International Journal of Intelligent Information Systems
2021; 10(6): 125-138
http://www.sciencepublishinggroup.com/j/ijiis
doi: 10.11648/j.ijiis.20211006.14
ISSN: 2328-7675 (Print); ISSN: 2328-7683 (Online)

Amharic Text Summarization for News Items Posted on Social Media

Abaynew Guadie*, Debela Tesfaye, Teferi Kebebew

Faculty of Computing, Jimma Institute of Technology, Jimma University, Jimma, Ethiopia

Email address:
abaynewhu@yahoo.com (A. Guadie), abaynew.guadie@ju.edu.et (A. Guadie)
*Corresponding author

To cite this article:
Abaynew Guadie, Debela Tesfaye, Teferi Kebebew. Amharic Text Summarization for News Items Posted on Social Media. *International Journal of Intelligent Information Systems*. Vol. 10, No. 6, 2021, pp. 125-138. doi: 10.11648/j.ijiis.20211006.14

Received: September 25, 2021; **Accepted**: October 19, 2021; **Published**: December 24, 2021

Abstract: This paper introduces Amharic Text Summarization for News Items posted on social media, to summarize the news items posted Amharic texts over a time posted documents from social media on Twitter and Facebook; The main problems of the social media posted texts are that most people would probably read their posted in Amharic texts with duplicate posted documents. However, to find the information the user is looking for she or he will have to find summary posted texts and read important portions of posts as Amharic documents to extract desired information on social media. Summarization is dealing with information overload presenting and posted with a text document for the current time representation of the posted documents to summarize. Our proposed approach has three main components: First, calculate the similarity between each posted document within the two pair of sentences. Second, clustering based on the similarity results of the documents to group them by using Kmeans algorithm. Third, summarizing the clustered posted document individually using TF-IDF algorithms that involve finding statistical ways for the frequent terms to rank the documents. We applied the summarization technique is an extractive summarization approach that is assigned an extract the sentences with highest ranked sentences in the posted documents to form the summaries and the size of the summary can be identified by the user. In the experiment one the highest F-measure score is 87.07% for extraction rate at 30%, in the clustered group of protests posts. The second experiment the highest F-measure score is 84% for extraction rate at 30%, in droughts post groups. In the third experiment the highest F-measure score is 91.37% for extraction rate at 30%, in the sports post groups and also the fourth experiments the highest F-measure score is 93.52% for extraction rate at 30% to generate the summary post texts. If the system to generate the size of summary is increased, the extraction rate also increased in posted texts. For this the evaluation system shown that a very good results to summaries the posted texts on social media.

Keywords: Amharic, Similarity, Text Summarization, Clustering, TF-IDF Algorithm, Social Media, Facebook, Twitter

1. Introduction

Amharic language is one of the main African languages and also it is the working language of the Federal Government of Ethiopia and widely spoken throughout the country [1, 2, 14]. It is an Afro-Asiatic language belonging to the Semitic group of its own unique alphabets [3]. Its speaker to post Amharic texts on Twitter and Facebook in the social network at this time, which is increasing volumes of posting data are available on the social media, which is observed on the growing online posts, websites, and digital storages in the language. These Amharic text documents are available digitally and the amount is highly increasing every day from a user posting to the social media on Facebook and Twitter. The task of developing for easy retrieval of relevant information is specially challenging on Amharic texts because there are only few recent and uncoordinated efforts of automation and language processing [2]. Currently, with the enhancement in the most people to use and posted and reposted many Amharic text documents in social networks, the volume of data one has to deal with has increased rapidly on Twitter and Facebook for the readers. Text summarization aims to investigate the summarization of user posts over time for the documents from the social media within the time period of each post as a stream posted documents on the social media. Therefore, documents shall be iterated over in news items order for the news items posted on Facebook and also its aims to create and evaluate the news items of posted texts summarization systems for the news post on the social media. It services such as Twitter and Face book generate to rapid access phenomenal volume of content for most real-world posts on a daily source [4].

In this paper, we address the problem of the summarizing user posts of interest in a human reader by extracting the most representative posts of the irrelevant Amharic tweet for the post news that could be summarizing without duplicate posted Amharic text documents [5]. Readers tend to focus on condensing the information provided to them by taking summarize posted texts and ignoring the details if they are not interested the posted texts. News posted Amharic text documents include summarizing posted political protests, natural disasters, bombing, earthquake [6], shooting, storm and social media for accident things to happen collecting information from a large set of documents into a single news posted document or generate a summary of user posts from Twitter and Facebook. A good way to get up to date, monthly and yearly information Amharic texts will be to get a stream of sentence length in posted documents about the situation as it develops [7].

2. Related Works

2.1. Text Summarization for News Items Tasks

The work of the Eidheim [8] to propose the temporal summarization task, which is to inform readers of important novel information about a particular topic. For that the DUC and TAC Update Summarization tasks were designed that as a single pass collection process, processing all the documents at once, while in that year TREC temporal summarization tracks the task designed requires generation of continuous and immediate updates. As with the earlier work, sentences are the unit of selection. The author experimented with several different methods of sentence selections based on language models, TF-IDF and word counting. One of primary issues the authors wanted to address the fact that earlier methods were based on the assumption that all the evaluated documents were relevant. The problems covered by his paper were very similar to the temporal summarization task and it shows that great care necessity be taken when selecting documents to ensure that the results are good. The author had been done only English news articles was considered when generating updates and the main reason for this is that the chunks and queries are all English, allowing other languages is a potential source of noise, but it is unlikely to contain any interesting information. It is also likely that a user monitoring an update stream would like the stream to be in his or her language, making it a reasonable decision to only look at documents in a specific language. For the sake of keeping the collection of documents small, non-English documents are removed by the preprocessor of the tweets [9].

2.2. Temporal Patterns Text Documents

The work of those Ks. R. Premlatha and T. V Geetha [10], for extracting the temporal patterns that is from texts requires to the expression for handling the types of temporal term categories were explicitly, implicitly and vaguely conveyed to the temporal information for the given document. The authors performed by using finite state automata in the natural language expression are converted into a calendar-based timeline. The authors give the temporal expressions the first the temporal explicitly are examples like on the 18th of March 2009, in Nov 2010, 12 Jul 2011 etc. We applied to use this temporal in the news items of posting term, text summarization explicit expression to identify the posted documents by using date, month and year format to group by monthly for the summary posted document in social media posts. Finally, the vague temporal expressions [11] is like July, after several weeks before etc. In the work of this article, for the topic-based clustering for the calendar model which go to enable the temporal intervals posted documents was carried out their work for the papers. Those authors to preprocess the used of the clustering such as stop word removal; stemming was performed the document represented using TF-IDF representation which is used for clustering for the similarity.

2.3. Text Summarization for Amharic Texts

In this author of Melse [12], had described and proposed to his papers for Amharic texts the two generic texts summarization approaches. For the first technique, to put the topics Latent Semantic Analysis (LSA) and the second technique mixes the graph-based ranking among the Latent Semantic Analysis algorithms to identify the topics of a document were used to select the semantically the main sentences for the summary generation. The author used the algorithm was Latent Semantic Analysis and graph based ranking algorithms to explore his work. His work evaluates to propose the performance of the summarization approaches and his prototype Amharic news text summarization system. For this author, to evaluated the summaries systems with manual summaries were generated by six independent human evaluators to be taken the evaluator for the experiments. He prepared the dataset corpus used for evaluating the summarization system was 50 Amharic news's from Ethiopian news reporter the items were in the range of 17-44 sentences. His results to evaluate by comparing the system summaries of their corresponding manual summaries of the results.

The work of Addis [13], had proposed an open source of customizing by Amharic texts automatic summarization using an open text summarizer tool of two ways of execution the two experiments, the first one experiment is done without changing the code of the tool and the second is done for the changing the Porter stemmed tool for the Amharic stemmed algorithm. His work uses the frequency of terms to determine the relative importance of a sentence in a text. His papers evaluation of the experiments was producing 90 news articles and to test its performance for the summaries for each rate at 10%, 20% and 30% extraction rates for the results. This author to evaluate the system was evaluated using the subjective and the objective evaluation.

