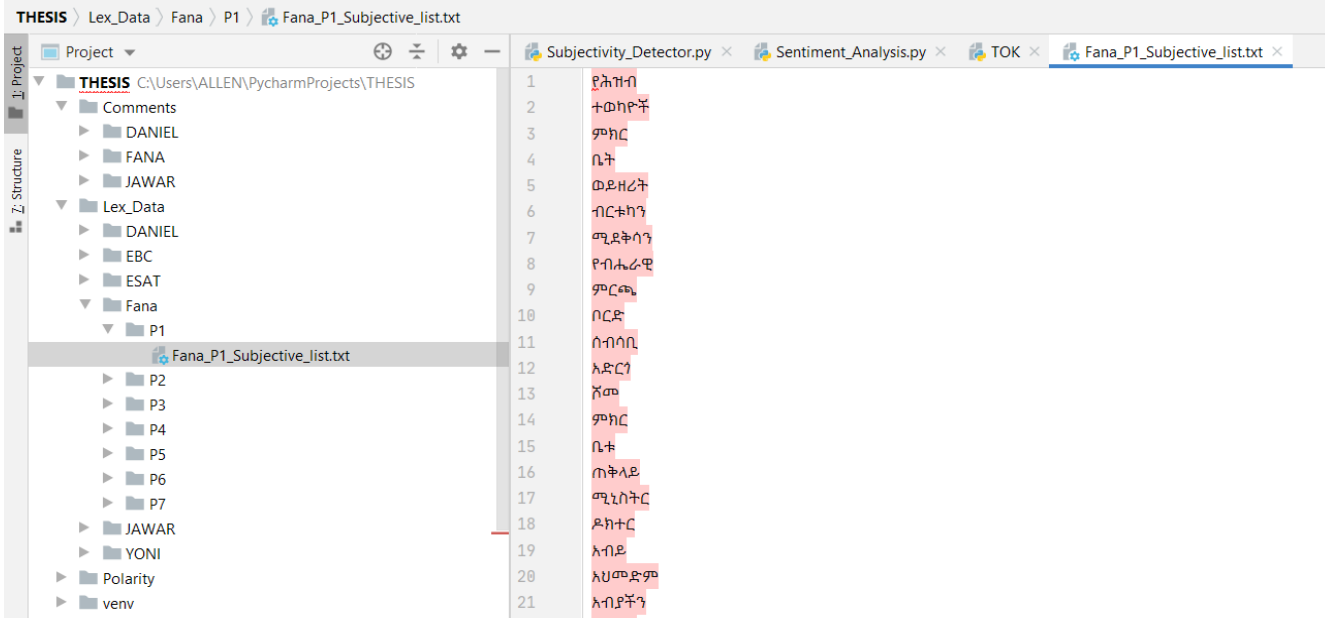
LEXICON DATASET

In this thesis work, we classified subjective sentences by developing a vector list of lexicon words form each belonging post content of the comments. From different works previously done [45], we learned the importance of lexicons to domains to improve construct validity when conducting dictionary-based automation and sense level lexicon developed [49]. To check if a sentence is subjective we trace the content and check if it contains two or more words from the list or dictionary of list words tokenized from the post which belongs to the lexicon of that post. Here, we first read and create a post content as a list of arrays by defining a function, and to read the content of the post we firstly try to divide the string of the data(post) into unigrams, tri-grams and tokens of the word.

This dataset consists of sentences together with subjectivity labels which are extracted from subjective words embedded in the comment. The Lexicon class loads an annotated dataset of subjective words(in other word contents of post) that have subjectivity and is used for subjective classifiers. For example sample subjective lexicon of the post looks like the following:

fana_list = ['የሕዝብ' 'ተወካዮች' 'ምክር' 'ቤት' 'ወይዘሪት' 'ብርቱካን' 'ሚደቅሳን' 'የብሔራዊ' 'ምርጫ' 'ቦርድ' 'ሰብሳቢ' 'አድርጎ' 'ሾመ' 'ምክር' 'ቤቱ' 'ጠቅላይ' 'ሚኒስትር' 'ዶክተር' 'አብይ' 'አህመድም' ‘አብያችን’ ‘አብይን’ '2012' 'ዓ.ም' 'የሕዝብ ተወካዮች' 'ምክር ቤት' 'ወይዘሪት ብርቱካን' 'የብሔራዊ ምርጫ ቦርድ' 'ጠቅላይ ሚኒስትር' 'ዶክተር አብይ' 'የሕዝብ ተወካዮች ምክር ቤት' 'ወይዘሪት ብርቱካን ሚደቅሳን' …………………………………]

So we’ll take this list of words which come from the post as the subjective lexicon of the post and by adding manually some related words. And then we trace subjectivity of the comment against this list.



```
የሕዝብ
ተወካዮች
ምክር
ቤት
ወይዘሪት
ብርቱካን
ሚደቅሳን
የብሔራዊ
ምርጫ
ቦርድ
ሰብሳቢ
አድርጎ
ሾመ
ምክር
ቤቱ
ጠቅላይ
ሚኒስትር
ዶክተር
አብይ
አህመድም
አብያችን
```

Figure 8 Structure of subjective lexicon from the GUI of PyCharm (Python Development tool)

The function works by taking the comment or sentence and comparing the words of the sentence that appear in each post with the subjective list in the dictionary file of the post-it belongs. Therefore, based on the presences of words in each sentence, each categorized corpus is split into subjective ones based on the lexical dictionary that we built by the above steps to the content of their belongings posts. Subjective sentences are further processed to polarity/ sentiments classification that is classified as positive, or negative opinions.

As we have discussed, in this study to develop a subjective lexicon, we first took the posts' content/sentence as input, then we tokenize it to get a list of words for the lexicon dictionary and we enhance manually the content to get rich words to be included in n-grams mode.

