DEBRE BERHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF

Course:

Title

By:

Name and ID

Summited to:

8 May, 2021

Debre Berhan, Ethiopia

Table of Contents

CHAPTER ONE 1

1. INTRODUCTION 1

1.1. Back ground 1

1.1.1. About Mereja TV 1

1.2. Motivation .... Error! Bookmark not defined.

1.3. Statement of the problem 1

1.4. Objectives 2

1.4.1. General Objective 2

1.4.2. Specific objectives 2

1.5. Significance of the study 2

1.6. Scope and limitations .... Error! Bookmark not defined.

1.7. Methodology 3

CHAPTER TWO 4

2. LITERATURE REVIEW 4

2.1. Introduction 4

2.2. Types of opinions 5

2.3. Basic Components of Opinion Mining 7

2.4. Steps of sentiment analysis 7

2.5. Level of opinion mining .... Error! Bookmark not defined.

2.6. Approaches of sentiment analysis 8

2.6.1. Natural Language Processing (NLP) 8

2.6.2. Lexicon Based approach 8

2.6.3. Dictionary-Based approach 9

2.6.4. Corpus-Based approach 9

2.7. Related works ... 9

CHAPTER THREE ... 11

3. THE AMHARIC LANGUAGE ... 11

3.1. Introduction ... 11

3.2. Writing System of Amharic ... 11

3.3. Word Formation ... 12

3.4. Morphology ... 12

3.4.1. Nouns ... 12

3.4.2. Pronouns ... 13

3.4.3. Verbs ... 13

3.5. Amharic punctuations ... 14

CHAPTER FOUR ... 16

SENTIMENT ANALYSIS TECHNIQUES AND IMPLEMENTATION ... 16

4.1. Introduction ... 16

4.2. Data collection and preparation ... 16

4.3. Tokenization ... 16

4.4. Normalization ... 16

4.5. Stop words ... 17

4.5.1. Stemming ... 17

4.5.2. Lemmatization ... 17

4.6. Performance measurement ... 18

CHAPTER FIVE ... 19

5. CONCLUSION ... 19

References ... 20

CHAPTER ONE

1. INTRODUCTION

1.1. Back ground

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. Natural Language Processing (NLP) deals with actual text element processing. The text element is transformed into machine format by NLP. Artificial Intelligence (AI) uses information provided by the NLP and applies a lot of math's to determine whether something is positive or negative [1]. Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis research has been mainly carried out at three levels of granularity: document level, sentence level, and aspect level.

1.1.1. About Mereja TV

Mereja TV is a daily program on Mereja TV where journalists and guests to discuss events and affairs happening in Ethiopia recently

1.2. Statement of the problem

Currently in Ethiopia Mereja TV viewed by a lot of people since it comes with hot and recent issues in the country. Viewers and non-viewers give comments depends on specific issues and almost half of them notice the program positively. Some says it is good to discuss problems in public but others say the editor takes one side so it exacerbate crisis of the country.

Nowadays although there are a huge important data on the web it is difficult to use because of their unstructured format. Different organizations develop a system to accept an opinion of their user since it helps them to satisfy needs of customer and improve their product and services. Opinions can influence every human being behavior and we want to know what others people think about every decision we made. Every institution needs to understand what their employees and customers feel by using different social media like Facebook, Twitter, Telegram, etc. to collect opinions of user but the problem is hugeness of the information. That is why we need to conduct this research and develop the sentiment analysis system that extracts the most useful information out of the collected data.

This study aim to answer the following research questions:

1. How to organize the flow of an idea for comments given on social media?
2. What kinds of comments are recommended by the viewers/followers and non-viewers of the Mereja TV?
3. To what extent my models improve sentiment analysis at sentence level?

1.3. Objectives

1.3.1. General Objective

The aim of this study is to develop sentence level Amharic sentiment analysis system for Mereja TV program.

1.3.2. Specific objectives

To achieve the general objective, the following specific objectives will be addressed.

- To review different literature to understand sentiment analysis deeply.
- To collect comments and apply pre-processing to make collected sentence clean.
- To create a dictionary for the negative, positive and neutral positive sentiments.
- To build Amharic opinion mining system prototype.
- To measure performance of the system.

