Research Article

Sentiment Analysis Model for Afaan Oromoo Short Message Service Text: A Machine Learning Approach

Ashebir Hunegnaw [a],Million Meshesha[b], EndalewAlamir[c],BahiruShiferaw[d]

[a,c] Department of Management Information Systems, Mettu University, Mettu, Ethiopia
[b] Department of Information science Addis Ababa University, Addis Ababa, Ethiopia
[d] Department of Computer science and Engineering Adama Science and Technology University

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: Thisstudy attempts to design a two-step approach for Afaan Oromoo text sentiment classification model, clustering followed by classification algorithms. A total of 1597 data which is collected from Oromia Broadcasting Corporate (OBN)"8331SMS database"from three domains (i.e. news, entertainment and general service domain) is used to conduct the experiment. First, text preprocessing is undertaken so as to clean data and prepare raw data for further processing. Then, clustering algorithm is used to find natural grouping of the unlabeled Afaan Oromoo text opinion documents. K-means and Gaussian Mixture (GMM) clustering algorithms were tested. GMM performs better and is selected to obtain the specific group of Afaan Oromoo documents. The result obtained from the clustering algorithm is directly exported to a CSV file and prepared for classification tasks. Three supervised learning algorithms, includingNaïve Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) in each domain are used to classify the sentiment of short Afaan Oromoo text. The result shows that SVM out performs NB and KNN with an accuracy of 91.66%, 93.76% and 92.87% for news, entertainment and general service domain respectively. This is a promising result to design sentiment analysis for comments given in Afaan Oromoo.

Keywords: Afaan Oromoo, Sentiment analysis, SMS text, Machine Learning

1. Introduction

Afaan Oromoo is one of the Afro-Asiatic languages which is mostly grouped under Cushitic family. It is one of the major indigenous African languages that is widely spoken and used in most parts of Ethiopia and some parts of the neighboring countries like Kenya and Somalia [1] [2]. It is spoken by more than 50 million people in Ethiopia, Kenya, Somalia, and Egypt and is the 3rd largest language in Africa following Kiswahili and Hausa [3]. It is used as first language for Oromo people and a number of members of other ethnicities who are in contact with the Oromo's and speak it as a second language. The number of sentiments on the web is increasing from day to day. This is also true for Afaan Oromoo language.This increase the amount of information available on the web, which made it impossible to read and objectively judge overall user opinion and sentiments manually. Therefore, there is a need ofcomputerized sentiment analysis, systems that reviews and changes it into a usable form [4] [5]. Sentiment analysis is the research which analyzes people'sthoughts, opinions, attitude and feelings in favor of the objectslike product, services, institutions, persons, topics and their attributes.

Sentiment analysis, usually studied at three levels that can be from document level to feature level [6]. Coarse-level analysis mainly concerned with finding the sentiment score of the entire document whereas feature level concerned with attribute level. Sentence-level sentiment analysis comes in between these two.

The most commonly used techniques for conducting sentiment analysis are lexicon-based approach, rule based approach and machine learning (ML) techniques [7][8][9]. The lexicon approach is based on a rule-based classifier, it counts positive and negative terms in review based on the sentiment dictionary and classifies the document as positive if it contains more positive words than negative and vice versa, as well as neutral if the number of positive and negative terms counts to equal [10] [11].In rule based approach we group the documents together, decide on categories and formulate the rules that define those categories; these rules are actually query phrases. It is accurate for small document sets. Here the rules are written by the writers and results are always based on what writers define. Since the rules fully depend on what the writers write but defining rules can be tedious for large document sets [9].

ML is a scientific study of algorithm and statistical models which enable ah system to learn from past data other than explicit programming [12]. The purpose of ML technique is to develop an algorithm so thatit optimize the performance of the system using example data.It consists of supervised learning and unsupervised learning method. Supervised ML requires the extra effort to labelgroups and assign labels to the documents in training dataset. Acquiring large corpora of labeled data is often a time consuming and expensive process in large and dynamic text databases. The Internet contains vast stores of data that could be useful for sentiment classification, but there is not an obvious way to use this data without first hand annotating each sentence. Classification

algorithms include SVM,NB, KNN, etc. It has generated a model of the training data, used to classify new unlabeled documents automatically [8].

Text clustering is the process of categorize a set of unlabeled text documents in such a way that texts in the similar group (called a cluster) are more similar to each other than to these in other clusters. Text clustering, is used for organizing aenormous number of text documents into well-organized form [13] [14].