The output of generating a list of words from the post:

```
[ 'ሰሞኑን', 'ያዙን', 'ልቀቁን', 'የሚሉት', 'የመደመር', 'መንጋዎች', 'የሀተታቸው', 'መጀመሪያም', 'መጨረሻም', 'ትግራይን', 'እንውረር', 'በዚህ', 'ወይ',
'በዚያ', 'መልኩ', 'እናጥቃ', 'የሚል', 'የቀን', 'ቅዥት', 'መሰል', 'ቅብዥርዥር', 'መሆኑ', 'ተስተውሏል', 'ሁለት', 'ጥያቄዎች', 'ወደአእምሮዬ', 'መጡ',
'ከስጋት', 'የመነጨ', 'ይሆን', 'ህወሓት', 'አራትኪሎን', 'ተቆጣጥሮ፣', 'መደመር', 'ተብየውን', 'የጨረባ', 'ተዝካር', 'ያፈርስብናል', 'የሚል', 'የምር', 'ስጋት',
'ይሆን', 'እንዲያ', 'ከሆነ', 'ህወሓት', 'ምን', 'እንደሚያስብ', 'ባላውቅም', 'የትግራይ', 'ህዝብ', 'ግን', 'ለዚህ', 'የሚባክን', 'ጉልበት', 'የለኝም', 'ብሎ',
'ደምድሟል።', 'የስከዛሬውም', 'በኪሳራ', 'ተመዝግቧል።', 'ስለዚህ', 'ስጋት', 'አትግባችሁ', 'ከአሸናፊነት', 'ስካር', 'የመነጨ', 'ይሆን', 'እንኳን', 'እናንተ',
'ብዙ', 'ኃያላን', 'ትግራይ', 'ላይ', 'ጦርነት', 'ከፍተው', 'በውርደት', 'ተሰናብተዋል',……………………………………….]
```

As we can see from the above output which is the study's take a list or dictionary of words where generate automatically will not be a sufficient list of words. The reason for this one is when we type comments we may miss some character or letter of the word, secondly since Amharic is a morphologically rich language we need to consider including a different description of words. For example, someone comment as: "የትግራይ ህዝብም ሆነ አስተዳደር ህገመንግስቱን ጥሳችዋል" as we can see from this comment the word "የትግራይ" does not include in the list of a word which generates from the post instead it appears as "ትግራይ". Therefore the subjective detector will not take this comment as subjective since the word is not listed in the list of words generated from the post contents or sentence. Additionally, from those words generated by the above step, we also add sentences' or posts' list of n-gram words to the lexicon dictionary. Generally, we can show here that to develop a subjective lexicon automatically has an effect on the accuracy of the analysis. Here we need to take token and n-gram words to enhance the content of the list of words taken to the lexicon of each categorized post.

```
def token_word()
```

```
wordcount = {}

file = open(r"C:\\Users\\ALLEN\\PycharmProjects\\THESIS\\Comments\\DANIEL\\P1\\P1.txt",
"rt",           encoding="utf-8-sig")
data = file.read()
words = data.split()
token = word_tokenize(data)
return token
print(token)
print('Number of words in text file :', len(words))
for word in data.split():
    if word not in wordcount:
        wordcount[word] = 1
    else:
        wordcount[word] += 1
for word1 in token:
    if word1 not in wordcount:
        wordcount[word1] = 1
    else:
        wordcount[word1] += 1
wordcount = sorted(wordcount.items(), key=lambda x: x[1], reverse=True)

for  k, v in wordcount[:111]:
    print(k, v)
    return (k,v)
```

The number of words in a text file: 111. As we can see from the above post we do have about 111 words and to take this list of words as a more relevant list for the lexicon of the post its output has low accuracy when we came to whether or not the given comment is subjective. So in this work, as it was done in [53], the list of words taken from the post was enhanced by developing n-gram words and enriched them by adding a family of the word manually. The n-gram outputs which were added as list words or dictionary represented in vector from the post content are as follow:

uni-gram: ['ሰሞኑን', 'ያዙን', 'ልቀቁን', 'የሚሉት', 'የመደመር', 'መንጋዎች', 'የሀተታቸው', 'መጀመሪያም', 'መጨረሻም', 'ትግራይን', 'እንውረር', 'በዚህ', 'ወይ', 'በዚያ', 'መልኩ', 'እናጥቃ', 'የሚል', 'የቀን', 'ቅዥት', 'መሰል', 'ቅብዥርዥር', 'መሆኑ', 'ተስተውሏል', 'ሁለት', 'ጥያቄዎች', 'ወደአእምሮዬ', 'መጡ', 'ከስጋት', 'የመነጨ', 'ይሆን', 'ህወሓት', 'አራትኪሎን', 'ተቆጣጥሮ፣

', 'መደመር', 'ተብየውን', 'የጨረባ', 'ተዝካር', , ……………………………………] this one obviously a token of the post.

bi-gram: ['ሰሞኑን ያዙን', 'ያዙን ልቀቁን', 'ልቀቁን የሚሉት', 'የሚሉት የመደመር', 'የመደመር መንጋዎች', 'መንጋዎች የሀተታቸው', 'የሀተታቸው መጀመሪያም', 'መጀመሪያም መጨረሻም', 'መጨረሻም ትግራይን', 'ትግራይን እንውረር', 'እንውረር በዚህ', 'በዚህ ወይ', 'ወይ በዚያ', 'በዚያ መልኩ', 'መልኩ እናጥቃ', 'እናጥቃ የሚል', 'የሚል የቀን', 'የቀን ቅዠርት', 'ቅዠርት መሰል', 'መሰል ቅብዠርዠር', 'ቅብዠርዠር መሆኑ', , …………………]

tri-gram: ['ሰሞኑን ያዙን ልቀቁን', 'ያዙን ልቀቁን የሚሉት', 'ልቀቁን የሚሉት የመደመር', 'የሚሉት የመደመር መንጋዎች', 'የመደመር መንጋዎች የሀተታቸው', 'መንጋዎች የሀተታቸው መጀመሪያም', 'የሀተታቸው መጀመሪያም መጨረሻም', 'መጀመሪያም መጨረሻም ትግራይን', 'መጨረሻም ትግራይን እንውረር', 'ትግራይን እንውረር በዚህ', 'እንውረር በዚህ ወይ', 'በዚህ ወይ በዚያ', 'ወይ በዚያ መልኩ', 'በዚያ መልኩ እናጥቃ', 'መልኩ እናጥቃ የሚል', 'እናጥቃ የሚል የቀን', 'የሚል የቀን ቅዠርት', 'የቀን ቅዠርት መሰል', 'ቅዠርት መሰል ቅብዠርዠር'…………….]

But in this work in addition to the above output of the list of which are automatically generated, we added manually some words that are similar and morphologically familiar to the word already listed above and to the post content [50][53]. As described with the above example after the lexicon was enriched manually, the content of the post in the list of words that were used for lexicon data of each respective post categories consisted 111 words of posts, 205 unigram and bi-gram words, 130 trigram words and around 60 manually annotated words. Finally, about 500 words were extracted from only one post for the lexicon dataset developed.