1.4. Significance of the study

Opinion Mining or Sentiment analysis involves building a system to explore user’s opinions made in blog posts, comments, reviews or tweets, about the product, policy or a topic. It aims to determine the attitude of a user about some topic [8]. The basic importance of sentiment analysis is ability of grouping polarity of a given text as positive, negative or neutral whether the opinion expressed at the document or sentence level. The main significance of this study is to identify the polarity of a given text whether it is positive, negative or neutral by gathering comments from Mereja TV websites about the program. Therefore this research can help founders of Mereja TV to improve their program by reducing their limitations and enables them to provide useful information that can teach the community. The result of the study can be used as resource and input for the development of future sentiment analysis system for Amharic language. Also it could answers peoples question in some manner. Generally this study will be useful for editors, founders, viewers of Mereja TV and for the overall society.

1.5. Methodology

Literature Review

Different materials from different sources that are related with sentiment analysis (published papers, journal articles and other materials) will be reviewed to understand sentiment analysis and its different approaches deeply.

Data Collection

Comments of viewers/non-viewers of Mereja TV will be gathered from social media using web scrapper to analyze their opinions. We will also conduct a closed ended questionnaire to collect data about what views and non-viewers feel about Mereja TV programs.

Tools

We will use python 3.9.2 to analyze the comments of Mereja TV viewers/non-viewers.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Introduction

Sentiments are feelings, thoughts; emotions of an individual for a particular event or topic e.g. love/hate, bad/good. Sentiment analysis is the study of the subjectivity (neutral vs. emotionally loaded) and polarity (positive vs. negative) of a text. It relies on sentiment lexicons, that is, large collections of words, each annotated with its own positive or negative orientation (i.e., prior polarity). The overall sentiment of a text is therefore computed upon the prior polarity of the contained words[9] . The sources of data for sentiment analysis (SA) are online social media, the users of which generate an ever-increasing amount of information. Thus, these types of data sources must be considered under the big data approach, given that additional issues must be dealt with to achieve efficient data storage, access, and processing, and to ensure the reliability of the obtained results. The problem of automatic sentiment analysis (SA) is a growing research topic. Although SA is an important area and already has a wide range of applications, it clearly is not a straightforward task and has many challenges related to natural language processing (NLP) [10]. The process of classifying a text written in a natural language as positive or negative opinion is difficult because of different factors. The most common factors are cultural factors, interpretation difference among individuals, reputation of reviewer, existence of spam and fake reviews etc. Opinion spamming has also become a major problem in sentiment analysis. People post fake opinions to promote or to discredit products, services, organizations etc. Such people are called opinion spammers. It is very important to detect spamming activities and ensure that the opinions on the Web are a trusted source of valuable information [11].

The major applications of Opinion mining and sentiment analysis are

1. Purchasing Product or Service: While purchasing a product or service, taking right decision is no longer a difficult task. By this technique, people can easily evaluate other's opinion and experience about any product or service and also he can easily compare the competing brands. Now people don't want to rely on external consultant. The Opinion mining and sentiment analysis extract people opinion form the huge collection of

unstructured content, the internet, and analyze it and then present to them in highly structured and understandable manner[12].

2. Recommendation Systems: the system identify which one should be recommended and which one is not recommended by classifying peoples thought or opinion as positive or negative.
3. Marketing research: we can use the result of sentiment analysis in marketing research. The recent attitude of community toward some new government policy and recent opinion of customer about some new product or service can be easily analyzed through sentiment analysis techniques.
4. Quality Improvement in Product or service: manufacturers can get a useful recommendation from their customer through sentiment analysis or opinion mining which can help them to improve their product or service.
5. Opinion spam detection: currently anyone can put whatever he/she want on the internet since it is available to everyone but this enables people to write spam content on the web to mislead others. Opinion mining can help us to know if the content is spam by classifying internet contents as spam and not spam.
6. Decision making: peoples want to know others opinion about anything they decide since others thought might help us to make a better decision. Opinion mining can proved well organized opinions of peoples which can help us while making a decision.
7. Policy making: policy makers can know citizen's opinion towards some policy by using sentiment analysis and can apply that information to create new citizen centered policy.
8. Detection of "flame": by using opinion mining we can control newsgroup and forums, social media, and blogs easily. It can also detect arrogant word used in emails or tweets on different web sources.

2.2. Types of opinions

We can classify opinions as regular and comparative. In addition to these two we can also classify as explicit and implicit depend on how they are expressed in text.