Currently the amount of Afaan Oromoo documents on the web rapidly increasing from time to time [5]. However, even if there are different Afaan Oromoo opinions that are given on the web; there are a very limited work done or it's at starting phase. It is very necessary to apply sentiment analysis in the domain, so as to extract useful information from opinion forwarded by people [15].Therefore, the goal of this study is to build a two-step approach for unlabeled Afaan Oromoo short SMS text as it takes advantages of both supervised and unsupervised learning techniques. Unsupervised learning is used to label the unlabeled opinion data where it minimizes the problem of manually annotating the large volume of data which is tedious, expensive, and error-prone as it requires skilled experts. The supervised learning technique uses labeled data set to construct classification model for opinionated Afaan Oromoo SMS text.

2. Related Works

Abreham[4]proposed opinion mining from Amharic entertainment texts by using three machine-learning algorithms at document-level.The researcher usedunigramtermsbyusing information gain (IG)featureselectiontechnique andallUnigramfeaturesasa feature. Selama[16]has conducted a research on document level sentiment mining for opinionated Amharic text in movies and newspaper domain using general and domain specific opinion terms.Lexica of sentiment terms were used to find and assign the first polarity value to the sentiment terms. All the words, which are not in the list of the dictionary, are ignored then the class will be unclassified.

Tulu [17] proposed feature level opinion mining model for Amharic language by employing manually crafted rules and lexicon.Mohammed [11] proposed designing a graph-based opinion mining model for opinionated text in English, Amharic and Afaan Oromoo languages. The proposed model extracts the summary of these opinions polarity from the corpus of opinion-oriented graph.

Wegderes[5] has studied feature based summarization of commented Afaan Oromoo news text. They used Lexicon and/or rule-based method for Sentiment Prediction/detection, which requires dictionary of words, or positive and negative seed list to predict opinion. Rule-based algorithm were applied to build the model used to detect aspect, predict sentiment polarity by cross checking the tokens with the lexical database and count the polarity of opinions available under each aspect to summarize aspect-based sentiment.

3. Materials and Methods

Data collection and preparation

The datasets (reviews) for conducting the experiment were collected from Oromia Broadcasting Network "8331SMS database". A total of 1597 opinions were used to conduct the experiment. Important preprocessing tasks such as text cleaning, normalization, tokenization, stop word removal, stemming, and feature extractions are performed using NLTK to clean the data and prepare for ML algorithms.

Development tool

Sklearn package is used for applying clustering and classification algorithms. Sklearnhave a lot of sufficient tooltobuild a machine learning and statistical model likeregression, classification, clustering and dimensionality reduction [18]. All programming has been done in the Python programming language. Because python is a more flexible and multipurpose programming language suitable to build machine learning model [19].

Machine Learning Algorithms

ClusteringAlgorithms

In this research two clustering algorithms, basic k-means algorithm and Gaussian Mixture model are applied to cluster the Afaan Oromoo SMS text polarity

K-means clustering is the best and known clustering algorithm due to its simplicity and efficiency. K-means partitioning or group given data set into K clusters where the value of K (i.e. Number of clusters) are defined by the user [20]. It is easy to understand. The performance of K-means depends on itsinitial centroid. Because it does not guarantee for an optimal solution. Due to its low computational requirements K-means clustering algorithm has been recognized to bebetterto handlehuge document datasets than hierarchical clustering algorithms [21].

The K-means algorithm, works as follows. Given a set of data objects D with a pre-defined number of clusters k, then, k data objects are randomly selected to initialize k clusters, each one being the centroid of a cluster. A centroid represents imaginary or reallocation representing the center of the cluster.The remaining objects are then allocated to the cluster represented by the nearest or most similar centroid. Next, new centroids are re-computed for each cluster and in turn all documents are re-assigned based on the new centroids. This step iterates until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids [22].

GMM, is an extended the ideas the k-means algorithm. It is also a powerful tool to estimatenumber of clusters in the same way as k-means in complex text documents [23].GMM is based on the probability density function f(x) which can be approximated toany degree of accuracy with the Gaussian probability density.

Classification Algorithms

In this study three popular supervised learning algorithms are tested, such as NB classifier, SVM and KNN. NB classifier is one of the well-known and simple probabilistic classifier. It is based on applying Bayes' theorem with strong independence assumption. The Multinomial NB classifier is suitable for text classification where it assumes that features are drawn from simple multinomial distribution. The parameters used in this study are alpha with value 1.0, fit_priori, True and class_priori, none.