Unlike previous studies, the development of a lexicon dataset which was enriched with a domain-based list of words for each subjective list of categorized posts made a subjective detector more accurate.

To prove this one, for example when we try to detect subjectivity of the comments give on Daniel’s post(P1) using the only automatically fetched list of words for subjective lexicon, from the total of cleaned pre-processed 102 comments only 12 comments were found to be subjective to the post they belong. But by taking the list of the words developed by both automatic and manual methods, 39 comments were found to be subjective to the post they belong.

Secondly, other similar studies [46][47][48][50] conducted based on subjective and sentiment lexicons such as AFINN lexicon, Bing Liu’s lexicon, MPQA subjectivity lexicon, Sent WordNet, VADER lexicon were the English language-based and used either directly or through translation, But this study developed its own Amharic language lexicon within the domain of the research area. This makes our study different from others.

Thirdly, previous studies prepared sentiment and subjective lexicon from the total data they get from web scrabbling. But this study applied a categorical lexicon development method that avoided the bias that could be created as a result of adding all the comments in one dataset. This is also another technique that differentiates this study from other similar studies. For example, let us take the above example by cases.

Case 1: automatic generation of lexicon dataset

To show this one let us take Daniel's post(P1) and that consists of three comments(C01.txt, C02.txt, and C03.txt)

C01.txt = 'አንተም በራስ መንገድ እኔም በኔ መንገድ ማን ቀድሞ እንድገባ እናየዋለን'
C02.txt = 'የደቡብ ክልል ፖሊስ ኮሚሽን በሰኔ ወር በሃዋሳ ተከስቶ በነበረው የዘር ማጥፋት ወንጀል ተሳታፊ የነበሩ ተጠርጣሪ ወንጀለኞችን ጉዳያቸው በዕርቅ ስላለቀ ለፍ/ቤት ማቅረብ አልቸልም'
C03.txt = 'አረ ባክህ ትግራይ ከመቼ ጀምሮነው ሀገር የሆነው በል አታቅራ ማቅራራት ሃላ ቀርነትነው ደርግ አሸነፍኩ ብሎ የለም'

and when we sum up content of these comments and add them in one dataset and try to generate list of words for lexicon data from the three comments cumulatively the result will be:

docs = 'አንተም በራስ መንገድ እኔም በኔ መንገድ ማን ቀድሞ እንድገባ እናየዋለን የደቡብ ክልል ፖሊስ ኮሚሽን በሰኔ ወር በሃዋሳ ተከስቶ በነበረው የዘር ማጥፋት ወንጀል ተሳታፊ የነበሩ ተጠርጣሪ ወንጀለኞችን ጉዳያቸው በዕርቅ ስላለቀ ለፍ/ቤት ማቅረብ አልቸልም አረ ባክህ ትግራይ ከመቼ ጀምሮነው ሀገር የሆነው በል አታቅራ ማቅራራት ሃላ ቀርነትነው ደርግ አሸነፍኩ ብሎ የለም' and a tokens of

['አንተም', 'በራስ', 'መንገድ', 'እኔም', 'በኔ', 'መንገድ', 'ማን', 'ቀድሞ', 'እንድገባ', 'እናየዋለን', 'የደቡብ', 'ክልል', 'ፖሊስ', 'ኮሚሽን', 'በሰኔ', 'ወር', 'በሃዋሳ', 'ተከስቶ', 'በነበረው', 'የዘር', 'ማጥፋት', 'ወንጀል', 'ተሳታፊ', 'የነበሩ', 'ተጠርጣሪ', 'ወንጀለኞችን', 'ጉዳያቸው', 'በዕርቅ', 'ስላለቀ', 'ማቅረብ', 'አልቸልም', 'አረ', 'ባክህ', 'ትግራይ', 'ከመቼ', 'ጀምሮነው', 'ሀገር', 'የሆነው', 'በል', 'አታቅራ', 'ማቅራራት', 'ሃላ', 'ቀርነትነው', 'ደርግ', 'አሸነፍኩ', 'ብሎ', 'የለም']

As we can see the first comment C01.txt content it doesn’t have subjective relationship with the post and if someone comment as “መንገድ ይገነባል” the subjective detector will take this comment as subjective to the post it belongs. But in this study we follow the reverse one to generate list of word and it will protect the detector from missing the target goal of the research.

Case 2: Manually

From the above example let us add morphs family of the word 'መንገድ’, ‘መንገዳችን‘ and 'ተሳታፊ' with ‘ስብሰባ ተሳታፊ’ based on the word if someone comments as “መንገዳችን ተጠናቀቀ” and ‘ስብሰባ ተሳታፊ’ the two words “መንገዳችን’ and ‘ተሳታፊ’ the subjective detector will take them as subjective to the post(Daniel’s post). Below is a list of tables that have each category's subjective list and the number of words within the dataset.

Lexicon data set of Daniel Posts	Number of words in the list generated automatically	Number of words in the list generated Manually	Total number of the word in each list
Daniel_P1_Subjective_list	446	30	476
Daniel_P2_Subjective_list	374	40	414
Daniel_P3_Subjective_list	356	40	396
Daniel_P4_Subjective_list	321	50	371
Daniel_P5_Subjective_list	441	30	471
Total subjective list	1938	190	2138

Table 5 A subjective list of words in Daniel’s post

Lexicon data set of Fana’s Broadcasting Posts	Number of words in the list generated automatically	Number of words in the list generated Manually	Total number of the word in each list
Fana_P1_Subjective_list	728	50	778
Fana _P2_Subjective_list	1022	50	1072
Fana _P3_Subjective_list	2170	30	2200
Fana _P4_Subjective_list	836	30	866

Fana _P5_Subjective_list	1007	40	1047
Fana _P6_Subjective_list	906	30	936
Fana _P7_Subjective_list	2200	80	2280
Total subjective list	8869	310	9179

Table 6 A subjective list of words in FANA's post

Lexicon data set of ESAT Broadcasting Posts	Number of words in the list generated automatically	Number of words in the list generated Manually	Total number of the word in each list
ESAT_P1_Subjective_list	600	50	650
ESAT _P2_Subjective_list	820	30	850
ESAT _P3_Subjective_list	747	40	787
ESAT _P4_Subjective_list	1127	20	1147
ESAT _P5_Subjective_list	526	30	556
ESAT _P6_Subjective_list	438	40	478
ESAT _P7_Subjective_list	587	30	617
Total subjective list	4845	240	5085