Regular opinion: is simply referred as opinion in literature and it can be grouped into two:

- Direct opinion: refers to an opinion expressed directly on an entity or an entity aspect, e.g., “The picture quality is great.”
- Indirect opinion: An indirect opinion is an opinion that is expressed indirectly on an entity or aspect of an entity based on its effects on some other entities. This sub-type often occurs in the medical domain. For example, the sentence “After injection of the drug, my joints felt worse” describes an undesirable effect of the drug on “my joints”, which indirectly gives a negative opinion or sentiment to the drug. In the case, the entity is the drug and the aspect is the effect on joints. Indirect opinions are often harder to deal with. For example, in the drug domain, one needs to know whether some desirable and undesirable state is before or after using the drug. For example, the sentence “Since my joints were painful, my doctor put me on this drug” does not express a sentiment or opinion on the drug because “painful joints” (which is negative) happened before using the drug [13].
- Comparative opinion: A comparative opinion expresses a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities. For example, the sentences, “Coke tastes better than Pepsi” and “Coke tastes the best” express two comparative opinions. It is usually expressed using the superlative or comparative form of an adjective or adverb, although not always [13].

Explicit opinion: An explicit opinion is a subjective statement that gives a regular or comparative opinion, e.g., “Coke tastes great,” and “Coke tastes better than Pepsi.” Explicit opinions are easier to detect and to classify than implicit opinions. Much of the current research has focused on explicit opinions. Relatively less work has been done on implicit opinions

Implicit (or implied) opinion: An implicit opinion is an objective statement that implies a regular or comparative opinion. Such an objective statement usually expresses a desirable or undesirable fact, e.g., “I bought the mattress a week ago, and a valley has formed,” and “The battery life of Nokia phones is longer than Samsung phones.” Explicit opinions are easier to detect and to classify than implicit opinions. Much of the current research has focused on explicit opinions.

2.3. Basic Components of Opinion Mining

There are mainly three components of Opinion Mining [14]

Opinion Holder: Opinion holder is the holder of a particular opinion; it may be a person or an organization that holds the opinion. In the case of blogs and reviews, opinion holders are those persons who write these reviews or blogs.

Opinion Object: Opinion object is an object on which the opinion holder is expressing the opinion.

Opinion Orientation: Opinion orientation of an opinion on an object determines whether the opinion of an opinion holder about an object is positive, negative or neutral.

2.4. Steps of sentiment analysis

There are some procedures we should follow for sentiment analysis: -

Goal setting: from the beginning the goal is to be set, including sentiment analysis goal and the scope for the content of the text.

Data collection: after the goal is set next thing is identifying data and collecting data from user generated contents existed in social media, forums or blogs.

Text processing: data on internet are unorganized and has a lot of noise so, it needs to be cleaned. Any irrelevant data should be identified and removed[15].

Sentiment detection: opinions which are expressed in sentence format analyzed and detected.

Presentation of output: when analysis is finished results are displayed on graphs like piechart, bar chart and line graph.

2.5. Approaches of sentiment analysis

Sentiment analysis played a vital role in different research area and there are a lot of methods to conduct research about sentiment analysis [17]. The most common ones are Natural Language Processing (NLP), Machine learning techniques (Naïve Bayes, Support Vector Machine, and k-Nearest Neighbor etc.), Lexicon-Based techniques (Dictionary-Based and Corpus-Based) etc.

2.5.1. Natural Language Processing (NLP)

Natural Language Processing (NLP) is part of Artificial intelligence which enables computers to analyze and understand human languages. It eases works of people and enables us to communicate with machine with our own language [18]. Natural language processing technique plays important role to get accurate sentiment analysis. NLP techniques like Bag of words, Hidden Markov model (HMM), part of speech (POS), co-reference resolution, N-gram algorithms, large sentiment lexicon acquisition and parsing techniques are used to express opinion for document level, sentences level and aspect level. Part of speech tagging (POS) is a process of assigning the part of speech categories to each and every word in the sentence. It is the basic stage in any NLP process and fundamental step of recognizing token's function in the given sentence. Different techniques have been applied for POS Tagging, with different level of performance[19] .

2.5.2. Lexicon Based approach

This technique is basically is unsupervised technique. It can apply Corpus-Based and Dictionary-Based approaches. There are step to be followed to apply this technique start with preprocessing then tokenization, sentiment score and classification [21]. Generally in lexicon-based approaches a piece of text message is represented as a bag of words. Then, sentiment value from the dictionary is assigned to all positive and negative words or phrases within the message. At last a combining function is applied to make the final prediction depend on the overall sentiment for the message.

