Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network

Surafel Getachew Tesfaye (surafel.getachew@aastu.edu.et)
Addis Ababa Scince and Technology univerisity https://orcid.org/0000-0001-6933-3442
Kula Kakeba
Addis Ababa Science and Technology University

Research

Keywords: Hate speech, Amharic posts, Recurrent Neural Network, Deep Learning, Long-Short Term Memory, Gated Recurrent Unit, Hyper-parameters

Posted Date: December 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-114533/v1

RESEARCH

Automated Amharic Hate speech Posts and Comments Detection Model using Recurrent Neural Network

Surafel Getachew Tesfaye* and Kula Kekeba Tune

*Correspondence:
surafel.getachew@aastu.edu.et
Department of Software Engineering, Big Data and HPC CoE,Addis Ababa Science and Technology University, Addis Ababa, Ethiopia
Full list of author information is available at the end of the article

Abstract

During the last few years, social activities over the internet especially on social media platforms increased drastically, but unfortunately, social networks have also become the place for hate speech proliferation by which most people's social lives are disturbed because of hate speech posts and conflicts triggered by those posts. Studies confirm that online hate speech has different offline consequences. Even though there are a lot of researches on automated hate speech detection most of them are for other language and there is a scarcity of labeled data to apply automated analysis and detection methods on Amharic dataset. Therefore the research on automatic detection of hate speech posts attracted our attention. As a solution to those problems, this research aimed to prepare a labeled huge Amharic dataset by collecting posts and comments from selected Facebook pages of activists that participated actively. Those Facebook data sets are labeled manually as hate and free based on the guidelines given from researcher and pre-processed by applying data cleaning and normalization techniques. In this research the recurrent neural network models for automated hate speech posts detection from Amharic posts on Facebook is developed by using Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) with word n-grams for feature extraction and word2vec to represent each unique word by vector representation. The experiment conducted on those two models by using 80% of the data set for training and 10% for validation to train the model and to select the best hyper-parameters combination for automated hate speech posts detection. The remaining 10% of the dataset used for testing the model after training. As a result LSTM based RNN of Batch size 128, and learning rate 0.001 with RMSProp optimizer and 0.5 dropout achieves an accuracy of 97.9% to detect posts as hate speech or free by training with 100 epochs. Which is assured by testing the models using models performance test and inference on user-generated data.

Keywords: Hate speech; Amharic posts; Recurrent Neural Network; Deep Learning; Long-Short Term Memory; Gated Recurrent Unit; Hyper-parameters

Introduction

There are now 5.15 billion unique mobile phone users in the world, and there are 3.96 billion social media users in the world totally among them 21.14 million users are from Ethiopia, which shows the rapid development of social networking websites and their usability by users from all over the world to easily expressing an opinion and communicating each other. Nowadays it is easy to spread individuals' idea within the population using a different languages through which the idea can reach

to the desired web technology users, resulting in a huge amount of user-generated data available for an enormous online audience. Those opportunities are widely used to express hateful statements to large groups or specific individuals with the malicious intention [1]. And also Facebook is a widely used social media site by most of the community members according to stat counter global status data from April 2019 till April 2020 from the overall 60.15% of social media usage is held by Facebook [2].

Facebook users in Ethiopia use the Amharic language to trigger deadly ethnic clashes among a population of 107 million through ugly Facebook content, which is a language supported by Facebook [3]. Amharic language is widely spoken language which is the working language of Ethiopia. It has 275 characters mainly consonant-vowel pair and it is the second-largest Semitic language. While using Facebook Ethiopian users mostly use Geez scripts to share more essential posts and to communicate through comment and replies, even though Amharic texts can be written by using the Latin alphabets [4].

The explosive growth in hate speech and its erosion to democracy, justice, peacebuilding, and public trust has increased the demand for automated hate speech detection and technological intervention because hate speech posts are more influential in Ethiopian communities as individuals become more sensitive to something which they think is mine due to that they follow posts about those things and participate through comments and share. Those posts shared by different users deliberately or unknowingly create a bad feeling on some side and which became a cause for conflicts, that is why the government of Ethiopia mostly blocking social media sites while rebellion arises to minimize their effect [5]. So that having organized knowledge about those hate speech posts to take some measures by the government need the application of automated hate speech posts detection based on the Amharic language because of the huge amount and unstructured data used in social media are difficult to analyze manually to solve those type of problems. Therefore, we will develop the model which can detect hate speech posts by applying deep learning techniques and algorithms.

Literature Review

Hate and Offensive speeches

In hate and offensive speech detection, there is no single definition that can agree everybody, due to this it became a topic to hotly debate by experts [6]. There are Legal definitions of hate speech from governments of different countries, by which governments are increasingly defining hate speech in their criminal codes in an attempt to directly regulate harmful rhetoric both on and offline [7]. The definition of hate speech varies from perspectives of different sources but the most important definition in our case is the definition given by Facebook and Ethiopian Hate speech law, those are:-

- Ethiopian Hate Speech and Disinformation Prevention and Suppression Proclamation Page 12339 under Proclamation No. 1185 /2020 “Hate speech” is the speech that intentionally promotes discrimination, hatred, or attack against a discernable group of identity or person, based on race, ethnicity, gender, religion or disability”.

- Facebook defines hate speech as "Objectionable content that directly attacks the people based on their protected characteristics such as religious affiliation, sexual orientation, caste, sex, race, ethnicity, national origin, gender, gender identity, and serious disease or disability". Generally, Facebook defines an attack as dehumanizing or violent speech, calls for exclusion statements of inferiority or segregation.

In addition to hate speech, there is language or expression that offends and negatively characterizes an individual or group of people. This kind of speech causes someone to feel upset, annoyed, insulating, angry, hurt, and disgusting [8].

The difference between hate speech and offensive language often based upon understated linguistic distinctions [9]. Offensive speech occurs when there is a low degree of hate speech characterizations occur. The speech may contain a target but do not directly incite violence, attack or diminish based on the target group characters because people often use offensive language to make a point in the debate, on heated conversation, and condemn another violent act performed by other people. Any speech that contains a sarcastic, mocking, and joke that offend can be considered as an offensive if it conducts using language that contains high negative, abusive, dirty words or phrases in both oral and text.