Table 7 A subjective list of words in ESAT's post

Lexicon data set of EBCZENA' Posts	Number of words in the list generated automatically	Number of words in the list generated Manually	Total number of the word in each list
EBCZENA _P1_Subjective_list	632	50	682
EBCZENA _P2_Subjective_list	459	20	479
EBCZENA _P3_Subjective_list	393	50	443
EBCZENA _P4_Subjective_list	746	30	776
Total subjective list	2230	150	2380

Table 8 A subjective list of words in EBCZENA's post

Lexicon data set of JAWAR's Posts	Number of words in the list generated automatically	Number of words in the list generated Manually	Total number of the word in each list
JAWAR _P1_Subjective_list	164	50	214
JAWAR _P2_Subjective_list	126	50	176
Total subjective list	290	100	390

Table 9 A subjective list of words in JAWAR's post

Lexicon data set of YONI's Posts	Number of words in the list generated automatically	Number of words in the list generated Manually	Total number of the word in each list
YONI _P1_Subjective_list	362	50	412
YONI _P2_Subjective_list	477	30	507
YONI _P3_Subjective_list	508	20	528
YONI _P4_Subjective_list	266	50	316
Total subjective list	1613	150	1763

Table 10 A subjective list of words in YONI's post

Finally, using the subjective detector algorithm discussed on step 7 comparing the list of words generated on step 6 (subjective list of words from the post content and add some useful words to enhance the word list) with corresponding lexicon data of the categories to determine whether the comment contains the subjective word that the post content has. After implementing each category's subjective comment detected through the algorithm the cumulative output of subjective comment in each category looks like the one presented in the following table. And this list of comments is passed to the polarity detection algorithm.

Categorized Dataset’s	Total Subjective comments from each category (sum of all subjective comments identified)
Daniel Berhane Posts	300
Jawar Mohammed Posts	18
Yoni Magana Posts	107
Fana broadcasting posts	635
ESAT posts	372
EBC posts	164
Total subjective comments	1596

Table 11 Total subjective comments extracted from each category

4.3. POLARITY DETECTION

As mentioned above, today, very large amounts of information are available in online documents, and a growing portion of such information comes in the form of people’s experiences and opinions. It has proven quite useful in such contexts to create summaries of people’s opinions that consist of subjective expressions extracted from their thought or just their thought’s polarity(positive or negative).

In this section, as we discussed in the previous chapter the study followed the hybrid sentimental classification approach. Therefore, the polarity classification started from labeling an opinionated subjective comments as either positive or negative. To do so first a polarity word was created as a generally positive and negative Amharic word of a dataset through manual annotation from the domain of this research perspective. And then the manual annotation was improved through automatically extracting sentimental words. Therefore based on the given lexicon of polarity it is possible to decide the polarity of the sentence.

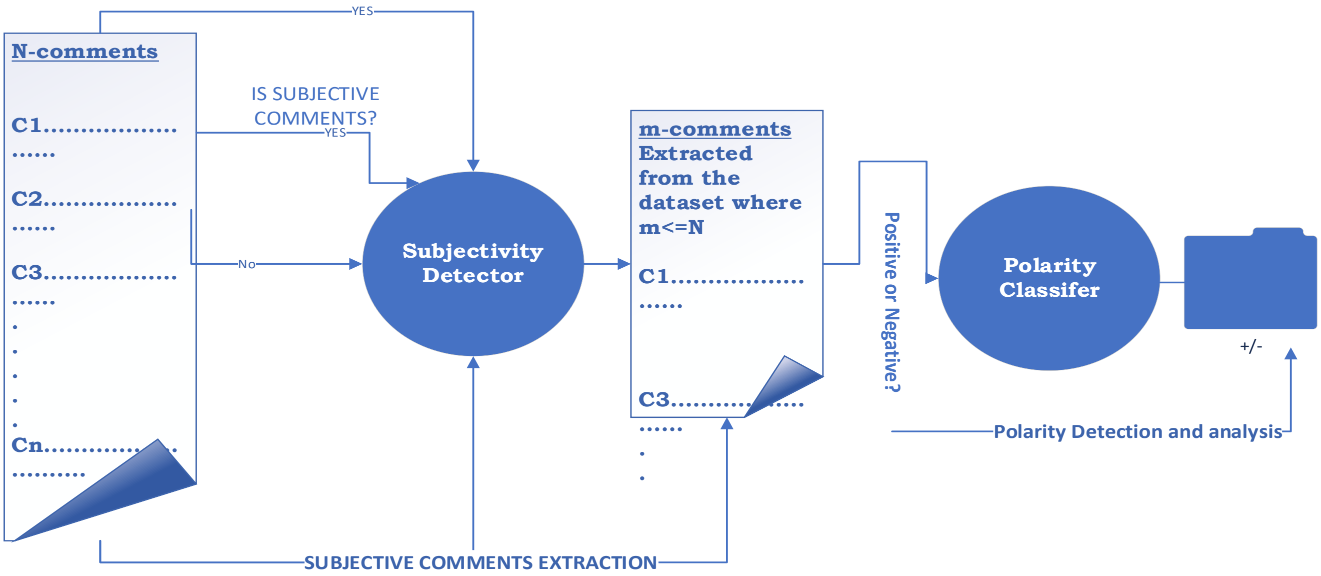


Figure 9 Polarity classification via subjectivity detection.

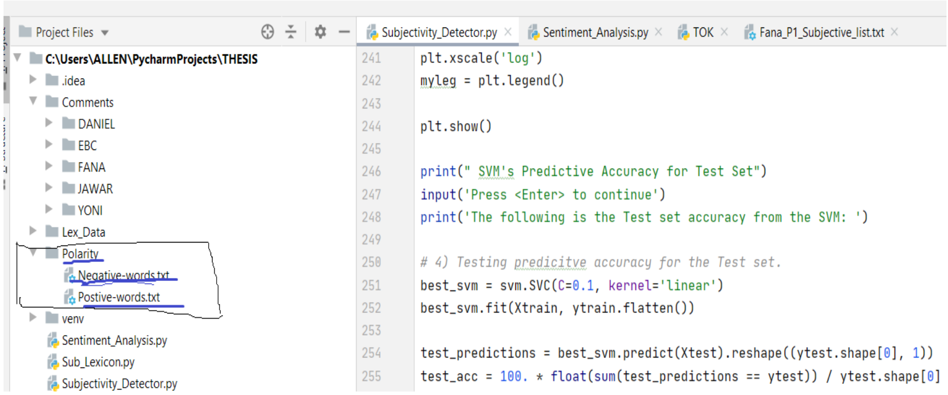
As it is depicted in the above figure the polarity detection is a phase next to the detection of subjective comments which was already been discussed in the previous section. A domain-based polarity dataset (negative and positive) was developed manually and most of them are words used for hate speech, and insulting, encouraging, appreciation words currently used on social media for the political domain of words. Therefore by comparing a list of words generated from subjective detection of each subjective sentence or comments against polarity words generated manually.