The sentiment lexicon can be created in three different ways i.e. manual, dictionary-based and corpus based (where corpus is a huge collection of text in a particular language). The manual is reliable but labor intensive. The dictionary-based sentiment lexicon is automatically expanded from a seed of words. Mostly, Word Net and SentiWordNet are used for this expansion of words. The corpus-based sentiment lexicon is expanded from a seed of words by using a corpus.

2.5.3. Dictionary-Based approach

It is based on the usage of terms (seeds) that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is WordNet, which is used to develop a thesaurus called SentiWordNet [2].

2.5.4. Corpus-Based approach

The corpus-based approach have objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical or semantic techniques. Methods based on statistics are like Latent Semantic Analysis (LSA). Methods based on semantic such as the use of synonyms and antonyms or relationships from thesaurus like Word Net may also represent an interesting solution.

2.6. Related works

Currently in field of Natural Language Processing (NLP) Opinion mining the most popular and hot research area. Many researchers are has been studied in area of opinion mining especially for English language. Related researches are done for other languages like France, Chinese, and India etc. Although they are less in number there are also some researchers conducted on local languages like Amharic and Afaan Oromo. Different researchers used different kinds of approaches such as Lexicon-Based, Natural Language Processing Toolkit, machine learning and others.

Author	Objectives/goals	Methods/ techniques	Data Resource /domain	Result
(Tulu Tilahun, *2013)*	Build Aspect level opinion mining and summarization model for Amharic language.	Lexicon/dictionary based	Amharic blog	Two experiments are conducted. 1st 95.2% of precision and 26.1% of recall 2nd 78.1% of precision and 66.8% of recall
(Yohannes Mekonnen, 2018)	build Amharic sentiment analysis system for the program that polarity	Dictionary and rule-based	Kana TV Facebook page	85% of accuracy, 95% of precision, 89% of recall and 91% of f-score.
(Yodit Teshome, 2019)	Design and implement a Sentiment mining model for the Amharic language	Lexicon-based approach	Amharic blog	92% precision and 82% recall for positive class, 94% precision and 96% recall for negative and 6% recall 90 % recall and precision for neutral.
(Chilot Dessalew, 2019)	Build an Aspect level sentiment mining model for Amharic News.	Support vector machine and naïve bayes (NB)	Amhara Mass Media Facebook page.	SVM algorithm out performs than NB
(Negessa Wayessa*, Sadik Abas, 2020)	Classifying opinion Affan Oromo text into very negative, negative, neutral, positive and very positive.	Support vector machine and random forest algorithm (RF)	Ethiopian governmental broad casting cooperation, Oromia broad casting network (OBN) twitter page.	90% of accuracy by using SVM and 89% by using RF.

CHAPTER THREE

3. THE AMHARIC LANGUAGE

3.1. Introduction

Amharic or (አማርኛ), is Semitic Language of the Afro-Asiatic Language Group which is related to Hebrew, Arabic, and Syrian. In addition to this language, Afan Oromo (Oromiffa), Tigrinya, Somali and many other languages are spoken in Ethiopia. Amharic is national language of Ethiopia. It is also the official language of the media, government, and cross-communication. The origins of the language and its people are traced back to the 1st millennium B.C. It is rumored that they are the descendants of King Solomon and the Queen of Sheba. Immigrants from southwestern Arabia crossed the Red Sea into present-day Eritrea and mixed with the Cushitic population. Thus, new languages formed as a result of this union, Ge'ez (ግዕዝ). Ge'ez was the classical language of the Axum Empire of Northern Ethiopia. It existed between the 1st Century A.D. and the 6th Century A.D. When the power base of Ethiopia shifted from Axum to Amhara between the 10th Century A.D. and the 12th Century A.D., the use of the Amharic language spread its influence, hence becoming the national language [30].