Amharic Language Processing

Natural language refers to the language spoken or written by human beings, as opposed to Artificial languages, computer understandable language, mathematical, or logical languages [10]. It is so important because it helps us to build hate speech posts detection models by processing the linguistic data sets by which it takes chunks of information as input in form of voice or text or both to manipulate them as per the algorithm of machine learning inside the computer. Thus the input speech, text, or image is going to be transformed into the output of an NLP.

Amharic language is a Semitic family language which is written from left-to-right by using unique geez script. It is one of the sub-Saharan countries Ethiopian's working language. Amharic lacks capitalization and in total has 275 alphabets mainly consonant-vowel pairs and it has its writing script known as "Fidel" having 34 consonants (base characters) and other six characters represented for each base characters formed from the combinations of vowel and consonant of base characters [11].

The Amharic language is one of under-resourced language which doesn't have more language processing tool and techniques like other languages have which could help computational linguistic researchers to go more detail and develop useful higher-level Internet or computer-based applications and models but, there are researches on Amharic language text mining. The Amharic writing system is more challenging than English to detect the intention of the text and to characterize, categorize or classify it because of factors like the redundancy of characters which has a similar sound. people use such characters for semantically same words interchangeably, writing mechanism of compound words using hyphen or space and sometimes they may merge but they have different lexicon when they put separately, spelling

variation of same words which can be written using different letters but pronounced as the same and inconsistency showed in abbreviating Amharic words and there are many ways of writing words especially loan words, words that are taken from other languages.[11] [12]

Hate Speech Detection Techniques

Hate speech detection problem has been studied by a different researcher, and they used different techniques to detect hate speech propagation on social media and other online web platforms. Just as there is no clear agreement on the definition of hate speech, there is no consensus concerning the most effective methods to detect on social media platforms.

Machine Learning for Hate Speech detection

The ability to learn from past observations and the increase in data volume, variety, and velocity leads to automation in text processing techniques which leads to machine learning which is a subpart of an Artificial Intelligence (AI) that helps systems to automatically learn and improve based on the training dataset without being explicitly programmed to do a specific task.

Machine learning approaches can be categorized into supervised, unsupervised, and semi-supervised approaches. But, the most common approach to be used in hate speech detection is the supervised method by which machine learning algorithms will apply what they have been learned using the training dataset to the new data to predict the desired output. This approach is domain-dependent because it depends on manual annotation of a huge volume of text for instant most of the research work reviewed for this research uses a classifier algorithm such as support vector machine (SVM), Random Forest (RF), Naïve Bayes (NB), Logistic regression (LR) and Deep learning method. Most of the research studies for automatic hate speech detection uses more than two classification algorithms for computational comparison reason and used to suggest which algorithm has high performance and accuracy for their proposed detection model [13] [14] [15].

A lot of researches on hate speech are being undertaken for English, other languages , and also for Amharic, Zwedie and Wang developed a model to detect Amharic text hate speech using Naïve Bayes and random forest machine learning techniques with word2vec and TF-IDF feature extraction methods using medium-sized data set and achieved an accuracy of 79.8% by which they classify the speeches as hate or not [5] , and also Yonas [8] studied about hate speech detection on Amharic posts on Facebook using three machine learning algorithms which are SVM, Naïve Bayes, and random forest by collecting 5000 datasets from posts of activists and most followed pages and word unigram, bigram, n-gram, TF-IDF and word2vec feature extraction using 5 fold cross-validation to achieve an accuracy of 79% to classify the speech as hate or offensive.

Hate Speech Detection Techniques

Hate speech detection problem has been studied by a different researcher, and they used different techniques to detect hate speech propagation on social media and other online web platforms. Just as there is no clear agreement on the definition of hate speech, there is no consensus concerning the most effective methods to detect on social media platforms.

Deep Learning for Hate Speech detection

Deep learning is a machine learning mechanism that helps computers to learn and to do what humans naturally could do. Like machine learning algorithms they can be learned through supervised, unsupervised, or semi-supervised ways. Nowadays, Deep Learning models shows excellence on analytics to be done on text and hate speech detection tasks specifically because of its dependency on the neural network classifiers with deep knowledge. It attempt to learn in a real sense to identify patterns in the provided text and tries to emulate the event in layers of neurons. The performance of deep learning models depends upon the best choice of neural network algorithm and its hyper-parameters as well as techniques to represent features. Some research works like [16] [17] [18] propose a deep learning neural networks based hate-speech text detection using Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), and Long short-term memory (LSTM) respectively, and also, Zwedie and Wang identify vulnerable community based on their ethnic identity using their hate speech detection technique to ban contents discussed already identified community by identifying the most hatred ethnic group among Amhara, Tigre and Oromo by utilizing classical and neural network machine learning libraries in python by preparing the word embedding from posts and comments extracted from selected Facebook pages for word2vec feature extraction by 300 embedding dimensions and using RNN neural network with 128 hidden layers, output dimension size of 50, 100 epochs of training and RMSProp optimizer [19].

The new deep neural network-based method was used to detect hate speech by Zang and Luo. To capture implicit features that are potentially useful for classification this neural network is proposed, and their methods were evaluated on the largest collection of Twitter datasets for hate speech detection, to show that they can be particularly effective in detecting and classifying hateful content (as opposed to non-hate), their results set a new benchmarking reference in this area of research [19].

RNN can be thought of as the addition of loops to the architecture through backpropagation in the training process, to update the network weights in every layer [21]. According to [24] the recurrent neural networks are a useful tool for hate speech detection because of their ability to remember the sequence of inputs by considering their order differs from the feed-forward model. They implement the model by combining both sequences of word embedding and social media features using Keras by fed sequence of word embedding to a recurrent layer whose output is concatenated with social features and also Mossie and Wang use the RNN variants such as, LSTM, Bidirectional-LSTM, Gated Recurrent Unit (GRU), and Bidirectional GRU models will be tested and evaluated for social media dark side

content detection using transfer learning emphasis on hate and conflict and these variants perform well in both short and long texts compared with CNN [17] [20].