4.3.1. POLARITY DATASET

The polarity dataset is a set of words with polarity labels that we created for our work on polarity classification. The work described in this section is based on two corpora we developed manually from different Amharic language sources (online, by manually fetch and from other works done on sentiment analysis) which are positive-word and negative-words.

Negative-words:	Positive-words:
(ሌብነት,ሌባ,ሌቦችህ,ሌቦች,የሌባዘር,ደነዝክ,ደነዝ,ደነዞች,እባብ,እባቦች,ቱሪናፋ,ባንዳ,ባንዳነት,ነፍሰ,ገዳይ,ልቅምቅም,መግደል,ይገደሉ,ሴይጣን,የሴይጣን,ድንጋዬች,ድንጋይ,ትል,እብደት,እብድ,ቅዠት,ቅዠታም,ቅማል,ቅማላም,ጥፍር,ጥፍራም,ዘቅጠህ,የዘቀጠ,የዘቅጥክ,ክፉ,ድንጋይራስ,ውሻ,ሆዳም,…………..)	(ፍቅር,በፍቅር,መዋደድ,ፖዘቲቭ,አንድነት,ትክክል,ትክክልነህ,ግሩም,መደመር,የመደመር,በመደመር,የአንድነት,አንድነት,በአንድነት,ፍቅር,የፍቅር,በፍቅር,መፋቀር,ስልፍቅር,የመደመር የፍቅር,የአንድነት,ሰላም,ለሰላም,ጎበዝ,ጥሩ,መልካም,ቀጥልበት,ሀሪፍ ………………)

Table 12 Polarity Dataset



```
plt.xscale('log')
myleg = plt.legend()

plt.show()

print(" SVM's Predictive Accuracy for Test Set")
input('Press <Enter> to continue')
print('The following is the Test set accuracy from the SVM: ')

# 4) Testing predicitve accuracy for the Test set.
best_svm = svm.SVC(C=0.1, kernel='linear')
best_svm.fit(Xtrain, ytrain.flatten())

test_predictions = best_svm.predict(Xtest).reshape((ytest.shape[0], 1))
test_acc = 100. * float(sum(test_predictions == ytest)) / ytest.shape[0]
```

Figure 10 Polarity Dataset arrangement

Therefore, the sentiment or polarity algorithm works based on the polarity dataset which was prepared manually as a positive-words and negative-words list. The subjective sentence which is extracted from the subjective detector steps was passed to the sentiment algorithm and the algorithm classified as a negative or positive comment by checking the lexicon of polarity words against the sentence or comment(here the comment is the subjective one).

4.3.2. FEATURE EXTRACTION

The feature extractor of this program takes in a tokenized subjective comments/sentence, checks what words it contains about the features that were extracted from our manually labeled data(positive and negative) which was developed manually and returns a dictionary of this information. Machines cannot understand the raw text and only see numbers, particularly statistical techniques such as machine learning can only deal with numbers. Therefore, we need to convert our text into numbers.

Different approaches exist to convert text into the corresponding numerical form. The Bag of Words Model and the Word Embedding Model are two of the most commonly used approaches. Some works [46] employed Particle Swarm Optimization (PSO) based feature selection algorithm for obtaining an optimized feature set for training and evaluation. System evaluation shows interesting results on the four emotion datasets i.e. anger, fear, joy, and sadness.

In this study, unlike [51][53], we used the TFIDF model to convert our text to numbers. The dictionary has Boolean values of what words the tokenized review contains with the training feature set. After we have our training set we generate our classifier model by utilizing the nltk.classify.apply_features function with our feature extractor function and list of tokens of comments as parameters.

The bag of words approach works fine for converting text to numbers. However, it has its drawback.[53]. It assigns a score to a word based on its occurrence in a particular document. It doesn't take into account the fact that the word might also be having a high frequency of occurrence in other documents as well. TFIDF resolves this issue by multiplying the term frequency of a word by the inverse document frequency. The TF stands for "Term Frequency" while IDF stands for "Inverse Document Frequency".

The term frequency is calculated as:

Term frequency = (Number of Occurrences of a word)/(Total words in the document)

And the Inverse Document Frequency is calculated as:

IDF(word) = Log((Total number of documents)/(Number of documents containing the word))

The TFIDF value for a word in a particular document is higher if the frequency of occurrence of that word is higher in that specific document but lower in all the other documents.

Since we can directly convert text documents into TFIDF feature values (without first converting documents to a bag of words features). We use the following script to convert directly:

```
# *-- coding=utf-8 --*
from nltk.corpus.reader import documents
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

tfidfconverter = TfidfVectorizer(max_features=1000, min_df=1, max_df=1)
comments = 'አንተም በራስ መንገድ እኔም በኔ መንገድ ማገ ቀድሞ እንድገባ እናየዋለን ' \
              'የደቡብ ክልል ፖሊስ ኮሚሽን በሰኔ ወር በሃዋሳ ተከስቶ በነበረው የዘር ማጥፋት   ወንጀል ተሳታፊ የነበሩ ተጠርጣሪ
ወንጀለኞችን ጉዳያቸው ' \
            'በዕርቅ ስላለቀ ለፍ/ቤት ማቅረብ አልችልም ስል ምላሽ ሰጠ የፌደራል ከፍተኛ ፍርድ ማጥፋት ወንጀል የተከሰሱ ግለሰቦችን
ክስ ይመለከታል'
X = tfidfconverter.fit_transform(comments).toarray()
print(X)
```

We use the Tfidf Vectorizer class from the sklearn.feature_extraction.text library and we use some parameters that are required to be passed to the constructor of the class. The first parameter is the max_features parameter, which is set to 1000. We set the max_features parameter to 1000, which means that we want to use 1000 most occurring words as features for training our classifier. The next parameter is min_df and it has been set to 1. So we only include those words that occur in at least 1 sentence. Similarly, for the max_df, feature the value is set to 1. Here 1 means that we should include only those words that occur in a maximum of 100% of all the sentences. Finally, the fit_transform function of the Tfidf Vectorizer class converts text sentence into corresponding numeric features.