3.2. Writing System of Amharic

Amharic is written from left to right with its own writing system structure is generally S+O+V (Subject + Object + Verb) or S+N+V (አበበ ጎበዝ ነዉ), or S+V (Subject + Verb e.g. አበበ መጣ). Amharic has its own writing system, a semi-syllabic system. There is no agreed translation of Amharic symbols to Roman characters (used in English). There are 33 consonant symbols that have seven variations. Variations are according to the vowel that is coupled with the consonant. About one quarter of Ethiopia's population is literate in the written form of Amharic. Sentence in Amharic can be a statement which is used to declare, explain an issue and the combination of phrases to create another phrase that can express a full idea about something is a sentence. From the grammatical structure point of view, Amharic sentence is a combination of noun phrase and verb phrase. The noun phrase comes first and then the verb phrase. Based on the number of phrases they contain sentences in Amharic are categorized under two basic categories simple sentences and complex sentence. Simple sentence only contains a single verb while complex sentence is constructed by combining more than one noun phrases and verb phrases. Since Amharic sentence formation follows its own structure, the syntax of the language also exhibited a unique structure

3.3. Word Formation

Word process of creating new words by compounding clipping and blending the existed words. The morphology of Amharic based on root and pattern system like other types of Semitic languages. Roots are the skeletal frames of words built with sequences of mostly three but also two to four consonants known as radicals. The patterns are the templates that determine the variation in length and repetition/ reduplication of the radicals and the insertion of vowels in the derived stem. With the exception of words borrowed from non-Semitic languages, the majority of Amharic words can be analyzed into roots-and-patterns. In addition to root-and pattern variation, Amharic utilizes multiple morphological processes such as affixation, reduplication, and compounding to form new words [31].

3.4. Morphology

In Amharic the most common usual clause order is noun + object + verb. Amharic is one of the most morphologically complex languages because of its Semitic characteristics. Amharic nouns and adjectives are marked for any combination of number, definiteness, gender and case. Amharic verb inflections and derivations are even more complex than those of nouns and adjectives. Several verbs in surface forms are derived from a single verbal stem, and several stems in turn are derived from a single verbal root. For example, from the verbal root sbr (to break), we can derive verbal stems such as sabr, sabar, sabr, sababr, tasababr, etc. From each of these verbal stems, we can derive many verbs in their surface forms. [32].

3.4.1. Nouns

Noun is a word that names something, such as a person, place, thing, or idea. Nouns may represent case, gender, number, definiteness, and direct object status by affixes prefixes and suffixes, predominately suffixes. Amharic nouns may have a masculine or feminine gender. Suffixes are added to denote a masculine or feminine noun gender. Some nouns may have both masculine and feminine gender, while other nouns may only have one gender. The feminine gender is used to indicate female as well as smallness. Plurals are indicated by the suffix. Affixes are added in the following order: gender, number, definiteness, case, and direct object status.

3.4.2. Pronouns

Pronoun is a word that can replace a noun or noun phrase. Sentences with no emphasized element do not require independent pronouns. The verb denotes the person, number, and gender. Object pronoun suffixes are affixed to verbs and indicate person, number, and gender of the object of a verb. Possessive suffixes are affixed to nouns to indicate possession.

Personal pronouns in Amharic language are ine (I), igna (we), ante(you {male, singular}), anchi (you {female, singular}), inante (you {gender neuter, plural}), irswo (you {gender neuter, polite}), irsu/ isu (he or it), irswa/ iswa (she), irsachaw/ isachaw (he or she {polite}), inarsu/ inasu (they). In addition, possessive and objective forms could be derived through prefixing and suffixing with ye- and -ne, respectively. For example, yeine (my) and inen (me) are derived from ine (I) through such affixation process. Amharic personal pronouns can exist in sentences in two forms: independent and hidden.

Independent pronouns are those which appear independently in the text whereas hidden pronouns are embedded in other words. For example, the subjective pronoun isu (he or it) and objective pronoun inen (me) do not appear independently but are embedded in the word yisebranal ([he/it] will break me). Moreover, both independent and hidden personal pronouns are subject to complex affixation processes in Amharic.

3.4.3. Verbs

Verbs are derived from roots and affixes to inflect person, number, gender, aspect, mood, voice, and polarity are added. Verbs agree with their subjects. Verb agreement with objects is optional. Verbs are placed at the end of the sentence. For Amharic, like most other languages, verbs have the most complex morphology. In addition to the affixation, reduplication, and compounding common to other languages, in Amharic, as in other Semitic languages, verb stems consist of a root + vowels + template merger. This non-concatenate process makes morphological analysis more complex than in languages whose morphology is characterized by simple affixation. The affixes also contribute to the complexity. Verbs can take up to four prefixes and up to five suffixes, and the affixes have an intricate set of co-occurrence rules. Grammatical features on Amharic verbs are not only shown using the affixes. The intercalation pattern of the consonants and the vowels that make up the verb stem will also be used to determine various grammatical features. For example, the following two verbs have the same prefixes and suffixes and the same root while the pattern in which the consonants and the vowels intercalate is different, resulting in different grammatical information [33].