The main idea of SVM algorithm is to fit a linear model to maptraining datasetthrough maximizing the margin of the algorithm. Itrepresents the value of the given parameter of hyper plane to the closest training patterns, given classes is maximized as much training patterns as possible [24]. The parameters used to implement SVM in this study are C which is the regularization parameter, C, of the error term, kernel type, degree of polynomial kernel and coefficient (gamma) for kernel is set first and then the model is created.

KNN is an effective and powerful classification and regression algorithm because it does not assume anything about the data, other than a distance measure can be calculated consistently between two instances. As such, it is called non-parametric or non-linear as it does not assume a functional form [28]. It works based on the minimum distance from the query instance to the training samples to determine the K-nearest neighbors. After gathering K nearest neighbors, it take a simple majority of these K-nearest neighbors to be the prediction of the query instance. Objects are categorized based on closest feature space in the training set. Euclidean Distance is typically used in computing the distance. The parameters includen-neighbor=5, weights=uniform, algorithm to compute the nearest neighbor, leaf_size=30, metric, metric_params and n_jobs=1.

Evaluation techniques

The performance of clustering algorithm is evaluated by using silhouette coefficient. The silhouette coefficient measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouettecoefficient ranges from −1 to 1, where a high value shows the object is well matched to its own cluster and low value shows it poorly matched to neighboring clusters [25]. The performance of the classification model is measured by classification metrics including accuracy, precision, recall and f-measure. In addition to these a test set of corpus of sentiments is prepared and classified in the correct classes as positive, negative and neutral.

4. TheProposed Architecture

The architecture of sentiment analysis for Afaan Oromoo SMS text is given in figurebelow.

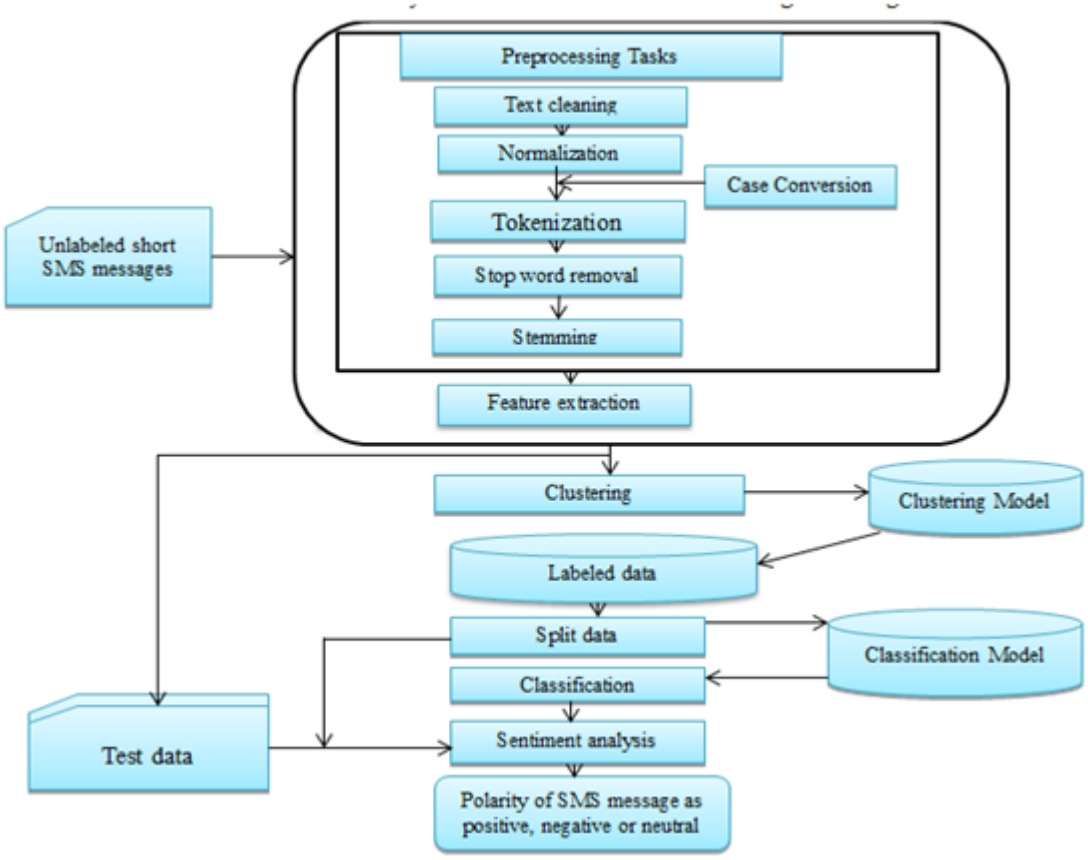

Figure 4. 1 The proposed architecture for sentiment analysis of Afaan Oromoo text