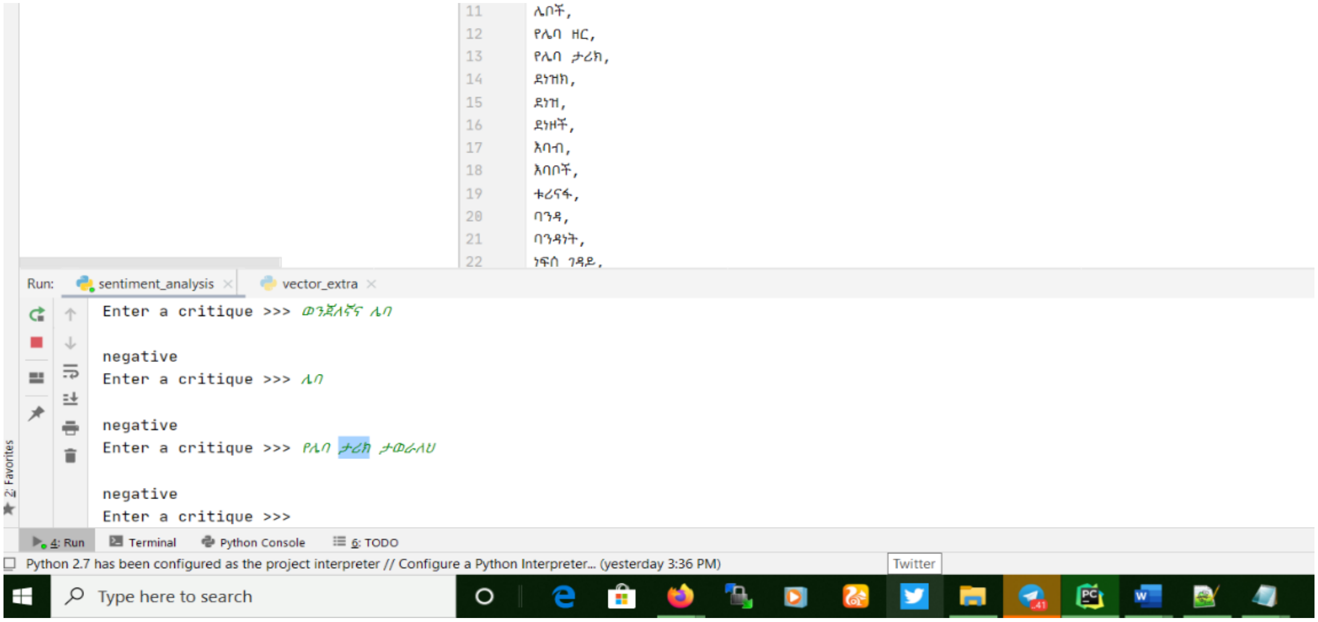
And finally, we need to divide our data into training and testing sets. To do so, we will use the train_test_split utility from the sklearn.model_selection library.

Using the following script:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

This divides data into a 20% test set and an 80% training set. To train our machine learning model using the random forest algorithm we will use the NaiveBayesClassifier class from the nltk library. The training class of this classifier is used to train the algorithm. We need to pass the training sets to this method. Let us look at the script:

```
training_set = nltk.classify.apply_features(extract_features, all_comments)
classifier = nltk.NaiveBayesClassifier.train(training_set)
return classifier
```



```
ሌቦች,
የሌባ ዘር,
የሌባ ታሪክ,
ደነዝክ,
ደነዝ,
ደነዞች,
እባብ,
እባቦች,
ቱሪናፋ,
ባንዳ,
ባንዳነት,
ነፍሰ ገዳይ,
```

```
Enter a critique >>> ወንጀለኛና ሌባ

negative
Enter a critique >>> ሌባ

negative
Enter a critique >>> የሌባ ታሪክ ታወራለህ

negative
Enter a critique >>>
```

Python 2.7 has been configured as the project interpreter // Configure a Python Interpreter... (yesterday 3:36 PM)

Figure 11 sample output after training our model to detect the polarity of word

Once we have our classifier model trained with the labeled dataset and extract some features we can predict the rest of the dataset in each category. Then we will take comments given on the belonging post as an input, tokenize it, extract its features using our feature extractor and pass the list of features into the classifier model's classify function. The result is a label which in this case can be 'positive' or 'negative'.

Therefore, after each subjective comment (1596 comments) within their categories evaluated through our polarity detector/sentiment, we get about in total of 469 comments are positive which is about 29.38 % and 70.61% of the total comments are negative which means that about 1127 comments.

Categorized Dataset’s	Total Subjective comments from each category (sum of all subjective comments identified)	Positive comments in each categorized dataset’s	Negative comments in each categorized dataset’s
Daniel Berhane Posts	300	38	262
Jawar Mohammed Posts	18	7	11
Yoni Magana Posts	107	78	29
Fana broadcasting posts	635	203	432
ESAT posts	372	102	270
EBC posts	164	41	123
Total comments (subjective)	1596	469	1127

Table 13 Total subjective comments and their polarity (positive and negative) extracted from each category.

Variation in results across different categories was caused by first, variations in the number of comments given under each post; second, exclusion of comments written in other languages than Amharic like Oromiffa, Tigrigna, English, and other Ethiopian languages; and third, unlike some other works this study didn’t use a multilingual approach that requires interpreter across languages which is limited in the case of Ethiopian languages.

4.4. DISCUSSION OF EXPERIMENT

The result shows that the accuracy of the classification tasks, i.e., Bag-of-Word (BoW), TIFD features and n-grams features was enhanced by enriching automatically generated lexicon of each subjective dataset and polarity dataset with manually generated list of words. This increased the accuracy of the classification of the subjectivity and polarity of the comments. For example, 111 words were automatically generated from Daniel’s post and after enriching the lexicon with manually added words the total number of list of words became 2138. This increased the probability of a word to appear in the given comment and protected the classifier from losing the content of the comment of individuals.