For example

- **sebr**-alehu (እሰብራለሁ)→ 1s pers. sing. simplex imperfective Gloss: 'I will break'
- **seber**-alehu (እሰበራለሁ)→ 1stpers.sing.passive imperfective Gloss: 'I will be broken'

In this next case, the difference in grammatical feature is due to the affixes rather than the internal root template structure of the word.

- te-**seber**-ku (ተሰበርኩ)→ 1st pers. sing. passive perfective Gloss: 'I was/have been broken'
- **seber**-ku (ሰበርኩ)→ 1st pers. sing. simplex perfective Gloss: 'I broke'

Adjectives are predominately derived from nouns, verbs, and other parts of speech [34]. Adjectives are words or constructions used to qualify nouns. Adjectives in Amharic can be formed in several ways: they can be based on nominal patterns, or derived from nouns, verbs and other parts of speech. Adjectives can be nominalized by way of suffixing the nominal article. Amharic has few primary adjectives. Some examples are *degg* 'kind, *dida* 'mute, dumb, silent', *bicha* 'yellow.

3.5. Amharic punctuations

The analysis of Amharic texts exposes that different Amharic Punctuation marks are used for different purposes. There are many punctuation marks of which only a few of them are commonly used. For example the sentence separator for Amharic text writing is four dots arranged in a square sequence as (።) and are referred as "አራት ነጥብ/arat neteb". The comma which is symbolized as (፣) and its equivalence punctuation in Amharic is "ነጠላ ሰረዝ/ netela serez” used to separate lists. The equivalence between the semicolon, which is symbolized as (,) which is used to separate phrasal lists. Compound sentences are referred as “ድርብ ሰረዝ/ dereb serez” which is denoted by the symbol (፤). Punctuation marks like the question mark (?) and the exclamation marks (!) are borrowed from the English language and used in Amharic language in similar way they are used in other foreign languages.

CHAPTER FOUR

SENTIMENT ANALYSIS TECHNIQUES AND IMPLEMENTATION

4.1. Introduction

In this study we have process the data before we compare each lexicon with positive and negative dictionaries. Processing include different tasks like splitting, stemming or lemmatizing, and extracting information from it. Language in its original form cannot be accurately processed by a machine, so you need to process the language to make it easier for the machine to understand. The first part of making sense of the data is through a process called *tokenization*, or splitting strings into smaller parts called *tokens*. After tokenization next important task is normalization. Normalization helps group together words with the same meaning but different forms and *stemming* and *lemmatization* are the two popular techniques of normalization.

4.2. Data collection and preparation

We collect data that are comments of viewer or non-viewers from Mereja TV face book, website and telegram. I collected 800 comments manually. Out of it 353 positive, 159 negative and other 288 were neutral.

4.3. Tokenization

Tokenization is the first step in text analytics. The process of breaking down a text paragraph into smaller chunks such as words or sentence is called Tokenization. Token is a single entity that is building blocks for sentence or paragraph.

Code

4.4. Normalization

Normalization generally refers to a series of related tasks meant to put all text on a level playing field: converting all text to the same case (upper or lower), removing punctuation, converting numbers to their word equivalents, and so on. Normalization puts all words on equal footing, and allows processing to proceed uniformly.

To remove punctuations and unnecessary characters code

4.5. Stop words

Stop words considered as noise in the text. Text may contain stop words such ,…………etc. In NLTK for removing stopwords

code

4.5.1. Stemming

Stemming is a process of linguistic normalization, which reduces words to their word root word or chops off the derivational affixes. For example, connection, connected, connecting word reduce to a common word "connect".