Tokenization

Tokenization, is the process of breaking textdocument into words, phrases, symbols and other meaningful elements called token.In natural language strings were broken into words, digits and punctuations. Thentokens were generated from tokenized strings. It was performed by using white space.

Algorithm 4.1 Algorithm toRemoveSpecialCharactersand Numbers

Open file containing corpus

Split documents at word level

For a word in wordlist

If a word is in r"[^a-zA-Z']+",' ',

Remove unwanted character

end if

end for

returnlist of cleaned text

close file

End algorithm

Stop word Removal

There is no structured stop word list prepared for Afaan Oromoo, SMS text document. Hence, Researchers manually prepared Afaan OromooSMS text document stop words lists depending on corpus using Afaan Oromoo dictionaries.These words are language specific words which carry no information like prepositions, conjunctions, articles, and particles. Some examples of Afaan Oromoo stop words are sun, kun, ani, etc.

Algorithm 4.2 Algorithm to Remove Stop Words

Open a file that contains stop word list

Read a sentence from the corpus

For each word in the corpus

If a word in list of stop words

Remove a word

else

Move to non-stop word list

end if

end for

return non-stop word listsEnd algorithm

Stemming

In Afaan Oromoo SMS text document, one word appears in different forms to refer singular or plural, and to show tense which should have one root or stem. According to Kekeba [26] suffixes are the predominant morphological features in Afaan Oromoo language. For the above reasons, in this study we have focused on Afaan Oromoo suffixes. We have removed Afaan Oromoo SMS text postfixes as follows. Postfixes of each word in Afaan Oromoo text documents corpus were identified. The length of postfixes to be removed from root words were decided by language expert. Then we wrote a code by using python programming language depending on the identified postfixes and length the postfixes.

Algorithm 4.3 stemming

Open corpus file

While not end of file

For each word in corpus

If word length≥ 3

If word end with suffix

Remove suffix

Return root_word

end if end if end for

end while

close file

Feature extraction

Feature extraction was applied after text preprocessing of Afaan OromooSMS text documents. Term weighting is used to represent documents. The term frequency is computed from number of times Afaan Oromoo SMS text word w terms found in the document d. Term weighting computed to decides the degree of importance of a given term to a given document.TF_IDF term weighting technique is used in this study. It is calculated as

$$TF*IDF(w,d,D)=TF(w,d)\times IDF(w,D)$$

Where w denotes the terms; d denotes each document; D denotes the collection of documents.

5. Result and Discussion

Two clustering algorithms; k-means and GMM clustering were applied after performing preprocessing steps. Clustering algorithm that achieved the best result is considered for labeling the SMS messagesand ready for the classification task. Then the three classification algorithms, NB, SVM and KNN were tested for short Afaan Oromoo SMS messages text classification. The documents were further divided into training dataset and test

dataset where 80% of dataset used for training and 20% the dataset used for testingfrom a total 1597 amount of data set.

Evaluation of Clustering Algorithms

Clustering algorithms is done through grouping the text into their polarity; dictionary containing document with their corresponding cluster number is created. After creating this dictionary it is converted into a data frame using panda's library. Then, the path where the data should be stored is specified with their file extension.Then the labeled data is ready for experimenting supervised learning algorithms.

Result of K-Means

The result of K-Means clustering for news domain shows that out of 602 total data set 106 texts were labeled as 0, 409 of them were labeled as 1 and 106 of them were labeled as 2. The result is exported as csv file. Then the domain expert evaluate the result of k-means clustering for news domain and labeled 0 as negative, 1 as positive and 2 as the neutral polarity. For entertainment domain out of 582 total data set 122 texts were labeled as 0, 134 of them were labeled as 1 and 334 of them were labeled as 2and exported to csv file. Domain expert evaluate the result of k-means clustering for entertainment domain and labeled 0 as negative, 1 positive polarity and 2 neutral polarity accordingly. General service domain shows that out of 416 total data set 305 texts were labeled as 0, 67 of them were labeled as 1 and 44 of them were labeled as 2. Then the domain expert evaluate the result of k-means clustering for general service domain and labeled 0 as positive, 1 negative polarity and 2 neutral polarity. The result of the k-means clustering for all domainsare shown in figure 5.1.