According to the work of Eyob [14], had proposed to topic-based Amharic text summarization to investigate the six algorithms to explore as the use of terms by concept matrix to implement for his thesis. His algorithms to take two common steps, the first step, to identified the keywords for the documents were used to select the term of use of the concept matrix to find the document. The second step, for the sentences to find the best keywords contains that were selected for presence in the summary. For this author to take the experiment with news articles in Amharic texts to explore the algorithms for selected the first sentence of the document for inclusion in the summary of the Amharic news texts. His evaluated in the paper by using for the precision/recall for summaries of 20%, 25% and 30% extraction rates to evaluate for the news articles. The author of comparing his system with the previous methods of developed for other languages based on topic modeling approaches summarization that had been used his Amharic data set of the papers.

Since the work of these authors of Kifle and Martha Y [15], had investigated for Amharic texts to an automatic single document summarization task for the graph based automatic Amharic text summarizer were proposed. They were to work the generic and domain independent graph-based model could successfully to make extracts from Amharic texts for their papers. These authors use the two graphs based ranking algorithms were introduced, their thesis used for PageRank and HITS that we're using the two sentence centrality measures and sum the relation between the sentences in a text for a graph that show for the results from their experiments. They worked to prepare the data sets for the experiments shown 30 news articles used on economics, politics, society, and sports were conducted for Amharic news articles from collected for Ethiopian reporter news web sites and Addis Admas to test its performance for the summaries for Amharic news articles [15].

2.4. Single and Multi-document Text Summarization for Posting Texts

In [16], multi-document summarization system is developed for the web context. Information may have to be extracted from many different articles and pieced together to form a comprehensive and coherent summary. One major difference between single document summarization and multi document summarization is the potential redundancy that comes from using many source texts. The solution presented in [16], is based on clustering the important sentences picked out from the various source texts and using only a representative sentence from each cluster. Multi-document summary when a summary of one topic is prepared for many different documents. For these reasons, multi-document summarization is much less developed than its single-document and various methods have been proposed to identify cross-document overlaps from different researchers. Summons (which is a paper issued by informing a person that a complaint has been filed against it), or order, a system that covers most aspects of multi-document tweet summarization [17], takes an information retrieval approach. Assuming that all input documents is parsed into templates (whose standardization makes comparison easier), summons clusters the templates according to their contents, and then applies rules to extract tweets items of major imports. In contrast, the problem of organizing information on multi document summarization so that the generated summary is coherent has received relatively little attention [18]. Multi-document summarization [19] poses interesting challenges to single post documents and also the information overloads faced by today's society poses great challenges to researchers that want to find a relevant piece of information. Automatic summarization is a field of computational linguistics which can help humans to deal with this information overload by automatically extracting the idea of documents. An important the study show that for the newspaper article type, even some very simple procedures provide essentially perfect results [20]. According to the work of these author Inouyes, and J. K. Kalita [21] their paper described that an algorithm for summarizing micro blog documents on Twitter. Firstly, they presented algorithms that produce single-document summaries, but later extend them to produce summaries containing multiple documents and also, they evaluated the generated summaries by comparing them to both manually produced summaries and, for the multiple post summaries, to the summary results of some of the leading traditional summarization systems. The author discussed to overflowing with information on twitter, they taught that just being able to search for a text tweet and receive user tweet the most recent posts whose text match the keywords are not enough. Summarization can represent the tweet document with a short piece of text covering the main topics for the user posts, and help users select through the Internet, the most relevant document, and filter out redundant information on the social media to find the relevant tweets [22-24]. So, the author to describe the document summarization has become one of the most important research topics in the natural language processing and information retrieval communities.

2.5. For the Summarize the User Posts on the Twitter and Facebook for Evaluation Techniques for Previous Studies

The most existing evaluations of summarization systems are fundamentally the typical; the evaluators create a set of summaries, one for each test text, and then compare the Summarizer's output, measuring the content overlap that often by sentence recall and precision. To simplify evaluating extracts [25, 26], independently developed an automated method to create extracts corresponding to abstracts for the Twitter. The two evaluation methods of the text summarization such as intrinsic and extrinsic evaluation methods. An intrinsic method is the evaluation of the system for summaries according to have the evaluators to take the ratio of some scale of interval the extraction rate (readability;

Informativeness; facility; coverage and others). It was prepared by creating the human or ideal summaries of the given input text document for comparing the summary of the summarizing system and the human summary of the evaluators. The evaluator used for measuring the average scores found from each evaluation criterion for the given document. The others evaluation method is extrinsic evaluation for measuring the acceptability and also the efficiencies of the automatic summaries of documents to achieve the tasks and easily to motivate, this evaluation method the main difficulty to ensure that are applied to correlate or link with the task of the performance efficiency for the summary documents. Some of the intrinsic evaluation methods to evaluate the summaries as follows.

2.5.1. Coherence and Structured for Summary

This is one of the evaluators to measure the extraction rate-based methods to use the flow of the information to be structured by using the cut and paste processes of the documents on phrases, sentences or paragraphs that produces the result in a summary extracted for the documents which results in coherence to get for the documents.

2.5.2. Informativeness for Summary

This is one of the summaries of comparing the system summary generated the documents within the input texts for summarized the key ideas about manual summary that are included in the automatically summaries and the summary information are presented in the input texts.

2.5.3. Sentence Precision and Recall

It is a standard measures for information retrieval in terms of sentences in the given document. So, precision of measures by using how many of the sentences in the system summary generated and also in the reference human summary are producing the results. An others hand, recall for the sentences to measure for how many of the sentences in the reference ideal summary are extracted from the system summary of the documents [27]. Twitter observatory that allows observing, searching, analyzing and presenting social media is introduced as a part of the research and illustrative examples of using his proposed pipeline show how the Twitter user interaction with the social media data [28]. According to this author Sharifi et al. [29] described that his algorithms process collections of short posts as specific topics on social media the well-known site called Twitter and create short summaries of those posts as the Twitter. The goal of his research is to produce summaries. His paper evaluated the summaries produced by the summarizing algorithms, compare them with human produced summaries and obtain excellent results.

Since this authors P. Meladianos [30] to deal with the task of sub-posts detection in evolving twitter posts using posts collected from the Twitter stream and also by representing a sequence of successive tweets in a short time interval as a weighted graph of words, they were able to identify the key moments sub-posts that combine a post using the concept of graph degeneracy on the twitter. They were selected a tweet to best describe each sub-post using a simple yet effective heuristic (which is used for experimental).

2.5.4. The Co-selection Summary Evaluations

This method was that to discuss here are the simplest of all evaluation measures co-selection sentences and this simplicity creates at a value score to check the intersection with the system and human summary of sentences [14]. For that the main problem is the difficulty in accounting for variations in what humans consider ideal summary sentences in the documents. The summary evaluation for using the co-selection measure were to take from information retrieval evaluations techniques and for describing the formula for recall (R), precision (P), and F-measure given to calculate the system and the human selected the fraction of sentences that the system has chosen from the total of sentences found in the ideal summary as follows to formula.

$$R = \frac{|System\ and\ Human\ \ choice\ Overlap|}{|Senetnces\ chose\ by\ human|} \tag{1}$$

Precision (P) measures the fraction of system summaries that are correctly chosen.

$$P = \frac{|System\ and\ Human\ \ choice\ Overlap|}{|Senetnces\ chose\ by\ system|} \tag{2}$$

F-score (F) is the harmonic mean of recall and precision.

$$F = \frac{2*P*R}{p+R} \tag{3}$$

There were many works being done on the area of summarization [31] on social media by Twitter in western world language for news items summarization. But most of the work being done for major technology languages like, English, Chinese, German and French [16].

Due to those authors [32] that described the shortness of tweets and also TwitIE makes the assumption that each tweet is written in only one language. The choice of languages used for categorization is specified through the arrangement file, complete as an initialization parameter. The authors take three tweets one English, one German, and one French. TwitIE TextCat was used to allocate automatically the language feature of the tweet text (denoted by the Tweet annotation). Those give a collection of tweets in a new language; it is possible to train TwitIE TextCat to support that new language as well and also that is done by using the fingerprint generation (which is patterned to generate) included in the language identification plugin. It builds a corpus of documents and reliable tweet language identification allows them to only process those tweets written in English with the TwitIE English and named entity recognizer [32].