4.5.2. Lemmatization

Lemmatization reduces words to their base word, which is linguistically correct lemma. It transforms root word with the use of vocabulary and morphological analysis. Lemmatization is usually more sophisticated than stemming. Stemmer works on an individual word without knowledge of the context. For example, the word "better" has "good" as its lemma. This thing will miss by stemming because it requires a dictionary look-up.

```
from nltk.stem import WordNetLemmatizer

from nltk.tokenize import word_tokenize

lemmatizer=WordNetLemmatizer()

input_str='C: /Users/SolomonDubale/Desktop/sent/oneandtwo.txt'

input_str=word_tokenize(input_str)

for word in input_str:

    print(lemmatizer.lemmatize(word))
```

4.6. Performance measurement

Accuracy, Precision, Recall, F-Score

We use the following performance metrics to evaluate our experiment result and our classifier accuracy, precision, recall, f-score.

Accuracy: One of the more obvious metrics, it is the measure of all the correctly identified cases.

Accuracy = TP + TN / (TP + FP+ TN+FN)

Precision: It is implied as the measure of the correctly identified positive cases from all the predicted positive cases.

Precision = TP / (TP + FP)

Recall: It is the measure of the correctly identified positive cases from all the actual positive cases.

Recall = TP / (TP + FN)

F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

F-score = 2*((Precision* Recall) / (Precision+ Recall))

Table 2: Experimental result of the system

Label	Precision	Recall	F1-score	Accuracy
-1	0.80	0.29	0.42	0.7701863354037267
0	0.73	0.84	0.78	
1	0.80	0.89	0.84	
Average	0.78	0.67	0.68	0.77

CHAPTER FIVE

5. CONCLUSION

Currently analysis of sentiments and opinions has spread across many fields such as Consumer information, Marketing, books, application, websites, and Social. It is becomes a hot area in decision-making. In this study we tried to make sentence level sentiment analysis in Amharic language. We collect 800 comments from face book and telegram that are opinions of viewers/non-viewers of Mereja TV. The collected data processed (punctuation removal) and normalized before we build the model. We use linear regression to build and classify and it achieved 0.77 accuracy, precision 0.78, recall 0.67 and f1-score 0.68. The most difficulty in this study was there is no prepared corpus that can be used for sentiment analysis purpose. So some sentences may not be correctly understood by the machine

Group contribution

Stud1	Sudent id Title selection, full documentation, data collection, punctuations and special character removal, splitting to sentence
2	Stud id Stop word removal code
3	Stud id Tokenization code
4	Stud id Stemming code

References

[1] O. Bharti and M. M. Malhotra, “International Journal of Computer Science and Mobile Computing SENTIMENT ANALYSIS,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 6, pp. 625–633, 2016, [Online]. Available: www.ijcsmc.com.

[2] V. A. and S. S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” *Int. J. Comput. Appl.*, vol. 139, no. 11, pp. 5–15, 2016, doi: 10.5120/ijca2016908625.

[3] A. Alsaeedi and M. Z. Khan, “A study on sentiment analysis techniques of Twitter data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/ijacsa.2019.0100248.

[4] D. M. E. D. M. Hussein, “A survey on sentiment analysis challenges,” *J. King Saud Univ. - Eng. Sci.*, vol. 30, no. 4, pp. 330–338, 2018, doi: 10.1016/j.jksues.2016.04.002.

[5] R. Ren *et al.*, *Introduction & The Problem of Sentiment Analysis*, vol. 933, no. November 2017. 2015.

[6] S. Kolkur, G. Dantal, and R. Mahe, “Study of Different Levels for Sentiment Analysis,” *Int. J. Curr. Eng. Technol.*, vol. 55, no. 22, pp. 2277–4106, 2015, [Online]. Available: http://inpressco.com/category/ijcet.

[7] C. Rica, “SENTIMENT ANALYSIS TO INFORM SDGs : Immigration in Colombia and.”

[8] C. S. and S. C., “Opinion Mining and Sentiment Classification: a Survey,” *ICTACT J. Soft Comput.*, vol. 03, no. 01, pp. 420–427, 2012, doi: 10.21917/ijsc.2012.0065.

[9] F. Calefato, F. Lanubile, F. Maiorano, and N. Novielli, “Sentiment Polarity Detection for Software Development,” *Empir. Softw. Eng.*, vol. 23, no. 3, pp. 1352–1382, 2018, doi: 10.1007/s10664-017-9546-9.

[10] N. C. Dang, M. N. Moreno-García, and F. de la Prieta, “Sentiment analysis based on deep

learning: A comparative study," *arXiv*, 2020.