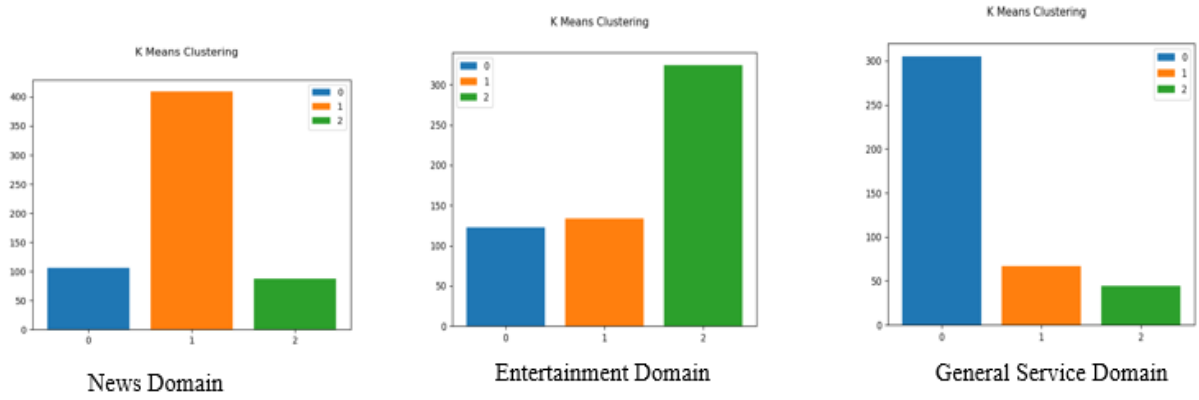

Figure 5.1. K-means clusteringAlgorithm Result

Clustering Result of Gaussian Mixture Model

The experimental result of Gaussian Mixture Model for news domain shows that out of the total 602 texts, 221 texts were labeled as 0, 202 of them were labeled as 1 and 179 of them were labeled as 2. Similarly the result of Gaussian Mixture Model is exported to csv file. Then the domain expert evaluate the result of GMM clustering for news domain and labeled 0 as negative polarity, 1 as positive polarity and 2 as neutral polarity. In theentertainment domain out of the total 582 texts, 144 texts were labeled as 0, 179 of them were labeled as 1 and 258 of them were labeled as 2. The result of Gaussian Mixture Model is exported to csv file and the domain expert evaluate the result of GMM clustering for entertainment domain and labeled 0 as negative polarity, 1 as neutral polarity and 2 as positive polarity.For general service domain out of the total 416 texts, 88 texts were labeled as 0, 124 of them were labeled as 1 and 204 of them were labeled as 2. The result of Gaussian Mixture Model is exported to csv file. And domain expert evaluate the result of GMM clustering for general service domain and labeled 0 as negative polarity, 1 as positive polarity and 2 as neutral polarity. The result of the GMM clustering for all domains are shown in figure 5.2.