LSTM and GRU

LSTM is a special type of recurrent neural network that addresses problems by conserving long term dependency more expertly in comparison to the basic recurrent neural network. To overcoming the vanishing gradient problem LSTM is particularly useful. Although it has a chain-like structure similar to RNN, LSTM uses multiple gates to carefully regulate the amount of information that is allowed into each node state. LSTM helps to compute the representation of a document from the representation of its words, with multiple abstraction levels. Each input word is represented by a low dimensional which is a 262- dimensional, continuous, and real-valued vector for the training of LSTM. In addition to LSTM, there is a bidirectional LSTM architecture that allows capturing long-range dependencies from both directions of a document by constructing bidirectional links in the network [21] [22]. Generally, LSTM can retain the knowledge of earlier states to be trained for tasks that require large memory or state awareness of tasks due to this reason it can overcome the limitation of RNN because it consists block of memory cells state by which signal flows by regulating input, forget and output gates to control what is stored, read and written on cell. LSTM is used by Google, Apple, and Amazon [23].

Gated Recurrent Unit (GRU) is a simplified variant of the LSTM architecture, which is a gating mechanism for recurrent neural network, but there are differences as follows: GRU contains two gates which are update gate and forget gate. Those gates does not possess any internal memory, and non-linearity is not applied. GRU is now popular in the community who is working with recurrent networks. The main reason for the popularity is the computation cost and simplicity of the model. The researchers used GRU to detect hate speech in the Indonesian language by using it to solve the vanishing gradient problem which comes with the standard recurrent neural network [24]. Generally, GRU combines the forget gate and input gate into one update gate and has an additional reset gate and it is increasingly popular for NLP [24].

Methodology

To develop hate speech posts detection models using a deep learning approach, the research began with the literature review covering the traditions that approached online hate speeches from the complementary perspectives, including the legal literature that studies how hate speech is addressed in different continents and countries specifically on social Medias. Our research flows depicts in Figure below.

Data-set Preparation

In this study, to detect hate speech texts from Amharic posts and comments on Facebook a new dataset is built because of the lack of published Amharic Facebook

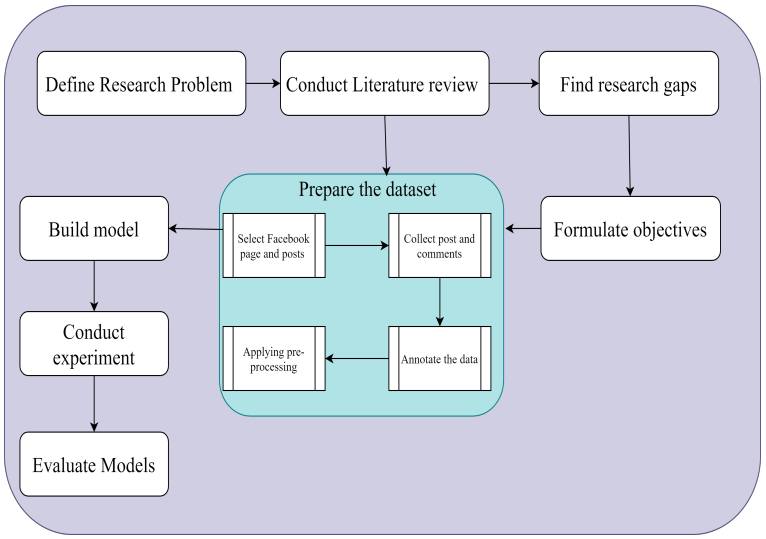

Figure 1 Research flow diagram

dataset. Even though there is publicly available dataset those are for the English language, so that to achieve the goal of building the data set the following tasks are performed.

1 Mostly followed pages of activists and news pages are selected because, over the previous years, speeches by activists in Ethiopia have spread rapidly across social media .
2 The posts and comments were collected from those pages manually because Facebook blocks data collection tools like netvizz.
3 Label those posts and comments collected from those pages as hate speech and free speech.

Facebook pages of activists, news pages, and broadcasting pages of channels are screened out because these pages typically post discussions spanning across a variety of political and religious topics. It's common for the social network communities to commonly posting on political and religious issues, Then the data collected from their posts and the comments of their followers [5] Based on the following metrics:-

- Facebook pages that have more than 50,000 like based on the study of Yonas, could help us to get more active public pages [8].
- Activists without public page but will have posts with more than 300 comments per post because it shows they have a lot of active participants for their posts.

Data Cleaning

We cleaned our data set by removing irrelevant characters, punctuation marks like ārati netibi , netela šerezi, diribi šerez and question mark (?), emojis, blank values, and URL.

Normalization

In Amharic language there are characters which represent the same consonants with the same pronunciation and it can be used interchangeably without any meaning

Sound characters	Normalized to
ሀ, ኀ, ሐ, ኃ, ኻ, ሓ, ሃ	ሀ
ሰ, ሠ	ሰ
ጸ, ፀ	ፀ
አ, ኣ, ዐ, ዓ	አ

Figure 2 Amharic character normalization

differences, therefore, we normalize those type of characters by converting into a single character as shown in figure 2.

Model building

To build the most efficient model for hate speech texts detection using deep learning neural network approaches 104,320 unique words extracted from 30,000 datasets to pass into an embedding layer to have a more efficient representation for our input data with one-hot encoded vectors. By using an embedding as input to the network learns a different embedding table on its own for learning semantic representations. After input words are passed to an embedding layer, the new embedding will be passed to LSTM cells and The LSTM cells will add recurrent connections to the network and give the ability to the system to include more information about the sequence of words in the data set. After all of this, the LSTM outputs will go to a sigmoid output layer. We're using a sigmoid function because hate speech and free becomes 1 and 0, respectively, and a sigmoid will output predicted values between 0 and 1. Finally, the hate speech text detection model looks like as shown in figure 3.

Experimental Results and Discussion

The experiment performed on the data set to measure the distribution of the dataset in 2 classes called hate speech and free. The data set has 2 labels which are hate speech and free by which 15949 of 30,000 are hate speech and others are free, which means 53.16% are hate speech text and 46.84% are free.