2.6. The Use of Social Media in the Summarization Posted Documents

Social media was defined by [5] as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user posted generate content. The use of social media has exploded to recent years and the high availability of such information is then used by many researchers. In this author to work [33], the author of described his goals is to design novel features “identifying valuable information on Twitter during natural disasters” that can be used as input to machine learning classifiers in order to automatically and accurately identifying informational tweets from the rest in a timely fashion. His approaches used for machine learning algorithms that would be discussed for URL extraction are used extensively in tweets to link to external sources that could not ordinarily acceptable to the long-restricted structure of a tweet. However, URLs found in tweets, are shortened in order to accommodate for the length restriction. Inherently, a few features can be extracted from the URL itself, considering that each shortened URL has a base domain and a randomly generated code appended to that (i.e., https: “t.co/ [code]”), which, when clicked on, will redirect to an actual web page [33].

2.7. Sentence Extraction by TF-IDF and Position Weighting

In this article [34] to describe the sentence extraction via using TF-IDF, they would be discussed the posted Japanese Newspaper in order to create a summary and his system is implemented with the sentence extraction approach and weighting strategy to mine from a number of documents. They created an experimental system for the Japanese Summarization to compute the importance value of each sentence based on Japanese newspaper terms. He's used the important sentences whose sum of characters exceed the restricted character amount are eliminated and the remaining sentences are then sorted as they appeared in the original document. The author of summarization to used single and multi-document summarization to implement the different evaluation between a long summary and its short summary is more remarkable. In this article a frequent term texts summarization algorithm is designed and implemented in Java and to designed algorithm is implemented using open-source technologies like Java, Porter stemmer etc. and verified over the standard text mining corpus. Japanese summarization to compute the importance values for each sentence based on Japanese newspaper term [35].

2.8. Hybrid TF-IDF Documents

TF-IDF stands for term frequency-inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus is occurring, it is the most common weighting method used to describe documents in the Vector Space Model (VSM), particularly on IR problems. In a hybrid TF-IDF algorithm development and the idea of the algorithm is to assign each word, sentences within a document a weight that reflects the words, sentences the most important to the document. The sentences are ordered by their weight from which the top sentences with the most weight is chosen as the summary. In order to avoid redundancy or duplicate posted words the algorithm selects sentences and tokenize, the next sentences or terms and checks it to make sure that it does not have a similarity with a given threshold with any of the other previously selected because the top most weighted tweets may be very similar [36]. The tf*idf of term t in document d is calculated as:

$$\text{TF-IDF (t, d, D)} = \text{tf}_{\text{t, d}}\ \text{X idf}_{\text{t, D}}$$

TF-IDF Algorithm (TF-IDF) is basically a mathematical statistic that is meant to show how important a word is to a document in a collection of documents. In information retrieval systems, it is used as a weighting factor. At the number of times a word appears in the document increases, the TF-IDF value increases proportionally. But this TF-IDF value is decreased by the frequency of the word with the collection. This helps to take into account the fact that some words appeared more frequently in general. It is a logarithmic obtained value and it is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient and the log of this term is calculated to a value obtained is the IDF.

$$Idf(t,D) = log\frac{N}{|\{d\in D:t\in d\}|} \tag{4}$$

Then the TF-IDF is calculated as

$$tfidf(t,d,D) = tf(t,d)x\ idf(t,D) \tag{5}$$

The inverse document frequency (IDF) is a measure of how much information on the documents provides, that is, whether the term is common or rare (some words such as common stop words are frequent that words do not help discriminate between one document over another) across all documents. Term Frequency -Inverse Document Frequency (TF-IDF) is a statistical weighting technique that has been applied to many types of information retrieval problems. For example, it has been used for automatic indexing [37], query matching of documents and automated summarization [5]. One other important contribution to the Hybrid TF-IDF equation is its normalization of words of a document occurs, which allows it to carefully control the overall target summary length. As described by Singhall et al. [38], had established TF-IDF is very sensitive to English text document length and often over weights terms from longer documents for capital and small characters. In our application of TF-IDF, we observed not the same effect in Amharic language like English words. Without a normalization word, established TF-IDF, given the most weight to the longer documents since the weight of a document is the simple sum of the weights of the composing words. The given below is the research study of the related literature about these approaches and systems for multi- document summarization.

2.8.1. Similarity Measure

The cosine similarity measure is one of the common techniques used to measure similarity between a pair of sentences vectors. Here sentences are represented as a weighted vector between the sentences in the posted documents on social media from Twitter and Facebook.

2.8.2. Cluster Based Method

Basically, clustering is to group similar sentences score values of their classes. For clustering of multi documents, these documents refer to the sentences and the cluster that a sentence belongs to be represented by classes of the group.

2.8.3. Word Frequency

The basic idea of using word frequency is that important words are found many times in the document. Tf and idf is one of the most common measure used to calculate the word frequency. After this we calculated the combine the tf and idf to score the important of the word to find in the documents to the summary.

2.8.4. Feature Based Method

The extractive summarization types include for identifying the relevant sentences from the text and put them together to create an accurate summary. Some of the features that are considered for the selection of sentences are the significant word location of sentence, length of sentence from the input documents etc.

3. Proposed Methodology

3.1. Data Sets

We used to collect training dataset on the Twitter and Facebook posted documents to Amharic text input corpus 4951 posted sentences in totals (protests (3943), droughts (667), floods (101) and sports (240)) in different news items. For those in the collected posted documents from Facebook and Twitter to obtain the sources in protests in totals 120,862 posts, in droughts posted is 43,774 posts, in sports posted is 10,299 posts and in floods is 1,209 are the training sets posts are identified by the format of date, month and years. We have the following training data sets and we could be tested the data sets are #protests, #droughts, #sports and #floods data on social media posted texts that could be found in the news items posted Amharic text documents over a time in social media. The data set to consist 4 news items posted texts to collect data onto the Facebook and Twitter posted text documents and selected to extract from summary the important sentences in the input to the summary.

Table 1. *Preparation of data sets pre-processes.*

Datasets items	Size (in sentences)	Similarity posts (in sentences)	Clusters (in documents), K=3 groups, Sentences		Amharic posted (dd-mm-yyyy) format
Protests	3,943	534,994	C1(doc1)	201,486	43,236
			C2(doc2)	166,797	39,805
			C3(doc3)	166,711	37,821
			Total	534,994	120,862
Droughts	667	224,971	C1(doc1)	84,380	16850
			C2(doc2)	70,429	12979
			C3(doc3)	70,162	13945
			Total	224,971	43,774
Sports	240	28,824	C1(doc1)	10,804	3470
			C2(doc2)	9010	3631
			C3(doc3)	9010	3198
			Total	28,824	10,299
Floods	101	4,960	C1(doc1)	1833	459
			C2(doc2)	1554	419
			C3(doc3)	1573	331
			Total	4,960	1,209
Totals	4,951	793,748			

In the above table 1, we made for training and testing the posted Amharic texts for social media on Facebook and Twitter posted in Amharic texts to take and analysis the samples to process the similarity and cluster for the similar posted sentences automatically. Based on these data to find the similarity with each posted Amharic texts with a pair of sentences on social media and identify the group of the similar clusters, the numbers of input k are three (created on distance values or nearest distance value to group the similar items of sentences to cluster and after cluster we had been found three clustered documents for each posted text documents for sentences to summarize individually clustered.

3.2. System Design for the Research Study

Step 1. Pre-processing of input post document: This would work on the Amharic posted documents for processing by the rest of the system. The preprocessing involves sentence segmentation in the document, tokenization of the segmented sentences, stop word removal of the list of words, stemming words.

Step 2. Similarity measure: Sentence similarity was computed as a linear combination of sentence similarity and word similarity. The cosine similarity measure is one of the common techniques used to measure similarity between a pair of sentences to be calculated the similarity between each sentence and after similarity to group the similar clusters to form.

Step 3. Clustering: For clustering of multi documents, these items refer to the posted documents for sentences and the cluster that a similar sentence belongs to represented by classes of the group in a shorter or nearest distance value calculated by the centroids using the formula of Euclidean distance function. Once we used with post text data, k-means clustering can provide a great way to organize the thousands-to-millions of words, phrases or sentences for posted documents. Since, the sentences for randomly for each document set for the input similarity results, posted documents in the sentences and to give clusters based on similarity values of news items posted Amharic texts on social media and each set clusters the similar posted documents for the user posts on the social media [29]. When we summarize the user posts by clustered document for each post of clusters by similar distance values Amharic text documents for the social media on the Twitter and Facebook posted data. Day to day user posts to arrange with the same cluster like political #Protests with #Protests similar contents, #droughts with #droughts, #floods with #floods, #sports with # sports, #health with #health the same posted documents or sentences and others user posts to arrange or order the tweets for the algorithm and summarize for each user posts. In our research the data onto social media could not be identified automatically protests at droughts, protests at sports posted or other data sets. So, we could have a different corpus for data sets to train and test the performance for Amharic text documents posted on social media.