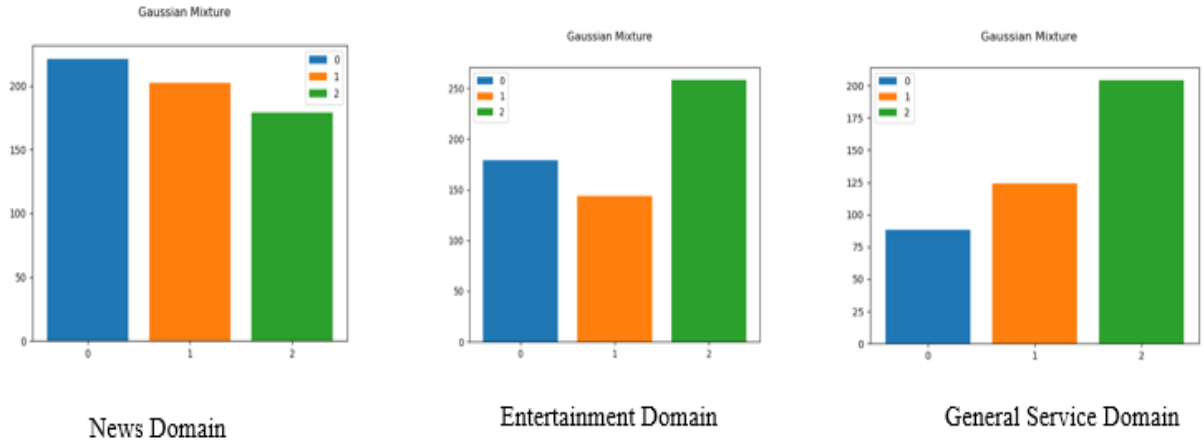

Figure 5.2. Gaussian Mixture Model Algorithm Result

Comparison of clustering algorithms

Table 5.1 Comparison of clustering algorithms in terms of Silhouette Values

Clustering Algorithm	Domain		
	News	Entertainment	General Service
K-Means	0.0909	0.3606	0.087
GMM	0.792	0.546	0.525

Since GMMclustering perform better result than K-Means, the result of GMM is directly exported to .csv file with their corresponding label as discussed in section 5.1. Having this result we have proceed to classification task.

Evaluation of Classification Algorithms

The input file now contains three attributes, text data (i.e. Commented opinions) and the class label of the comment, which is neutral, positive and negative polarity.

Classification Result of Naïve Bayes Algorithm

The experimental resultofNB algorithm in each domain is shown in table 5.2.

Table 5. 2 Experimental Result of NB for each Domain

	Precision	**Recall**	**F1-Score**	
	News Domain			
Positive	0.92	0.94	0.92	0.928
Neutral	0.94	0.93	0.93	
Negative	0.95	0.89	0.91	
	Entertainment Domain			
Positive	0.95	0.97	0.96	0.920
Neutral	0.89	0.87	0.89	
Negative	0.94	0.96	0.95	
	General Service Domain			
Positive	0.94	0.86	0.92	0.873
Neutral	0.89	0.84	0.87	
Negative	0.86	0.95	0.91	

Classification Result of SVM

The experimental result of SVM algorithm in each domain is shown in table 5.3.

Table 5. 3 Experimental Result of SVM for each Domain

	Precision	**Recall**	**F1-Score**	**Accuracy**
	News Domain			
Positive	0.91	0.99	0.94	0.916
Neutral	0.97	0.91	0.93	
Negative	0.93	0.94	0.92	
	Entertainment Domain			
Positive	0.99	0.96	0.97	0.937
Neutral	0.93	0.92	0.92	
Negative	0.95	0.94	0.93	
	General Service Domain			0.928
Positive	0.99	0.95	0.97	
Neutral	0.89	0.85	0.87	
Negative	0.91	0.95	0.93	

Classification Result of K-Nearest Neighbor

The experimental result of SVM algorithm in each domain is shown in table 5.4.

Table 5. 4 Experimental Result of KNN for each Domain

	Precision	**Recall**	**F1-Score**	**Accuracy**
	News Domain			
Positive	0.84	0.94	0.89	
Neutral	0.95	0.90	0.92	0.885
Negative	0.89	0.91	0.90	
	Entertainment Domain			
Positive	0.97	0.88	0.94	
Neutral	0.81	0.81	0.81	0.877
Negative	0.85	0.93	0.89	
	General Service Domain			
Positive	0.79	0.97	0.87	
Neutral	0.85	0.93	0.88	0.838
Negative	0.98	0.88	0.90	

Table 5. 5 Comparison of the Performance of Classification Algorithms

Algorithm Name	**Naïve Bayes**			**Support Vector Machine**			**K-Nearest Neighbor**		
	News	**Entertainment**	**General Service**	**News**	**Entertainment**	**General Service**	**News**	**Entertainment**	**General Service**
Accuracy (%)	90.83	92.05	87.38	91.66	93.76	92.87	88.53	87.72	83.80
Weighted Average precision (%)	93.6	92.7	89.66	93.66	95.66	93	89.3	87.7	87.3
Weighted Average recall (%)	92	93.3	88.3	94.6	94	91.7	91.7	87.3	92.6
Weighted Average F-measure	92.3	93.3	90	93	94	92.3	90.3	88	88.3
Duration (minute)	00.41	00.33	00.39	00.33	00.31	00.33	00.34	00.36	0.37