Experimental setup

To implement a recurrent neural network that performs hate speech text detection models in PyTorch which is a framework for Python based on Torch, which is a library for deep learning. To produce results different parameter combinations were performed for both LSTM and GRU. The datasets splits into three sets called training, validation, and test set by split fraction of 80:10:10 and different batch size, learning rate, embedding dimension, optimizer, number of an epoch, and dropout. Because, the successful implementation of neural network algorithms strongly depends on the parameter settings used. The value of each parameter used to produce

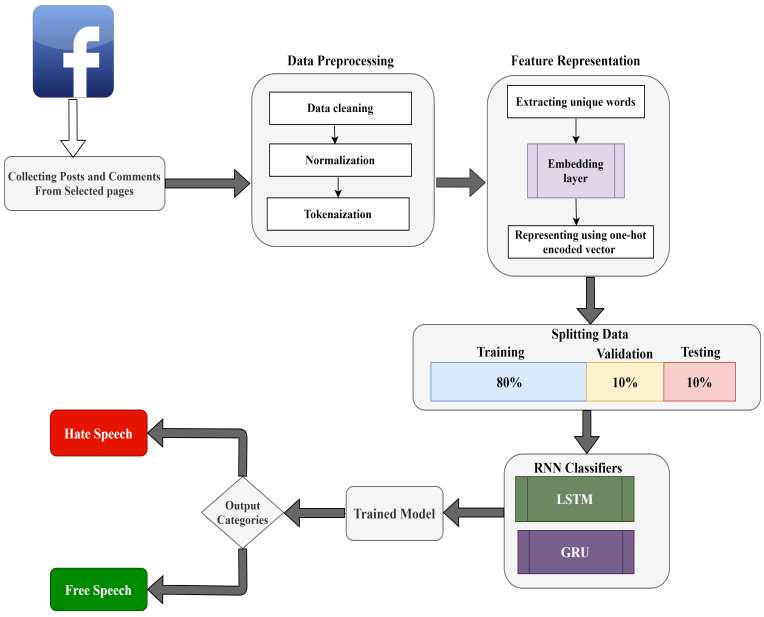

Figure 3 Methods of building hate speech detection models

good results cannot be determined manually. Therefore, the random search technique is used to tune the hyper-parameter by choosing some parameters because in many cases, it is not the models that require improvement and tuning, but the parameters. We select the following hyper-parameter values for LSTM-RNN and GRU-RNN models to experiment by interchanging their values as shown in figure 4.

Training

After splitting our dataset into training, test, and validation by split ratio of 80:10:10, which means the 30,000 data set 27,000 is used for training and validation and the remaining 3000 is used for testing. And then 27,000 data set is split by 90:10 split fraction to use 24,000 for training and 3000 for validation.

Testing

Our model is tested by 2 ways. The first one is test of model using a testing dataset which is 10% of the dataset by which we'll see how the trained model performs on all of the defined test data. In this case, the model Attempts to predict the labels for the given testing datasets and uses the test labels to calculate the accuracy of those predictions. The second way to test the model is an inference on user-generated data by which the model provided the text file that consists of the post at a time without a label and sees what the trained model predicts. By this time the model will accept new input text data from the user and predicting an output label.

Parameter combinations	Selected hyper-parameters and values				
	Batch size	Learning rate	Optimizer	Epoch	Dropout
PC1	32	0.1	Adam	20	0.1
PC2	64	0.01	Adam	50	0.25
PC3	128	0.001	Adam	100	0.5
PC4	32	0.1	Adamx	20	0.1
PC5	64	0.01	Adamx	50	0.25
PC6	128	0.001	Adamx	100	0.5
PC7	32	0.1	SGD	20	0.1
PC8	64	0.01	SGD	50	0.25
PC9	128	0.001	SGD	100	0.5
PC10	32	0.1	RMSProp	20	0.1
PC11	64	0.01	RMSProp	50	0.25
PC12	128	0.001	RMSProp	100	0.5
PC13	32	0.1	Adadelta	20	0.1
PC14	64	0.01	Adadelta	50	0.25
PC15	128	0.001	Adadelta	100	0.5

Figure 4 Selected hyper-parameters and their values

Results

The inter-annotator agreement for 1000 posts among 5 annotators has resulted in Fleiss's kappa of 0.664, which is a very good level of agreement and it is a substantial result because we have much more annotator and more data set than usual which could affect the inter-annotator agreement among annotators. Then after experimenting with the LSTM based RNN model by 15 different parameters combination different test accuracies achieved as the bar chart in the figure 5 demonstrates. The x-axis represents individual parameters combination and the y-axis represents the result of test accuracy.

For the second experiment, another RNN model was built by using a gated recurrent unit as a base and applied previously described 15 parameters combinations on our algorithm and evaluated it based on the evaluation metrics, by which parameter combination 12 perform 88% for our data set which is better than all other parameter combinations, but it is much smaller than the best result achieved by LSTM based RNN which is 97.9% test accuracy achieved by parameter combinations 12 as shown in Figure 6.

Discussion

Based on the result of the experiment, the recurrent neural network showed that this type of neural network could be used for text analysis of Amharic datasets. The result over the models covering two classes hates speech and free showed some very good results. Over the Amharic post and comments datasets, the fifteen parameter combinations have been evaluated using test accuracy, precision, and recall. Based on the result of the experiment Parameter combination 12 of the LSTM model is the best combination that has the highest percentage on the test dataset among those 15 combinations of the model, by which the model takes RMSProp optimizer,