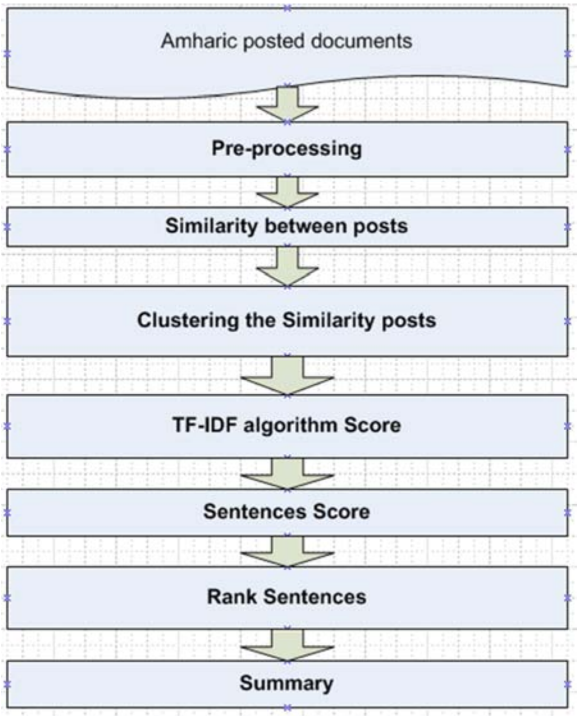


Figure 1. *The Architecture of the system design.*

Step 4. TF-IDF algorithm Score: The TF-IDF value was composed of two primary parts. The term frequency (TF) and inverse document frequency (IDF). TF component assigns more weight to words that occur frequently to a document. Because important words are often repeated in one document and a single document that encompasses all the documents together [4]. Therefore, TF-IDF gives the most weight towards that occurs most frequently within a small number of documents and the least weight to terms that occur infrequently within the majority of the documents.

$$\mathrm{Tf}\,(\mathrm{t,d}) = \frac{\text{term counts for the stem}}{\text{total terms in the document}} \quad (6)$$

$$\mathrm{IDF}\,(\mathrm{t, Docs}) \;= \log 10\,(\frac{\text{Docs}}{\text{t appeared in all Docs}}) \quad (7)$$

$$\mathrm{Tfidf}\,(\mathrm{t, d, Docs}) \;=\; \mathrm{tf}\,(\mathrm{t,d}) * \mathrm{idf}\,(\mathrm{t, Docs}) \quad (8)$$

Where, t is term or word, d is document, Docs = all documents in the corpus.

Step 5. Sentence scoring: this step after the TF-IDF stem word score, each term of obtaining the score its values if this sentence determines the score of each sentence and several possibilities exist. The score can also be made to the number of sentences in which the words in the sentence appear in the document. Scoring each sentence is to rank each sentence we need to score each sentence using the TF-IDF values calculated before. Rather than simply taking the sum of all the values of a given word with one sentence sequentially. These include only summing TF-IDF value score of the term where the word is a noun, verb and others, but stop words could not be added for the sentences that could be scaled down the stop words in a sentence.

$$Score(q,d) = \sum_{t\in q} tf\_idf\ t,d \quad (9)$$

Step 6. Sentence Ranking: after the sentences scores for the list of stem words, the sentences will be ranked according to the sentence scores values and any other measures like the position of a sentence in the document can be used to control the ranking. After applying the sentences scored, we can finally sort our sentences in descending (top value to low value) orderly the sentences score values to sort. For example, even though the scores are high, we would be putting first ranks sentences and compare to each score value of the sentences higher score values to order the sentences.

Step7. Summarizing Extraction: - After ranked the sentences list, if the user to select the input on the size of the summary, the sentences will be picked from the ranked lists orderly. The news items text summarization is the process of automatically creating a compressed version of a given text that provides useful information about the user in the social media. The length of a summary depends on the user's needs.

4. Experimentation and Results of Evaluations

4.1. Similarity Measures Between Sentences

Sentences are made up of words, so it is reasonable to represent a sentence using the words with the sentence. The most relevant research area, to our task is the Amharic text summarization news items posted on social media. The cosine similarity measure is used to measure similarity between a pair of sentences and also after similarity to group the similar clusters to form. Given two pairs of sentences:

S1= {w11, w12... w1m1} S2= {w21, w22... w2m2}

Where wi, j is the j^{th} word of Si (i=1, 2), mi is the number of words in Si. A joint-word set S = S1 U S2 is then formed into distinct words with S1 and S2.

S =S1 U S2 = {w1, w2... Wm}, that has m distinct words. The joint word sets S contains all distinct words of S1 and S2. And also, we could be intersected the similarity in words with the two sentences word similarity.

$$\begin{pmatrix} s1,1, s1,\dot{2}, \ldots . s1, n \\ s2,1, s2,2, \ldots . s2, n \\ \cdots \quad \cdots \qquad \cdots \\ \cdots \quad \cdots \qquad \cdots \\ \cdots \quad \cdots \qquad \cdots \\ Sp, 1, Sp, 2, \ldots . Sp, n \end{pmatrix}; S\mu\mathrm{j} = \begin{Bmatrix} \omega\mu i, j \; when\; word\; j \;\; \in S\mu \\ 0 \quad otherwise \end{Bmatrix} \qquad (10)$$

Where each μ row contains the $\omega\mu i, j$ weighting of the word j in $S\mu$ sentence. Example.

S1= 9-Feb-2016 ኢትዮጵያውያን ሰላማዊ ሰልፍ በዋሽንግተን ዲሲ #OromoProtests #Ethiopia Protests #Oromo #Ethiopia "በኦሮሚያ፥ በጎንደርና በጋምቤላ በዜጎች ላይ እየደረሰ ያለው በደል ይቁም!"

S2= 2-Feb-2016 ኢትዮጵያውያን ዛሬ በዋሽንግተን ዲሲ ሰላማዊ ሰልፍ አደረጉ "በኦሮሚያ፥ በጎንደርና በጋምቤላ በዜጎች ላይ እየደረሰ ያለው በደል ይቁም" "አምባገነኖች የልማት ሃጋር አይደሉም" የዛሬው የዋሽንግተን ሰልፈኞች። etc. Two sentences similarity is calculated as formula: S1= sentence1, S2 = sentence2,∩ = in tersection,∪ = union Sim = similarity between two sentences

$$\mathrm{Sim}(\mathrm{S1}, \mathrm{S2}) = \frac{2{*}match\ (\mathrm{s1} \cap \mathrm{s2})}{\mathrm{len}(\mathrm{S1} \cup \mathrm{s2}\ )} \qquad (11)$$

In the work of Simpson [39], his presented the method of the syntactic similarity with the two strings of capturing the similarity between words that was concerned to work. The work of the author to compute the similarity between two sentences and he basically to capture for the semantic similarity between the two-word senses for the strings of the path length similarity.