[11] D. Kulkarni and P. S. F. Rodd, "A Survey on Opinion Mining problem and levels of Analysis," pp. 12390–12394, 2015, doi: 10.15680/IJIRSET.2015.0412134.

[12] H. R. P, "Opinion Mining and Sentiment Analysis Challenges and Applications," *Int. J. Appl. or Innov. Eng. Manag.*, vol. 3, no. 5, pp. 401–403, 2014.

[13] Y. Lin, X. Wang, and A. Zhou, "Opinion spam detection," *Opin. Anal. Online Rev.*, no. May, pp. 79–94, 2016, doi: 10.1142/9789813100459_0007.

[14] A. Dhokrat, C. N. Mahender, and S. Khillare, "Review on Techniques and Tools used for Opining Mining," *Int. J. Comput. Appl. Technol. Res.*, vol. 4, no. 6, pp. 419–424, 2015, doi: 10.7753/ijcatr0406.1001.

[15] M. Godsay, "The Process of Sentiment Analysis: A Study," *Int. J. Comput. Appl.*, vol. 126, no. 7, pp. 26–30, 2015, doi: 10.5120/ijca2015906091.

[16] G. Beigi, X. Hu, R. Maciejewski, and H. Liu, "An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief BT - Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence," pp. 313–340, 2016, [Online]. Available: https://doi.org/10.1007/978-3-319-30319-2_13.

[17] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," *Procedia Comput. Sci.*, vol. 87, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.

[18] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *arXiv*, no. Figure 1, 2017.

[19] Vishaal Jatav, R. Teja, S. Bharadwaj, and V. Srinivasan, "Improving part-of-speech tagging for NLP pipelines," *arXiv*, 2017.

[20] W. M. Soon, D. C. Y. Lim, and H. T. Ng, "A machine learning approach to coreference resolution of noun phrases," *Comput. Linguist.*, vol. 27, no. 4, pp. 521–544, 2001, doi:

10.1162/089120101753342653.

[21] N. Nandal, J. Pruthi, and A. Choudhary, “Aspect based sentiment analysis approaches with mining of reviews: A comparative study,” *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 95–99, 2019.

[22] M. R. Hasan, M. Maliha, and M. Arifuzzaman, “Sentiment Analysis with NLP on Twitter Data,” *5th Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2019*, pp. 1–4, 2019, doi: 10.1109/IC4ME247184.2019.9036670.

[23] S. Taj, B. B. Shaikh, and A. Fatemah Meghji, “Sentiment analysis of news articles: A lexicon based approach,” *2019 2nd Int. Conf. Comput. Math. Eng. Technol. iCoMET 2019*, pp. 1–5, 2019, doi: 10.1109/ICOMET.2019.8673428.

[24] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, “COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model,” *IEEE Access*, vol. 8, pp. 138162–138169, 2020, doi: 10.1109/ACCESS.2020.3012595.

[25] TULU TILAHUN, “Opinion Mining From Amharic Blog,” *Unpubl. Masters Thesis,Addis Ababa Univ.*, 2013.

[26] N. Wayessa and S. Abas, “Multi-Class Sentiment Analysis from Afaan Oromo Text Based On Supervised Machine Learning Approaches,” vol. 7, no. 7, pp. 10–18, 2020.

[27] Y. Teshome, “Debre berhan university college of computing,” no. June, 2019.

[28] D. Chilot, “Public sentiment analysis for Amharic news,” 2019.

[29] Y. Mekonnen, “DICTIONARY AND RULE BASED (HYBRID) APPROACH OF AMHARIC SENTIMENT ANALYSIS FOR KANA TV,” 2020.

[30] A. The *et al.*, “Visit our web site for up to date information and new products.”

[31] B. T. Ayalew, “THE SUBMORPHEMIC STRUCTURE OF AMHARIC: TOWARD A PHONOSEMANTIC ANALYSIS,” 2013.

[32] T. Dawit and Y. Assabie, “Amharic Anaphora Resolution Using Knowledge-Poor Approach,” pp. 278–289, 2014.

[33] W. Mulugeta, M. Gasser, and B. Yimam, “Incremental learning of affix segmentation,” *24th Int. Conf. Comput. Linguist. - Proc. COLING 2012 Tech. Pap.*, vol. 1, no. December 2012, pp. 1901–1914, 2012.

[34] A. Wimsatt and R. Wynn, “Amharic Language and Culture Manual,” 2011.