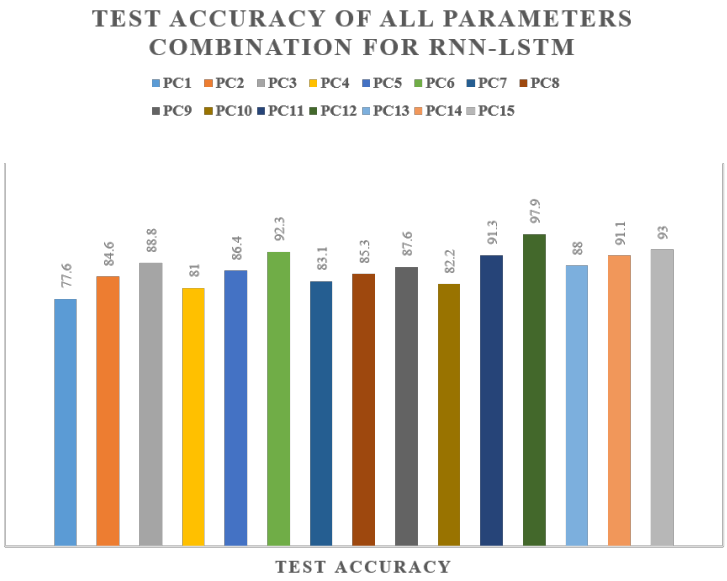

Figure 5 Test accuracy of 15 parameters combinations for LSTM

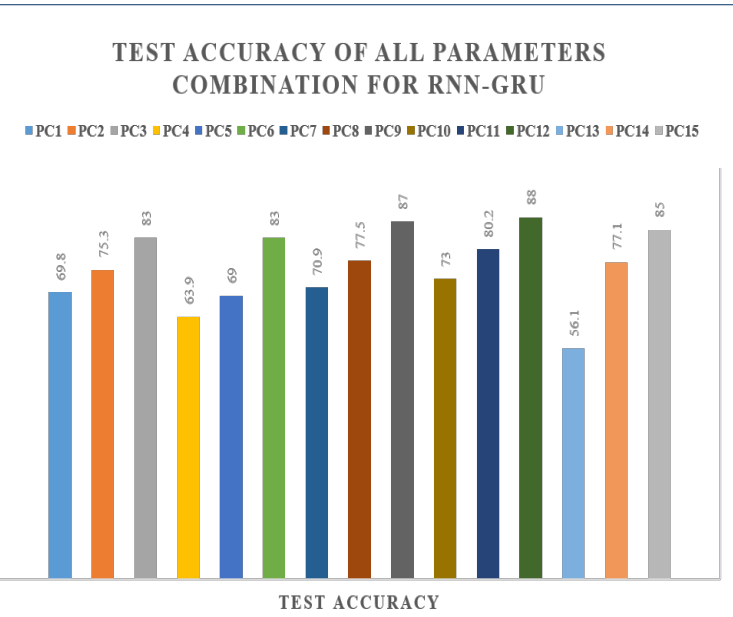

Figure 6 Test accuracy of 15 parameters combinations for GRU

0.001 learning rate, and 128 batch size and dropout value of 0.5 with 100 epochs to achieve 97.9%. Since the LSTM based RNN models show a good result on both validation data and test data, this indicates that the neural network has learned from the training data and can take the knowledge with them when new data has to be analyzed as we test it using inference on user-generated data. Finally, the LSTM based RNN model for hate speech text detection from Amharic posts can detect hate speech texts by training with the dataset from Facebook, which can solve the problem that caused by hate speech posts from Facebook by detecting and reporting them to Facebook by collaborating with the ministry of peace.

Conclusions

Nowadays, in social media, there are a huge amount of data exchanged among users daily and poses a lot of influence on users life positively or negatively, among those huge amounts of data hate speech texts on Facebook takes the major role to affect users life negatively by imitating them for conflict based on their race, religion, ethnic group or any other attributes which could differentiate individuals.

In this research, a model that can detect hate speech texts using recurrent neural networks is built. To successfully implement the model we start with defining hate speech text which is the text consist of both hate and offensive language based on their formal definition and we explore existing techniques for hate and offensive speech detection. The dataset is collected and labeled manually because of strict social media rule which prevents data collection tools, by which 30,000 data set is collected and annotated to the binary class of hate and free speech. Based on this data set the LSTM and GRU model is trained and tested by splitting the data set into a training, validation, and test set using the split ratio of 80:10:10. The experiment performed using this dataset with different parameters on GRU and LSTM based RNN model by feature representation of word2vec resulted in better test accuracy of 97.9% by RNN-LSTM.

Finally, we find out that using deep learning neural network models for Amharic text data analysis it is possible to detect hate speech posts from Facebook, and LSTM results in better efficiency than GRU based on our dataset. And the change of neural network hyper-parameters makes the change in accuracy of the deep learning model.

Acknowledgments

First and foremost I would like to praise the almighty God without the need of any reason and St. Maryam (Dingel Mariam) for the strength she gives to me from the beginning to the end of this research work and in my whole life too. I gratefully would like to give my special thanks to my advisor Dr. Kulla Kekeba for his brilliant guidance, constructive suggestion, and encouragement during this research. Secondly, I want to acknowledge my friends for their great help in the data collection process as it was so tidy process because of the block of previously available automated data collection tools. Finally I would like to thanks my family's as the whole for their love, endless support and encouragement throughout my life.

. . .

Funding

The journal of big data is funding for African country to publish the research.
I have allowed to not pay article processing fee because of lack of Fund for publication.

Abbreviations

. . .

Table 1 Abbreviations and Acronyms

Abbreviations and Acronyms	
AI	Artificial Intelligence
CNN	Convolutional Neural Network
DNN	Deep Neural Network
GRU	Gated Recurrent Unit
LR	Logistic Regression
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
RF	Random Forest
RMSProp	Root Mean Square Propagation
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency

Availability of data and materials
Data collected for this study is published on Mendeley. The data available at doi: 10.17632/ymtmxx385m.1

Ethics approval and consent to participate
The text written and acknowledged is true and it is the fact we have got by experiment.

Competing interests
We confirm that there are no competing interests. The article is not under any other review process and is not in the subject of any other submission.

Authors' contributions
Contribution to the knowledge. This study has the following contributions:
- Preparation of corpus or data sets for Amharic hate speech detection.
- Development of two models for amharic hate speech detection.

Authors' information
The Author is student at Addis Ababa Scince and Technology university and Lecturer at Wollo university. . .

Author details
Department of Software Engineering, Big Data and HPC CoE,Addis Ababa Science and Technology University, Addis Ababa, Ethiopia.