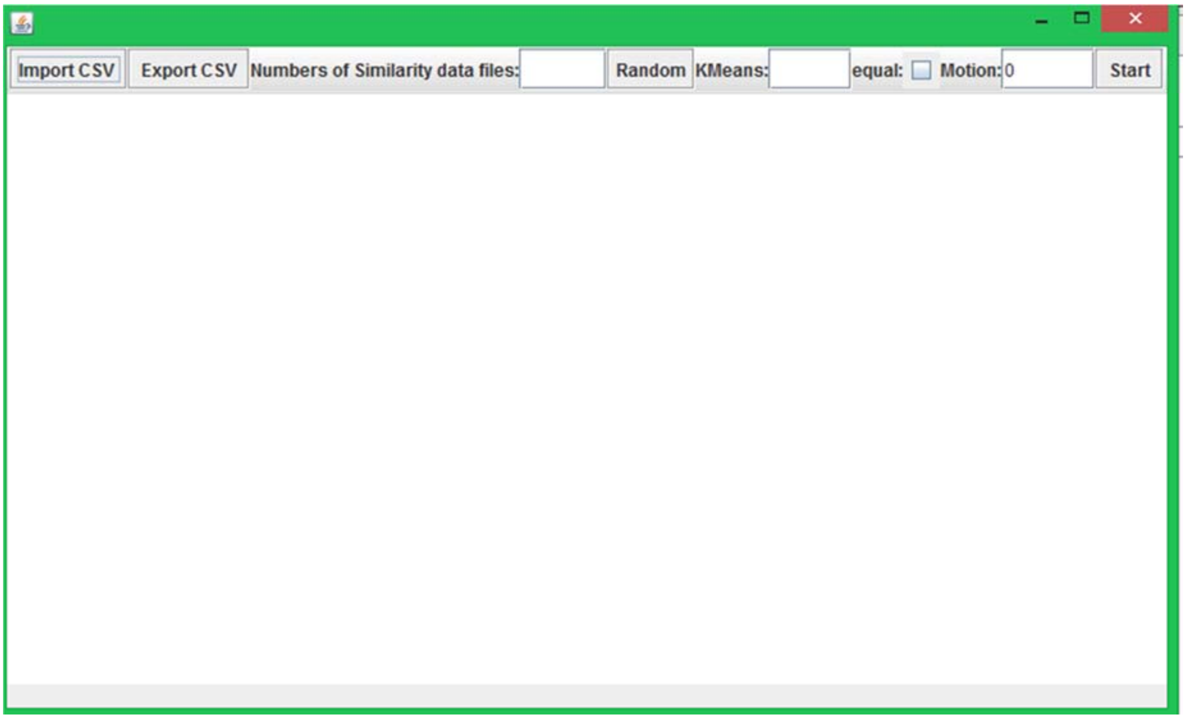


Figure 2. *User interfaces with display clusters.*

4.2. Clustering Based Algorithms Using K-means

Clustering is to group similar posted sentences into their classes. It is a process of creating groups of similar items or posted documents. Clustering to find clusters of data posts that are similar in some nearest distance values of one another. The members of a cluster are more like each other than they are like members of other clusters. The main aims at clustering algorithms are similar to one another within the same cluster and dissimilar to the objects in other clusters. The Kmeans data sets to access and Microsoft Excel have been used for initial preprocessing and storing the data, and in particular to analysis the data sets. K-means Algorithm:

Input: input the number of cluster k with centroid (mean).

Process: Step 1: partition the data into k- cluster or k- non empty subsets.

Step 2: compute the mean for each partition.

Step 3: assign each object of the cluster of its nearest centroid (mean).

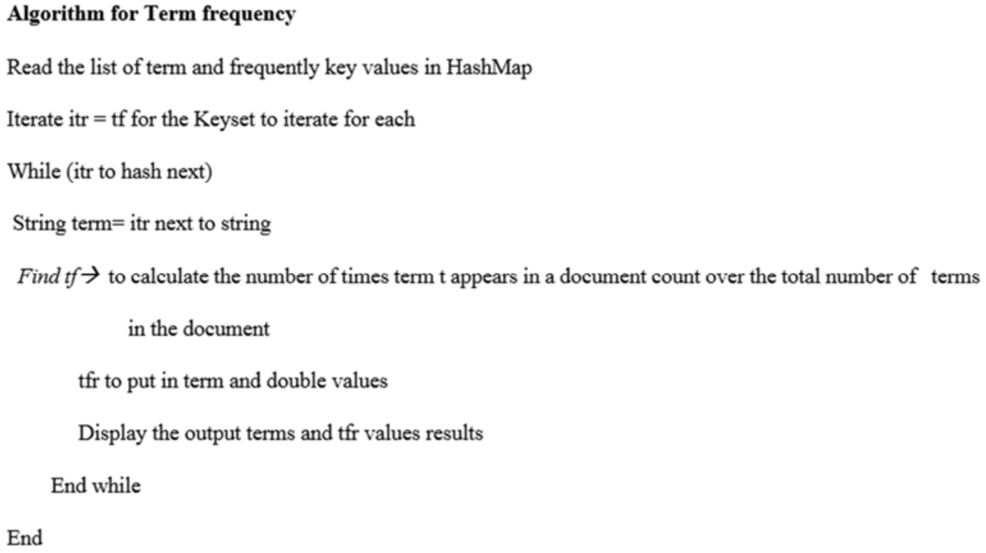
Step 4: Step 2 and Step 3 continuous until no change in the mean values and also sum of sequence of distance is minimized within the clusters.

Output: number of clusters of partitioned data objects.

4.3. Hybrid TF-IDF Algorithm Calculates

The TF-IDF is one of the techniques that are to pick the most frequently occurring terms (words with high term frequency or tf). However, the most frequent word is a less useful since some words like the stop word occur very frequently to all documents. Hence, we also want a measure of how unique a word is i.e., how infrequently the word occurs to all documents (inverse document frequency or IDF). Hence, the product of Tf x IDF of a word gives a product of how frequently, this word is in the document multiplied by how unique the word is written the total corpus of documents.

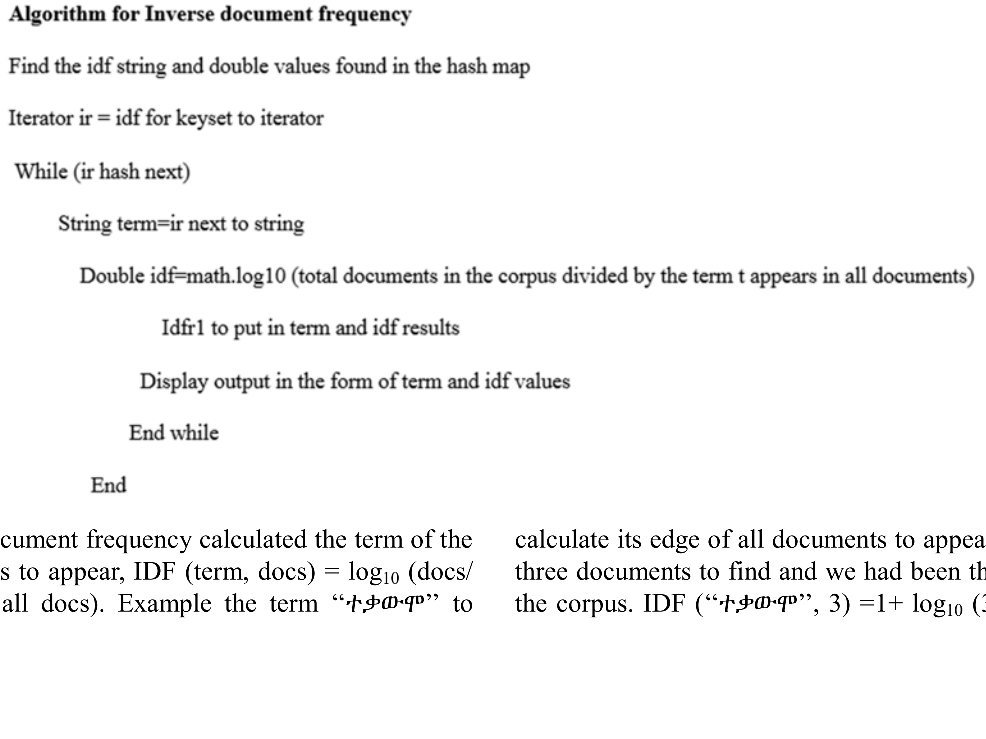
Term frequency (*ft*): which measures how frequently a term occurs of a document. Since every document is different to length, it is possible that a term would appear much more time for the longer document than shorter ones. TF (t, d) = (Number of times term t appears in a document) / (Total number of terms in the document).



```
Algorithm for Term frequency
Read the list of term and frequently key values in HashMap
Iterate itr = tf for the Keyset to iterate for each
While (itr to hash next)
 String term= itr next to string
 Find tf→ to calculate the number of times term t appears in a document count over the total number of  terms
          in the document
      tfr to put in term and double values
      Display the output terms and tfr values results
    End while
End
```

The term frequency calculates tf (t, d) = terms count/ total of the terms of document1 words, Example the term (“ተቃውሞ”) to the word occurs in document one is 1060 and total document contain words 4,802,014, to calculate tiff as follows. So, tf (“ተቃውሞ”, d) = 1060/4,802,014=2.2074071420866327E-4.