As we can see from the result of this research work from a total of 3636 comments less than half of them (1596 – 44%) were found to be subjective to the post they belong and this was achieved through enhancing the lexicon dataset and increasing accuracy of the detector. This indicates that given Ethiopia's political discourse on social media, more than half of the followers commented on the issue posted without understanding the content of the post. In addition, from these subjective comments only 29.38% of the comments were positive and above 70% of comments constitute hate speeches, insultings and negative thoughts. This implies that from less than half of subjective sentence 70% of the comments were found to be negative responses.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1. CONCLUSION

In this study, Facebook was taken as a source of social media content and a categorical dataset was created through querying using a common set of domain-based topics. Two groups were created consisting public media and individual activists and comments were classified under each category based on their belongingness to the post. A lexicon was also developed for each category. Then, subjective expression was detected based on the occurrence of the list of words within the comment in the corresponding lexicon and sentiment expressions were detected using domain-based polarity dataset. .

We conclude that although the gap in the general proportion of subjective sentence and non-subjective sentence in a given comment, remained narrow, 44% and 56% respectively, there are marked differences between sentiment or polarity of the comments across all the categories , which is about 70% of the comments are negative and the remaining are positive comments.

The major concern of most of the previous studies is building a classification model and comparing various algorithms based on their performance. In addition, most of the classification models are developed from the linguistic perspective. In this study subjectivity and polarity analysis was conducted based on actual posts and comments belonging to the posts within the polical discourse domain using Amharic Language. In addition, this study has scientifically proved that the judgments by scholars, politicians, and individuals regarding the misuse of the social media is true.

One of the main contributions of this study is to deliver Amharic language corpus on the political domain using the categorical concept, Lexicon of subjective dataset and polarity dataset. Eventhough the stated factor was challenging while conducting the study, the study developed the first lexicon and polarity dataset in the Ethiopian language. The lexicon dataset for subjectivity was developed through the same approach except that the dataset was enhanced by learning from testing the classifier after it is trained. In addition, how SVM and Naïve Bayes model is good enough to perform subjectivity and sentiment classification of Amharic dataset was tested.

The model developed by this study can be used by a social media company like Facebook, to have content and belongingness detector addons on their webpage and applications. Other concerned bodies like policymakers, politicians, activists etc. can use the model to measure the extent of subjectivity and polarity in the comments belonging to a particular post and provide more information to their followers in an interactive way and make the idea they share more clear and understandable.

5.2. CHALLENGES OF SENTIMENT AND SUBJECTIVITY ANALYSIS

There are several defined elements in a piece of text that factor into sentiment analysis: the object, the attributes, the opinion holder, the opinion orientation, and the opinion strength. To obtain complete, accurate, and actionable information from a piece of text, it’s important to not only identify each of these five elements individually but to also understand how they work together to provide the full context and sentiment. Because keyword processing only identifies the sentiment reflected in a particular word, it typically fails at providing all of the elements necessary to understand the complete context of the entire piece.

Some challenges were faced in this sentiment analysis and subjectivity detection work. From those challenges, the first one was the process of collecting different posts within the domain of political issues and their belonged comment given in Amharic. Collecting Amharic comments about political issues is a difficult task because it is not easy to do it manually. To overcome this problem, the study followed the approach of doing it using Facebook API and using python scripts. The other big challenge in collecting the posts and comments was identifying the Amharic comment from the unprocessed data. The other challenge was a manual annotation of word from the subjectively treated sentence for sentiment analysis because it is difficult to decide whether a given comment is positive or negative without considering the post it belongs and the intent it tries to describe, it depends on the context and the posts it belongs.

5.3. FUTURE WORK

Although the provided datasets, algorithms or classifiers, analyses, and methodologies are quite a good many different works, tests, and experiments have been left for the future due to lack of time (i.e. the experiments with real data are usually very time consuming, requiring even days to finish a single run). Future work concerns a deeper analysis of particular mechanisms, application of across different languages, using translators, new proposals to try different methods, or simply like other works comparison of the model using already developed dataset.

This thesis has been mainly focused on only domain based datasets, single language detection, and Navies classifier, but this makes the scope of the research narrowed and the following ideas could be tested:

1. It could be interesting to take the whole comments given on the belonging post including non- Amharic language as a source of the data to preprocess like Oromiffa, Tigrigna, etc.
2. It could be also interesting to compare different algorithms using the prepared dataset.
3. More categorical sources could also add value to the work of analysis like across some different Ethiopian languages.
4. It could also interesting to compare the model using the English language where almost every researcher works on.
5. Comparing the performance of the classifier across different Ethiopian languages could also make relevant for the analysis.
6. Concerning the result for both subjectivity analysis and sentiment analysis, we can also expect to improve them by having richer graphs and tables with more descriptive attributes.

REFERENCE

[1] Iginio Gagliardone, 2018: Mediatization of politics on social media in Ethiopia.

[2] J. K. Ahkter, S. Soria, "Sentiment analysis: Facebook status messages," Technical report, Final Project CS224N, Stanford University, 2010.

[3] National Abat, 2015: accessed on 03/04/2015,https://ecadforum.com/ethiopia-misuse-of-social-media-and-threat-it-imposes-to-our-coexistence.

[4] Emma Haddia, XiaohuiLiua, Yong Shib, 2017: The Role of Text Pre-processing in Sentiment Analysis.

[5] Bing Liu, 2012: Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.

[6] M.Abdul-Mageed, M. Diabc, and S. Kübler,2013, "SAMAR: Subjectivity and Sentiment Analysis for Arabic Social Media," Computer Speech and Language, Vol. 28, No. 1.

[7] AlessiaD'Andrea, Fernando Ferri, PatriziaGrifoni, TizianaGuzzo, 2016: Approaches, Tools, and Applications for Sentiment Analysis Implementation.

[8] Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Stanovich, Denis Stukal, and Brendan Nyhan, 2015: Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.

[9] Megan Fountain, 2017: Social Media and its Effects in Politics: The Factors that Influence Social Media use for Political News and Social Media use Influencing Political Participation.

[10] Katharine E. Van Wyngarden, 2016: Young adults' political engagement using Facebook.

[11] Phillip Smith: Sentiment Analysis: Beyond Polarity.

[12] Ellen Riloff and Janyce Wiebe, 2015: Learning Extraction Patterns for Subjective Expressions.

[13] F. Anta, L. N. Chiroque, and P. Morere, A. Santos, "Sentiment analysis and topic detection of Spanish tweets: A comparative study of NLP techniques," La Revista de Procesamiento de Lenguaje Natural, vol. 50, pp. 45–52, 2013.

[14] B.Pak, and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of the Seventh International Conference On Language Resources and Evaluation (LREC), pp. 1320–1326, Valletta, Malta, May 19–21, 2010.