References
1. "Global digital overview," DataReportal – Global Digital Insights, 2020. [Online]. Available: https://datareportal.com/global-digital-overview
2. "Social media stats worldwide — statcounter global stats," StatCounter Global Stats, 2019. [Online]. Available: https://gs.statcounter.com/social-media-stats
3. M. Fick and P. Dave, "Facebook's flood of languages leave it struggling to monitor content," *Reuters, April*, vol. 23, 2019.
4. G. Hudson, "Linguistic analysis of the 1994 ethiopian census," *Northeast African Studies*, vol. 6, no. 3, pp. 89–107, 1999.
5. Z. Mossie and J.-H. Wang, "Social network hate speech detection for amharic language," *Computer Science & Information Technology*, pp. 41–55, 2018.
6. R. Cohen-Almagor, "Fighting hate and bigotry on the internet," *Policy & Internet*, vol. 3, no. 3, pp. 1–26, 2011.
7. H. Miklos, "Hate speech and the coming death of the international standard before it was born (complaints of a watchdog)," *The content and context of hate speech*, pp. 12–18.
8. K. Yonas, "Hate speech detection for amharic language on social media using machine learning techniques," Ph.D. dissertation, ASTU, 2019.
9. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *arXiv preprint arXiv:1703.04009*, 2017.
10. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.
11. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145–153.
12. T. Tilahun, "Linguistic localization of opinion mining from amharic blogs," *International Journal of Information Technology & Computer Sciences Perspectives*, vol. 3, no. 1, p. 890, 2014.
13. N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 69–76.
14. A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach," *arXiv preprint arXiv:1809.08651*, 2018.
15. N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggression and violent behavior*, vol. 40, pp. 108–118, 2018.
16. G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in twitter data using recurrent neural networks," *Applied Intelligence*, vol. 48, no. 12, pp. 4730–4742, 2018.
17. Z. Mossie, "Social media dark side content detection using transfer learning emphasis on hate and conflict," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 259–263.

18. Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019.
19. M. Corazza, S. Menini, P. Arslan, R. Sprugnoli, E. Cabrio, S. Tonelli, and S. Villata, "Comparing different supervised approaches to hate speech detection," 2018.
20. F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 2017, pp. 86–95.
21. A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53 040–53 065, 2019.
22. J. Patihullah and E. Winarko, "Hate speech detection for indonesia tweets using word embedding and gated recurrent unit," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 1, pp. 43–52, 2019.
23. B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," in *Proceedings of the 10th ACM conference on web science*, 2019, pp. 173–182.
24. A. Kaur and D. Chopra, "Comparison of text mining tools," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2016, pp. 186–192.