Inverse document frequency (IDF): The inverse document Frequency is the number of times a word occurs to a corpus of documents. IDF, which measures how important a term is. While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "ነበር", "ናቸው", "እና", "ይህ", "ያ", may appear a lot of times but have little importance. Thus, we need to weight down the frequent terms while scale up the rare ones, by computing the following: IDF (t, Docs) = $\log_{10}$ (Total number of documents of the corpus/ Number of documents with term t in it appears on all documents).

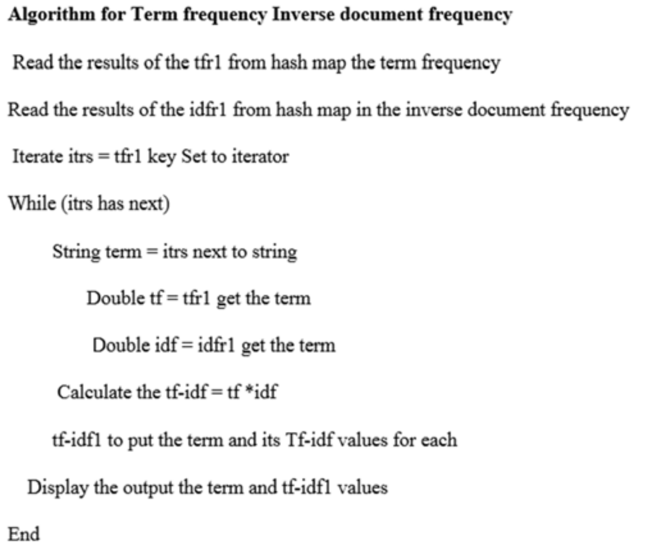


```
Algorithm for Inverse document frequency
Find the idf string and double values found in the hash map
Iterator ir = idf for keyset to iterator
 While (ir hash next)
      String term=ir next to string
        Double idf=math.log10 (total documents in the corpus divided by the term t appears in all documents)
                Idfr1 to put in term and idf results
              Display output in the form of term and idf values
             End while
       End
```

The inverse document frequency calculated the term of the protest documents to appear, IDF (term, docs) = $\log_{10}$ (docs/ term appears in all docs). Example the term “ተቃውሞ” to calculate its edge of all documents to appear the term with in three documents to find and we had been three documents in the corpus. IDF (“ተቃውሞ”, 3) =1+ $\log_{10}$ (3/3) =1+0=1. The

IDF value is zero means if the term occurs to all documents in the corpus occurred or frequently appear in all documents in the corpus, so this term is usefulness of the documents the IDF to scale down the term values.

TF-IDF (Term frequency- Inverse document frequency): It is a way to score the importance of words (or "terms") in a document based on how frequently they appear across multiple documents. The TF-IDF is used to weight words, according to how important they are. Words that are used frequently in many documents will have a lower weighting while infrequent one term will have a higher weighting. Intuitively, if a word appears frequently in a document, it's important. It gives the word a high score. But if a word appears in many documents, it is not a unique identifier. It gives that the word a low score in the documents. Therefore, common words like “እና", "ስለ"," ወይም", which appear in many documents, will be scaled down.



```
Algorithm for Term frequency Inverse document frequency
 Read the results of the tfr1 from hash map the term frequency
Read the results of the idfr1 from hash map in the inverse document frequency
 Iterate itrs = tfr1 key Set to iterator
While (itrs has next)
      String term = itrs next to string
           Double tf = tfr1 get the term
            Double idf = idfr1 get the term
       Calculate the tf-idf = tf *idf
      tf-idf1 to put the term and its Tf-idf values for each
   Display the output the term and tf-idf1 values
End
```

The TF-IDF value score, example the term (“ተቃውሞ”) to the word occurs in the document the multiplication of tf and idf. TF-IDF(term,d)=tf(term,d)*idf(term,docs)=2.2074071420866327E-4*1.0 = 2.2074071420866327E-4.

4.4. News Items Posted Text Summaries

We implement the news items posted texts in the current time to post texts in social media like protests, droughts, sports, floods and others to post in the social media on Facebook, Twitter and others social media, but we focused on the post texts on Facebook and Twitter post Amharic texts to summary by using date, month and years to summaries the poster texts to give to the readers within a short period of time to summaries the posted documents. News items posted to identified by the time of the intervals of the period posted texts in the social networks/media like to use the date, weeks, months, years, one year’s etc. posted texts to summarize the texts. For example, if the user selected for political protests summary to need within the monthly to group for multi documents within a one month and to want to summarize by entering the month and year into the system to find automatically the system group posted text document based on the user asked the month to group the same month and years rank each posted sentences for a multi documents post and also give to the summarize sentences to the reader. If the compression ratio to calculate the selected posted sentence for monthly in to post text documents divided by the total monthly posted text documents the give the percentage. Example protests posted texts to display each clustered text with console in Java programming (NetBeans).

For clustered document 1

How many sentences do you want select in doc1 to the summarization: 2?

Enter date of summary time Month-Year: Mar-2016

Total numbers of posted text documents in doc1:104

The compression ratio of doc1 is: 98.07692307692308

1) 8-Mar-2016 በኦሮሚያ ዘገባ ለማቅረብ የተንቀሳቀሱ ጋዜጠኞች በቁጥጥር ስር ውለው ተለቀቁ #Ethiopia #FreePress #OromoProtests #Journalism: 1.8259625047899546E-5
2) 12-Mar-2016 በኢትዮጵያ ኦሮሚያ ክልል ውስጥ ከተቀሰቀሰው ተቃውሞ ጋር በተያያዘ የኦሮምያ ዳያስፖራ ማኅበረሰብ ሰላማዊ ሰልፍ በዋሽንግተን ዲሲ #Diaspora #Oromo #OromoProtests በኦሮሚያም ሆነ በሌሎች የሃገሪቱ ክልልች ለተፈጠረው የሕዝብ ቅሬታ፥ መንግሥታቸውና ፓርቲያቸው ሙሉ ኃላፊነት እንደሚወስዱ ጠቅላይ ሚኒስትር ኃይለማርያም ደሳለኝ አስታወቁ: 1.60538780117E-5

==

For clustered document 2

How many sentences do you want select in doc2 to the text Summarization: 1?

Enter date of summary time Month-Year: Jun-2016

Total number posted text documents in doc2: 40

The compression ratio of doc2 is: 97.5

15-Jun-2016 ቤንች ማጂ ዞን የርእሰ መምሕር ቢሮ በመታሸጉ ትምሕርት መስተጓጎሉን ወላጆች በተቃውሞ ተናገሩ: በደቡብ ክልል ቤንች ማጂ ዞን የሚገኝ የሁለተኛ ደረጃ ትምሕርት ቤት ርእሰ መምሕር ቢሮ በመታሸጉ: 0.102

===

For clustered document 3

How many sentences do you want select in doc3 to the summarization: 3?

Enter date of summary time Month-Year: Aug-2016

Total numbers of posted text documents in doc3: 90

The compression ratio of doc3 is: 96.666667

1) 23-Aug-2016 መንግስት እየወሰደ ያለው የሃይል እርምጃ ተቃውሞን ከማባባስ ውጭ መፍትሄ አያመጣም ተባለ ኢሳት (ነሃሴ 16 ፥2008) የኢትዮጵያ መንግስት በኦሮሚያና በአማራ ክልሎች ሰላማዊ ሰልፈኞች ላይ እየወሰደ ያለው የሃይል እርምጃ ህዝባዊ ተቃውሞውን ከማባባስ ውጭ መፍትሄ አያመጣም ሲሉ የኦባማ አስተዳደር የዲሞክራሲና የሰብዓዊ ጉዳዮች ከፍተኛ ባለስልጣን ገለጹ የኢትዮጵያ ቀጣዩ ብሄራዊ ስራ የፖለቲካ ምህዳሩን ማስፋት እንደሆነ ገልጸዋል:2.014347526732104E-4
2) ሃላፊው በቅርቡ በአማራና በኦሮሚያ ክልሎች የተከሰቱ ህዝባዊ ተቃውሞዎች ለኢትዮጵያና ለአሜሪካ ከፍተኛ ፈተና እንደሆኑ ቶም ማሊኖውስኪ ባሰራጩት ጹሁፍ ገልጸው፣ በህገመንግስቱ የተገለጹ የአስተዳደርና ችግርና የፖለቲካ ብዝሃነት ማነስ ላይ ተመስርተው የቀረቡ ጥያቄዎች ናቸው ሲሉ በጹሁፋቸው አብራርተዋል:1.704464172197319E-4
3) በራስ ተነሳሽነት ለሰላማዊ ሰልፍ በነቂስ የወጣው የቡሬ ከተማና አጎራባች መንደሮች ህዝብ፣ የብዓዴን የድርጅት አርማም እንዳቃጠለ፣ “ወያኔ