[15] Saif M.Mohammad, Challenges in Sentiment Analysis,2015.

[16] Selam Gebremeskel, 2010: sentiment mining model for opinionated Amharic texts.

[17] Seid Muie, 2016,"Brife Analysis of Amharic NLP".

[18] Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, Tiziana Guzzo. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. International Journal of Computer Applications , 0975 – 8887.

[19] Assabie, Y. (2015). Natural Language Processing. Addis Ababa University, Addis Ababa, Ethiopia.

[20] Bo Pang, Lillian Lee. (2008). A Sentimental Education: Sentiment Analysis Using Subjectivity.

[21] Carmen Banea, Rada Mihalcea, Janyce Wiebe. (2010). Multilingual Subjectivity: Are More Languages Better?

[22] Caroline Laganas, Kendall McLeod, Elizabeth Lowe. (2017). Political Posts on Facebook: An Examination of Voting, Perceived Intelligence, and Motivations. Pepperdine Journal of Communication Research.

[23] Chirag Shah, Tayebeh Yazdani nia. (2011). Politics 2.0 with Facebook – Collecting and Analyzing Public Comments on Facebook for Studying Political Discourses. The Journal of Information Technology and Politics.

[24] Dey, A. (2016). Machine Learning Algorithms: A Review. International Journal of Computer Science and Information Technologies, 1174-1179.

[25] Faranak Ebrahimi Rashed, Neda Abdolvand. (2017). A Supervised Method for Constructing Sentiment Lexicon in the Persian Language. Journal of Computer & Robotics, 11-19.

[26] Janyce Wiebe and Ellen Riloff. (2005). Subjective and Objective Sentence Classifiers from the annotated text.

[27] Judith Hurwitz, D. K. (2018). Machine Learning. Hoboken: John Wiley & Sons, Inc.

[28] Kajaree Das, Rabi Narayan Behera. (2017). A Survey on Machine Learning: Concept, Algorithms, and Applications. International Journal of Innovative Research in Computer and Communication Engineering.

[29] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.

[30] Li, S. (2013). Sentiment Classification using Subjective and Objective views. International Journal of Computer Applications, 0975 – 8887.

[31] Luciano Barbosa, Junlan Feng. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data.

[32] Nick Yin Zhang, Celine Yunya Song. (2018). The Role of Social Media in Political Campaigns: A Sentiment and Engagement Analysis of Political News Feeds and Facebook Comments in Three Political Events in Hong Kong.

[33] Rada Mihalcea and Carmen Banea, Janyce Wiebe. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections.

[34] Rahul Rajput, Arun Kumar Solanki. (2016). Review of Sentimental Analysis Methods using Lexicon Based Approach. International Journal of Computer Science and Mobile Computing, 159 – 166.

[35] Rebecca F. Bruce, Janyce M. Wiebe. (n.d.). Recognizing Subjectivity: A Case Study of Manual Tagging. 1999.

[36] Rehling, Chapman & Hall/CRC. (n.d.). Handbook of Natural Language Processing: Machine Learning & Pattern Recognition 2nd Edition.

[37] Sara Douglas, Misa Maruyama, Bryan Semaan, Scott P. Robertson. (2014). Politics and Young Adults: The Effects of Facebook on Candidate Evaluation.

[38] Shoshana Hebshi, Etse Sikanku, Erin O'Gara. (2012). The role of online social networking in the 2008 Democratic presidential primary campaigns.

[39] Umman Tugba Gürsoy, D. B. (2017). Social Media Mining and Sentiment Analysis for Brand Management. An Online International Research Journal, 2311-3170.

[40] Wyngarden, K. E. (2012). new participation, new perspectives? young adults' political engagement using Facebook.

[41] Yogen, Y. (2017). To Study The Social Media Sentimental Analysis Using Facebook As Platform. d.y. Patil University.

[42] Younis, E. M. (2015). Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study. International Journal of Computer Applications, 0975 – 8887.

[43] Kumar, R. (2011). Research methodology a step-by-step guide for beginners.

[44] Martin Haselmayer and Marcelo Jenny. (2016). Sentiment analysis of political communication: combining a dictionary approach with crowd coding. Springerlink.

[45] Soroka, S., Young, L., & Balmas, M. (2015). Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. The ANNALS of the American Academy of Political and Social Science, 659(1), 108–121.

[46] Md Shad Akhtar, Palaash Sawant, Asif Ekbal, Jyoti Pawar, Pushpak Bhattacharyya. (2017).Measuring the Intensity of Emotions using Sentence Embeddings and Optimized Features.

[47] Zhang L., Liu B. (2014) Aspect and Entity Extraction for Opinion Mining. In: Chu W. (eds) Data Mining and Knowledge Discovery for Big Data. Studies in Big Data, vol 1. Springer, Berlin, Heidelberg, pp. 1-49.

[48] Rajkumar S. Jagdale, Vishal S. Shirsat. (2017). Review of Sentiment Lexicons.

[49] Yoonjung Choi and Janyce Wiebe (2014) +/-EffectWordNet: Senselevel Lexicon Acquisition for Opinion Inference, Proc. of EMNLP 2014.

[50] Zhang, F., Zhang, Z. & Lan, M. 2014. Ecnu: A Combination Method and Multiple Features for Aspect Extraction and Sentiment Polarity Classification. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).

[51] M. Lutfullaeva, M. Medvedeva, Candidate of Physic-Mathematical Sciences, Associate Professor, E. Komotskiy, K. Spasov, Ph.D. (2018). Optimization of Sentiment Analysis Methods for classifying text comments of bank customers.

[52] Zhon, Z. (2016). Topic-oriented Sentiment Analysis on Blogs and Microblogs. RMIT University.

[53] Yin Zhang · Rong Jin · Zhi-Hua Zhou. (2015).Understanding Bag-of-Words Model: A Statistical Framework.

[54] Wiebe, J.M., Bruce, R.F., O'Hara, Development and use of a gold-standard data set for subjectivity classifications, pp 246–253.

[55] Pang, B., Lee, (2013).Sentiment analysis using subjectivity summarization based on minimum cuts. pp 271.

[56] Lambov, D.,(2009) Dias, G., Noncheva., High-level features for learning subjective language across domains.

[57] Beata Bergman Klebanov, Jill Burstein, Nitin Madnani, Adam Faulkner, Joel Tetreault. (2013). Building Subjectivity Lexicon(s) From Scratch For Essay Data.