Figures

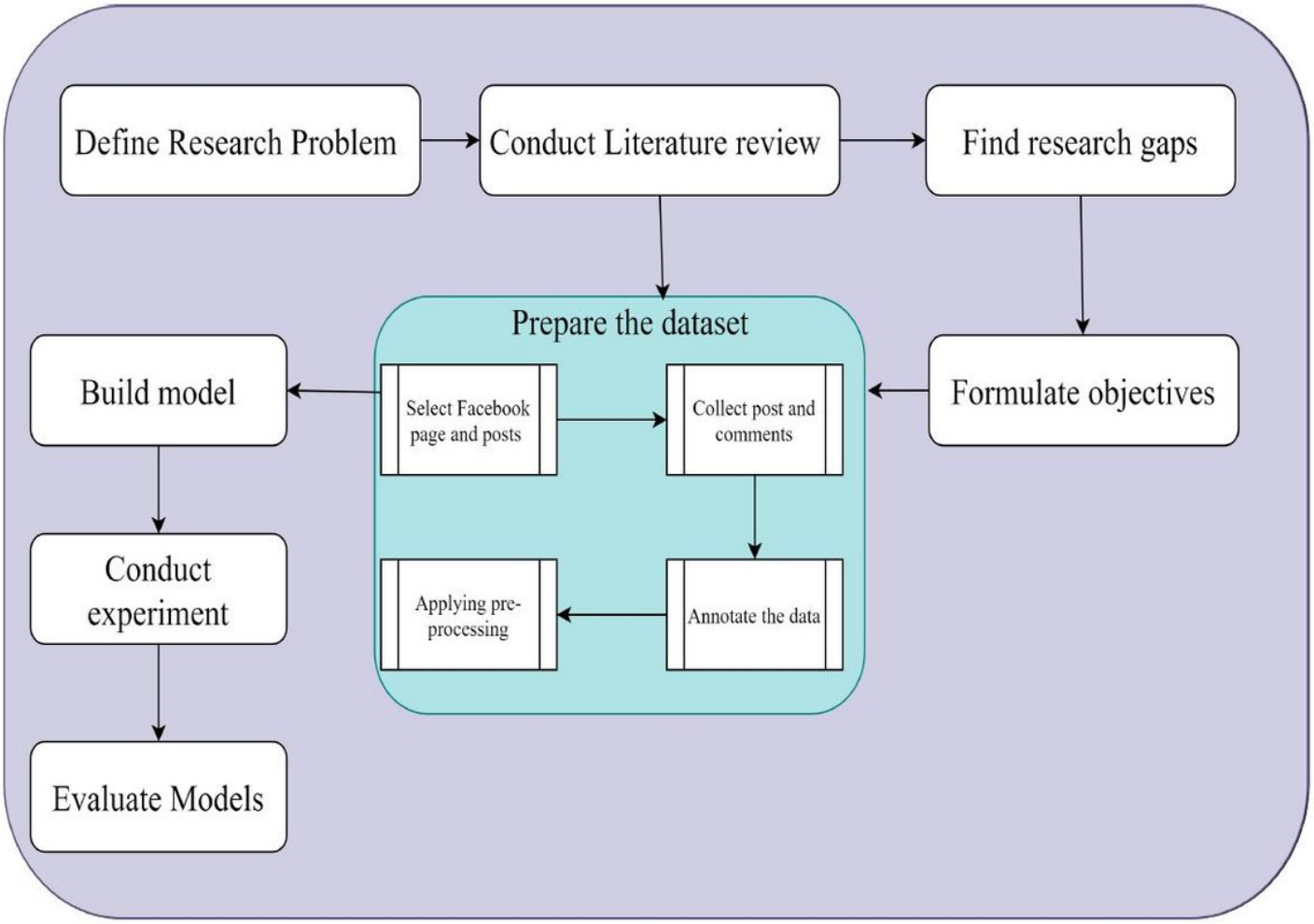

Figure 1

Research flow diagram

Sound characters	Normalized to
ሀ, ኀ, ሐ, ኃ, ኻ, ሓ, ሃ	ሀ
ሰ, ሠ	ሰ
ጸ, ፀ	ፀ
አ, ኣ, ዐ, ዓ	አ

Figure 2

Amharic character normalization

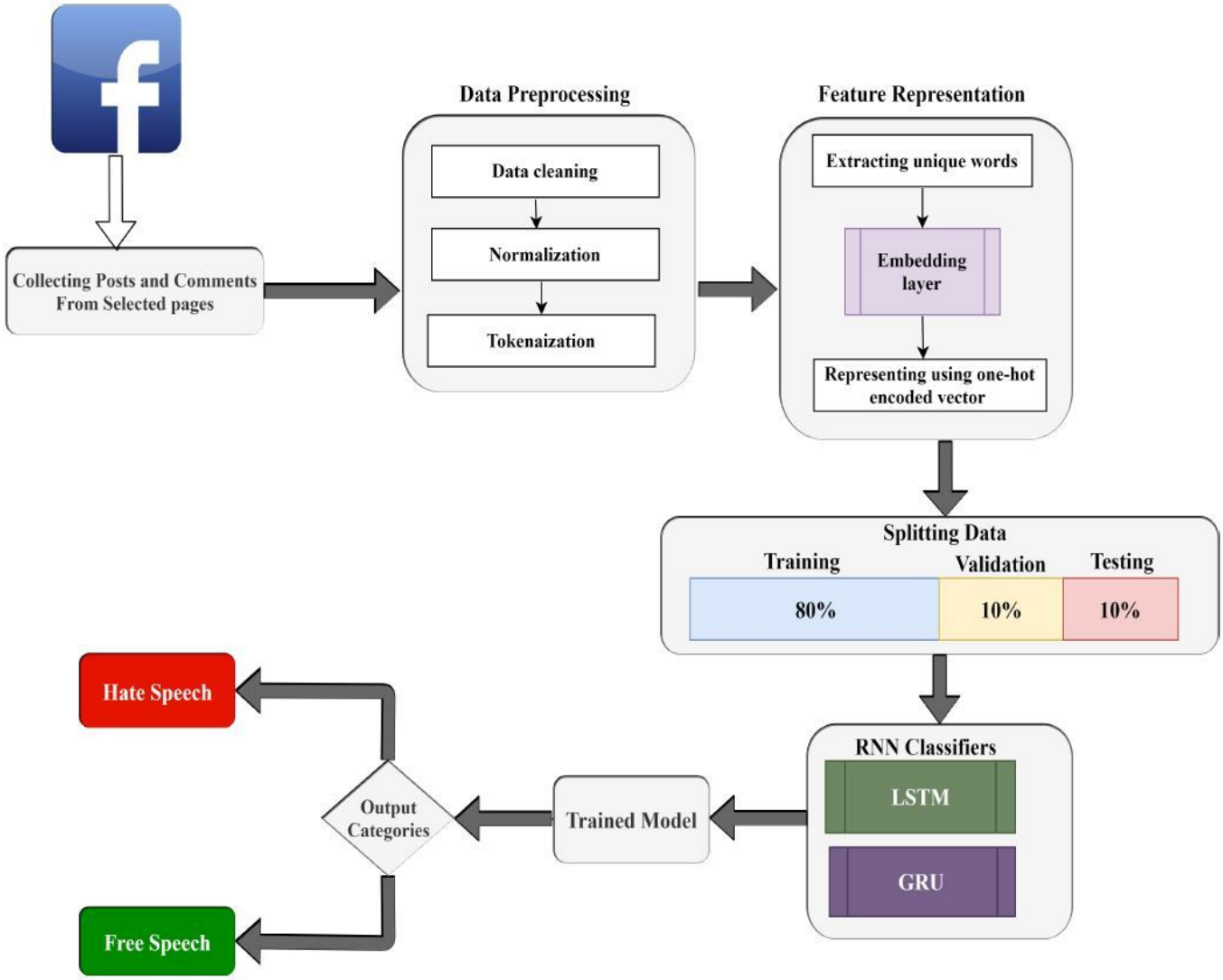

Figure 3

Methods of building hate speech detection models

Parameter combinations	Selected hyper-parameters and values				
	Batch size	Learning rate	Optimizer	Epoch	Dropout
PC1	32	0.1	Adam	20	0.1
PC2	64	0.01	Adam	50	0.25
PC3	128	0.001	Adam	100	0.5
PC4	32	0.1	Adamx	20	0.1
PC5	64	0.01	Adamx	50	0.25
PC6	128	0.001	Adamx	100	0.5
PC7	32	0.1	SGD	20	0.1
PC8	64	0.01	SGD	50	0.25
PC9	128	0.001	SGD	100	0.5
PC10	32	0.1	RMSProp	20	0.1
PC11	64	0.01	RMSProp	50	0.25
PC12	128	0.001	RMSProp	100	0.5
PC13	32	0.1	Adadelta	20	0.1
PC14	64	0.01	Adadelta	50	0.25
PC15	128	0.001	Adadelta	100	0.5