Discussion of the Result

In In this study, three supervised learning algorithms (NB, SVM and KNN) were used to classify the sentiment of Afaan Oromoo text document in news, entertainment and general service domain. The comparison of all three algorithm results in terms of their accuracy and CPU run time is shown below in table 5.5.SVM is the best algorithm that achieved the highest accuracy of 91.66%, 93.76% and 92.87% for news, entertainment and general service domain respectively as compared to NB and KNN. So the model built by SVM is selected as the best algorithm in this study for sentiment analysis. In order to test performance of the selected model,120 new datasetwere prepared which includes equal number of positive, negative and neutral comments from news, entertainment and general service domain. The datasetwere labeled by domainexpert. After building and selecting the model it should be saved to predict the new data. Application program interface (API) was created using Flask(i.e. the Python micro framework for building web applications)to test performance of the model. Flask API shows the predictive capabilities through HTTP request. HTML form is created and POST method is used to transport the data to be predicted from HTML form to server. Finally Flask's debugger activated by running the app and make predictions on the saved SVM model. 40 comments were used for each class randomly to test the performance of the SVM model. The performance of the SVM model was shown in table 5.6 below.

Table 5. 6 Confusion Matrix of the Selected SVM Model

	Positive	**Neutral**	**Negative**
Positive	37	2	1
Neutral	3	36	1
Negative	2	1	37

As shown in the confusion matrix, some positive comments were labeled as negative and/or neutral and vice versa. The first problem observed was people sometimes use positive words in negative comment, but the word is preceded or succeeded by any other negative word in Afaan Oromoo such as "hinjaalanne" which means I didn't like it. A human would easily detect the true sentiment of the review, but it is difficult to machine. There are some comments which are classified out of their polarity because as negation did not handled in this work.In addition to this, we have seen that all algorithms used in this research achieved less performance in general service domain when compared with news and entertainment domain. Because as the comments are given on general services of the organization like live streams, video quality, sound quality and other issues, the machine learning techniques are challenged to mis-cluster and mis-classify. That is happening because of machine learning is domain dependent [27].

6. Conclusion and Future Work

Internet technology has changed the way people express their views and opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc. The explosion of the World Wide Web provide an increasing amount of data and information coming from diverse sources. Due to this, sentiment analysis becomes more difficult for users to select and convert data and information into useful knowledge. Supporting the target people to organize the vast and large volume of information is became a critical issues. Therefore, there is a need for automated sentiment analysis systems that reviews and changes it into a usable form. Many researchers have been proposed to analyze, document polarity with different machine learning approaches and good results were obtained.

In this study, unlabeled text documents collected from OBN "8331 SMS database" was used for classifying the commented documents as positive, negative or neutral. A two-step approach was applied to classify the Afaan Oromoo documents into their respective polarity. K-means clustering and GMMalgorithms was applied to label or group the Afaan Oromoo text document. The experimental result shows that GMM clustering algorithm registered better performance. The result obtained from clustering was directly exported to csv file with their corresponding class label that gained from clustering. The result obtained from clustering is used to classify the Afaan Oromoo text document using three supervised learning algorithms (NB, SVM and KNN).Experimental result shows that SVM registered best accuracy of 91.66%, 93.76% and 92.87% for news, entertainment and general service domain respectively. Hence, because of the promising result obtained we selected SVM algorithm for classification. Text classification using unigram method is considered in this study.To extend this research in the future bigram techniques will be considered for overcoming the challenge of negation.