ሌባ” በማለት ተቃውሞ ሲያሰሙ እንደዋሉ ለማወቅ ተችሏል:1.598495843953218E-4

=the three documents are completely finished the process=

4.4.1. System Summarization for Posting Texts for Each Experiment

As to describe to the four experimentations for each post documents by summarizing extracted rate under each experiment, the four experiments are referred as E1, E2, E3 and E4. All of these experiments would be conducted for all the posted Amharic input texts for datasets with the summary extraction percentages of 10%, 20% and 30%. Those summarize extraction percentages are selected to be observed the effect of the summarization processes on different ranges of summarization posted text documents. The total numbers of automatic summaries of each post texts in protests, droughts, sports and floods are 120,862, 43,774, 10,299 and 1,209 respectively posted Amharic texts as training sets on social media, having 30 posts texts automatic summaries in each experiment for as test set for each clustered post text. For this is 30 posts texts are selected randomly in our training data set corpus of the clustered post documents as to a summary.

System summary is the number of sentences extracted for each post texts using the selected percentage are to determine by multiplying the number of sentences in the post texts in a sentence level by the percentage and rounding it's to the nearest zero post texts in a sentence. That is if the original one clustered posts text has 67 sentences, within a 10% summary that has 7 sentences, within 62 sentences in a 10% summary that has 6 sentences and so on to find for each sentence in a document. It is depending on the extraction rate post texts and the number of sentences in the document, this Summarizer selects the first "$n\ text\ file$" sentences of the documents.

$$n = (extraction\ rate\ x\ number\ of\ sentences) \qquad (12)$$

Those lists of the number of sentences extracted using each percentage rate for each posted text file is presented for each experiment in table 2.

Table 2. *The number of sentences extracted using the selected percentages for each posted text.*

File name	Post texts	No. of totals	No. sentence in System and Ideal summary under E1, E2, E3 and E4		
Protests posts		sentences	10%	20%	30%
Prot1.txt	30	53	5	11	16
Prot2.txt	30	62	6	12	19
Prot3.txt	30	53	5	11	16
Average score	30	56	5	11	17
Drought posts					
Drot1.txt	30	67	7	13	20
Drot2.txt	30	62	6	12	19
Drot3.txt	30	59	6	12	18
Average score	30	63	6	12	19
Sport posts					
Spot1.txt	30	42	4	8	13
Spot2.txt	30	38	4	8	11
Spot3.txt	30	34	3	7	10
Average score	30	38	4	8	11
Floods posts					
Flod1.txt	30	56	6	11	17
Flod2.txt	30	43	4	9	13
Flod3.txt	30	30	3	6	9
Average score	30	42	4	9	13

4.4.2. Manual Evaluation for Each Experiment

From in the manual evaluations are prepared for 40 summaries texts to produce from 30 posted texts in news items posted Amharic texts on the social media in the protests, droughts, sports and floods post clusters texts at the three-summary extraction percentage rate used in this the experiments. As discussed in the earlier sections, the six manual evaluation, quality measured criteria used to grade the summaries on the scales to give from 1-5. For these grades show that the level of the criteria under the respect of the summary information is achieved in and for each summary in the documents.

4.4.3. Comparison of Each Experiment

From the four experiments are implemented in the posted texts on social media in protests post (E1), droughts post (E2), sports post (E3) and floods post (E4) for Twitter and Facebook the users posted Amharic texts. We were tested for each post to take 30 posted texts for each text file and 40 summaries to select in ranked sentences for each experiment both the automatic summary and human summary. The main difference for each experiment for the use of the porter stemmer in the different post texts to remove the affixes by using the Java tools. For this each summary to evaluated by the human/ideal summary in all experiments to generate the summaries the post texts in the social media. During the experiments in the four experimentations that observed the difference between their performance on the effectiveness and efficiency to compared to extract the summary and extraction percentages. For the systems evaluated in 1st, 2nd, 3rd and 4th experiments are effective for all groups of post texts on the extraction percentages and the size of the summary the average score for each experiment is greater than 70% to score the performances. In all the experiments the evaluation results that could be shown the automatic summaries for using the Amharic stemmers to perform the extraction rate of the size of the posted texts with the better consistency in higher rates. In the first experiment protest posts (E1) the average results of the performances as well as in the table, the extraction rates at 10% to compare the other's experiments, it performs better results than E2 in the groups to show that as 0.19%, 7%, 12.11% for protests post group and the droughts group 7.19%, 12.3% and also in the sports group 5.11% posted texts to change. We compared for the first experiments to the second experiments are more condensed or compress for the summary, but in the third and the fourth experiments are less compressing the posted documents at 10% extraction rate. For the others at 20% extraction rates for the size of the summary in the first experiments (E1) is better results than E2 in the group of post texts, such as 1.81%, 3.9%, 4.21% and in the droughts 5.71%, 6.02% and for in sports post groups is 0.31%. We compare the first experiments for the others experiments in the extraction rate at 20%, it is greater than E2 for compress the summary, but it is less than others 20% extraction rate for the post texts. The size of the summary increase, the extraction rate is increasing in the summary. In all the experiments at 20% summary the highest performance to compress the summary in the E4, E3, E1, E2 and 0.31%, 5.71%, 1.81% and 3.9% respectively. At 30% extraction rate for the summary to compare for each experiment (E1) is the best result outperformed than E2 in the groups of posts for 3.07%, 4.3%, 6.45% and in the droughts post 7.33%, 9.52% and in the sports 2.15%. In the experiments at 30% summary the highest performance to compress the summary in the E4, E3, E1, E2 2.15%, 7.37%, 3.07% and 4.3% respectively. In the average score for experiment E4 is more condensed the posted documents for the summary than the others experiment, which is 93.52% at 30% extraction rate.

Table 3. *The average performance comparison results for all experiments.*

Group File	**Average of F-measure in percentage (%) for all E1, E2, E3, E4**																	
At10%	**E1**	**E2**	**Differ**	**E1**	**E3**	**Differ**	**E1**	**E4**	**differ**	**E2**	**E3**	**differ**	**E2**	**E4**	**differ**	**E3**	**E4**	**differ**
Protests (E1)	74	73.81	-0.19	74	81	7	74	86.11	12.11									
Droughts (E2)										73.81	81	7.19	73.81	86.11	12.3			
Sports (E3)																81	86.11	5.11
Floods (E4)																	86.11	
At20%																		
Protests (E1)	85.35	83.54	-1.81	85.35	89.25	3.9	85.35	89.56	4.21									
Droughts (E2)										83.54	89.25	5.71	98.25	89.56	6.02			
Sports (E3)																89.25	89.56	0.31
Floods (E4)																	89.56	
At30%																		
Protests (E1)	87.07	84	-3.07	87.07	91.37	4.3	87.07	93.52	6.45									
Droughts (E2)										84	91.37	7.37	84	93.52	9.52			
Sports (E3)																91.37	93.52	2.15
Floods (E4)																	93.52	

4.5. Discussions Results

In this paper, we reported the evaluation of the results our research news items posted Amharic text summarization on social media for posting text on Twitter and Facebook and how to process, the steps minimize the error and to increase the products of the results to summarize Amharic text our research algorithm is hybrid TF-IDF. The evaluation of the Amharic text summarize the Intrinsic evaluations are made by comparing for automatic summary within the human/ reference summary made manually prepared the summary posted texts on social media. It is measuring the automatic system summary performance on its own posted texts. We evaluated the process of the system an intrinsic method by using the subjective (qualitative) and objective (quantitative) for posting texts in social media to summaries by automatic and manual summary. The subjective evaluations are used to measure the Informativeness, coherence's structure and linguistic quality measuring criteria of the automatic summary generate for the posted documents. For the others evaluation method is objective evaluations that are measured the summarizer of the performance in the extraction summary and also the identification of the post documents of salient sentences with the given post texts. For this to measure the performance of the standard for precision and recall measure in the given input texts, human summary and Summarizer extract percentages how to close the extracts for each other's for the system. Whereas, in the extrinsic evaluations is completed based on the evaluating how the automatically summaries the systems are good enough to be accomplished the purpose of some other specific tasks. As evaluation of the experiments for each training data sets to compress the text summarization by automatically if the user selected the size of the summary of the document and to condense the document. We tested the best compression rate for each training data for protests posts higher or better compress ratio sequentially to condense the objective summary. For the average score for the performance measuring the three extraction percentages at 10%, 20% and 30% are presented for the protests post texts in the protest group 74%, 85.35% and also 87.07% summarize the protests post texts in social media.