Figure 4

Selected hyper-parameters and their values

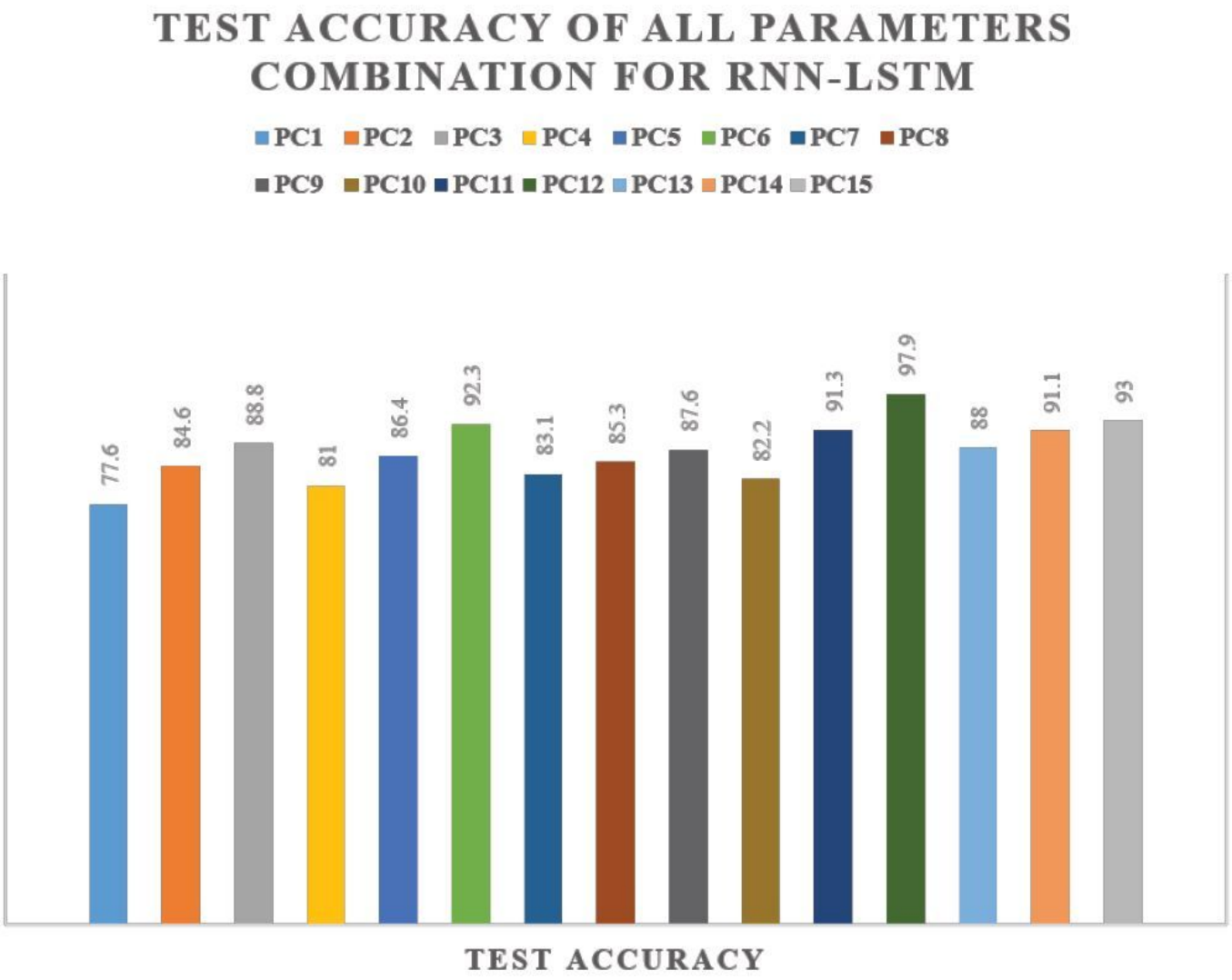

Figure 5

Test accuracy of 15 parameters combinations for LSTM

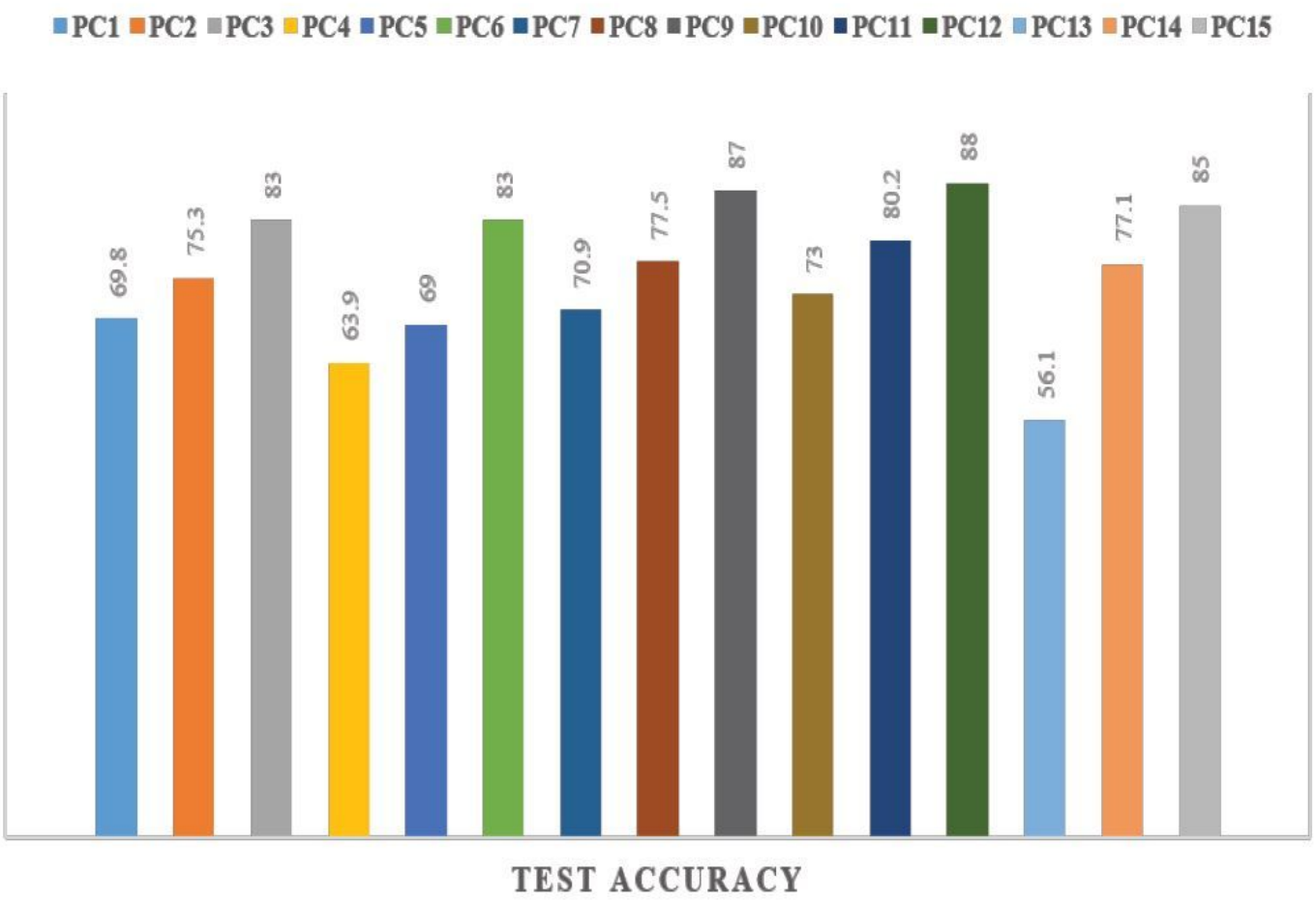

Figure 6

Test accuracy of 15 parameters combinations for GRU