References

1. T. Guta, "Afaan Oromo Search Engine," Unpublished Msc thesis, Addis Ababa University, Addis Ababa, 2010.

2. B. G. Erena, "Afaan Oromoo," 25 March 2018. [Online]. Available: https://scholar.harvard.edu/erena/oromo-language-afaan-oromoo. [Accessed 14 June 2019].
3. G. G. Eggi, "Afaan Oromo Text Retrieval System," Unpublished Msc thesis, Addis Ababa University, Addis Ababa, 2012.
4. Getachew, "Opinion Mining From Amharic Entertainment Texts," Unpublished Msc. Thesis, Addis Ababa University, Addis Ababa, 2015.
5. W. Tariku, "Aspect Based Summarization of Opinionated Afaan Oromoo News Tex," Unpablished, Debre Berhan, 2016.
6. Suchdev R., Pallavi K., Rahul R., and Sridhar S., "Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach," International Journal of Computer Applications, vol. 103, no. 4, p. 0975 – 8887, 2014.
7. Yusof, N.N., Mohamed, A. and Abdul-Rahman, S., "Reviewing classification approaches in sentiment analysis," in International conference on soft computing in data science, Springer, Singapore, 2015.
8. Bhuvaneswari, K., and Parimala R., "Sentiment Reviews Classification using Hybrid Feature Selection," Int. J. Database Theory Appl, vol. 10, no. 7, pp. 1-12, 2017.
9. D. Kalita, "Supervised and Unsupervised Document Classification: A Survey," International Journal of Computer Science and Information Technologies, vol. 6, no. 2, pp. 1971-1974, 2015.
10. Yang, Min, Tu W., Lu Z., Yin W., and Chow K. P., "LCCT: A semi-supervised model for sentiment classification," in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, 2015.
11. M. Tune, "Designing a graph based opinion mining model for opinionated text in English, Amharic and Afaan Oromo Languages," Unpublished Msc thesis, Debre Berhan university, Debre Berhan , 2013.
12. Hurwitz, J., and Kirsch D., Machine Learning For Dummies®, IBM Limited Edition 75: J. Wiley, 2018.
13. Grace, G. Hannah, and Kalyani D., "Experimental estimation of number of clusters based on cluster quality.," arXiv preprint arXiv.1503.03168, 2015.
14. V. Goel, "/applying-machine-learning-to-classify-an-unsupervised-text-document-e7bb6265f52," Medium, 2 November 2018. [Online]. Available: https://towardsdatascience.com/applying-machine-learning-to-classify-an-unsupervised-text-document-e7bb6265f52. [Accessed 15 August 2019].
15. Kharde, Vishal, and Prof Sonawane, "Sentiment analysis of twitter data: a survey of techniques," International Journal of Computer Applications, vol. 139, no. 11, p. 0975 – 8887, 2016.
16. S. Gebremeskel, "sentiment analysis for opinionated Amharic text," Unpublished Msc. Thesis, Addis Ababa university, Addis Ababa, 2010.
17. T. Tilahun, "Opinion mining from Amharic blog," Unpublished Msc. Thesis, Addis Ababa University, Addis Ababa, 2013.
18. K. Jain, "scikit-learn-python-machine-learning-tool," Analytics Vidhya, 5 January 2015. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/. [Accessed 1 October 2019].
19. Tagliaferri L., Morales M., and Birbeck E., "Python Machine Learning Projects," in Python Machine Learning Projects, New York, USA, DigitalOcean, 2015, p. 7.
20. Unnisa, Muqtar, Ayesha Ameen, and Syed Raziuddin, "Opinion mining on Twitter data using unsupervised learning technique," International Journal of Computer Applications , vol. 148, no. 12, pp. 975-8887, 2016.
21. Ortigosa-Hernández, J., Rodríguez, J.D., Alzate, L., Lucania, M., Inza, I. and Lozano, "Approaching SentimentAnalysisbyusingsemi-supervisedlearning," Neurocomputing ., vol. 92, no. 1, pp. 98-115, 2012.
22. Huang, "Similarity measures for text document clustering," Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, vol. 4, pp. 9-56, 2008.
23. J. VanderPlas, "05.12-gaussian-mixtures.html," Python Data Science Handbook, 09 January 2018. [Online]. Available: https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html. [Accessed 01 October 2019].
24. Wondwossen Philemon and Wondwossen Mulugeta, "A Machine Learning Approach To Multi-Scale Sentiment Analysis Of Amharic Online Posts," HiLCoE Journal of Computer Science and Technology, vol. 2, no. 2, pp. 80-87, 2014.
25. MediaWiki, "Silhouette_(clustering)," Wikipedia, 09 August 2019. [Online]. Available: https://en.wikipedia.org/wiki/Silhouette_(clustering). [Accessed 21 September 2019].
26. Tune, Kula Kekeba, and Vasudeva Varma, "Oromo-English Information Retrieval Experiments at CLEF," Working Notes of Cross Language Evaluation Forum Workshop 2006, Spain, 2007.
27. S. Schrauwen, "Machine learning approaches to sentiment analysis using the dutch netlog corpus," Computational Linguistics and Psycholinguistics Research Center, Antwerp, Belgium, 2010.

28. deparkes, "machine-learning-vs-rules-system," 24 10 2017. [Online]. Available: https://deparkes.co.uk/2017/11/24/machine-learning-vs-rules-systems/. [Accessed 12 6 2019]..