The droughts training data sets, we tested the best summarize document sequentially to condense the objective summary. For the average score for the performance measuring the three extraction percentages at 10%, 20% and 30% are presented for the droughts post texts in the droughts group 73.81%, 83.54% and also 84% summarize the droughts post texts in social media.

Sports training data sets, we tested the summarize cluster document sequentially to condense the objective summary. For text summarization in this post texts. For the average score for the performance measuring the three extraction percentages at 10%, 20% and 30% are presented for the sports post texts in the sports group 81%, 89.25% and also 91.37% summarize the sports post texts in social media. In Floods training data sets, we tested the summarize document sequentially to condense the objective summary. For the average score for the performance measuring the three extraction percentages at 10%, 20% and 30% are presented for the floods post texts in the floods group 86.11%, 89.56% and also 93.52% summarize the floods post texts in social media.

Future Work

1) Apply proposed algorithm to summary the identifying the news item posted texts automatically grouped on social media post news rather than to prepare manual the corpus separately posted news items on summarization for others local language.
2) Applying for the tasks of multi-documents posted Amharic texts and sequential update summarization tasks are one possible future work to do the research on social media posts.
3) Apply more Amharic lexicon rules and the dictionary file to use a dictionary to control over and under stemming the Amharic words.

5. Conclusion

In this paper, we introduced Amharic text summarization for news items posted on social media. In the information overload on social media are different news items posted in Amharic texts. For this is a single or multi document to need summarization that contain in the frequent terms or repeated posting documents to summarize using hybrid TF-IDF algorithm is introduced. As discussed in the previous chapters, in this research, hybrid TF-IDF algorithm was recommended for Amharic and for under resourced languages in social media posted text documents. The TF-IDF is used to process for counting words in a document as well as throughout a corpus of documents to the end of sorting documents in statistically relevant ways. The growing of the web for social media a huge amount of information is posted and the posted documents coming from different sources at this time, it becomes very difficult for summarizing the documents to find specific information they want to summarize. It provides support for those users to get that relevant information in posted Amharic texts and summary is today the mandatory to get relevant important information for the documents for social media.

Acknowledgements

I would like to thank our Advisors and families for their given comments, and also those who helped us during this research study.

References

[1] Q. Guo, F. Diaz, and E. Yom-Tov, “Updating users about time critical events,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7814 LNCS, pp. 483–494, 2013.

[2] E. N. Agency, “Amharic Text classfication,” 2000.

[3] T. Xu, P. Mcnamee, and D. W. Oard, “HLTCOE at TREC 2013 : Temporal Summarization,” 2013.

[4] D. Chakrabarti and K. Punera, “Event Summarization Using Tweets,” pp. 66–73, 2011.

[5] F. Chong and T. Chua, “Automatic Summarization of Events from Social Media,” 2009.

[6] H. Sayyadi, M. Hurst, A. Maykov, and M. Livelabs, “Event Detection and Tracking in Social Streams*,” pp. 311–314, 2009.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: real-time event detection by social sensors,” WWW ’10: Proceedings of the 19th international conference on World wide web, p. 851, 2010.

[8] H. L. Eidheim, “Temporal Summarization of Time Critical Events,” no. June, 2015.

[9] P. W. Mcburney and C. Mcmillan, “Automatic Source Code Summarization of Context for Java Methods,” vol. 5589, no. c, pp. 1–18, 2015.

[10] K. R. Premlatha and T. V. Geetha,“Extracting Temporal Patterns and Analyzing Peak Events,” Thesis, Department of computer science, Anna University, Vol 2, July-Dec 2010.

[11] J. Makkonen and H. Ahonen-myka, “Utilizing Temporal Information in Topic Detection and Tracking,” 2005.

[12] T. Melese, “Automatic Amharic Text Summarization using Latent Semantic Analysis,” Thesis, Department of Computer Science, Addis Ababa University, 2009.

[13] A. Addis, “Automatic Summarization for Amharic Text Using Open Text Summarizer”. Thesis, Department of Information Science, Addis Ababa University, 2013.

[14] D. Eyob, “Topic-based Amharic Text Summarization Topic-based Amharic Text Summarization,” no. March, 2011.

[15] K. D. Dessalegn and M. Y. Tachbelie, “Graph-based Automatic Amharic Text Summarizer,” vol. 8, February-2017.

[16] H. Saggion, D. Radev, S. Teufel, and S. M. Strassel, “Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment,” no. 2000, 2001.

[17] D. R. Radev and A. Arbor, "Centroid-based summarization of multiple documents : sentence extraction, utility-based evaluation, and user studies," 1989.

[18] R. Barzilay and K. R. Mckeown, "Inferring Strategies for Sentence Ordering in Multidocument News Summarization," vol. 17, pp. 35–55, 2002.

[19] "COMPARATIVE EVALUATION OF MODULAR AUTOMATIC SUMMARISATION SYSTEMS USING CAST Constantin Or ˘," 2006.

[20] D. Marcu and M. Rey, "Discourse-Based Summarization in DUC-2001 Summarizing document," 2001.

[21] D. Inouye and J. K. Kalita, "Comparing Twitter Summarization Algorithms for Multiple Post Summaries," 2010.

[22] Z. He, C. Chen, J. Bu, C. Wang, and L. Zhang, "Document Summarization Based on Data Reconstruction," pp. 620–626, 2011.

[23] P. Paroubek, S. Chaudiron, L. Hirschman, L. Cnrs, B. Université, and P. Xi, "Principles of Evaluation in Natural Language Processing," vol. 48, pp. 7–31, 2007.

[24] K. Tao, F. Abel, Q. Gao, and G. Houben, "TUMS : Twitter-based User Modeling Service," pp. 1–15, 2010.

[25] W. Chung, H. Chen, L. G. Chaboya, C. D. O. Toole, and H. Atabakhsh, "Evaluating event visualization : a usability study of COPLINK spatio-temporal visualizer," vol. 62, pp. 127–157, 2005.

[26] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin, "Earlybird : Real-Time Search at Twitter," 2011.

[27] M. Potthast, B. Stein, F. Loose, and S. Becker, "Information Retrieval in the Commentsphere," vol. V, no. January, 2012.

[28] I. Novalija, M. Papler, and D. Mladeni, "TOWARDS SOCIAL MEDIA MINING : TWITTEROBSERVATORY," pp. 2–5.

[29] B. Sharifi, M. Hutton, and J. K. Kalita, "Experiments in Microblog Summarization," 2004.

[30] P. Meladianos and I. R. C. Athena, "Degeneracy-based Real-Time Sub-Event Detection in Twitter Stream," 2015.

[31] F. Ibekwe-sanjuan, S. Fernandez, E. Sanjuan, and E. Charton, "Annotation of Scientific Summaries for Information Retrieval," pp. 1–14, 2002.

[32] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani, "TwitIE : An Open-Source Information Extraction Pipeline for Microblog Text," 2013.

[33] B. Truong, C. Caragea, A. Squicciarini, and A. H. Tapia, "Identifying Valuable Information from Twitter During Natural Disasters," 2014.

[34] Y. Seki, "Sentence Extraction by tf / idf and Position Weighting from Newspaper Articles," 2003.

[35] N. K. Nagwani, "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm," vol. 17, no. 2, pp. 36–40, 2011.

[36] H. Saggion, T. Poibeau, H. Saggion, T. Poibeau, A. Text, and S. Past, "Automatic Text Summarization : Past, Present and Future To cite this version : Automatic Text Summarization : Past, Present and Future," 2016.

[37] J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, "Automatic Text Summarization using a Machine Learning Approach," 2003.

[38] S. Eyassu, "Classifying Amharic News Text Using Self-Organizing Maps," 2000.

[39] T. N. Dao and T. Simpson, "Measuring Similarity between sentences," 1983.