SENTIMENT ANALYSIS OF MYANMAR NEWS AND COMMENTS USING SUPPORT VECTOR MACHINE

THEIN YU

UNIVERSITY OF COMPUTER STUDIES, YANGON

April, 2023

Sentiment Analysis of Myanmar News and Comments Using Support Vector Machine

Thein Yu

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

April, 2023

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

28.4.2023

Date

Thein Yu

ACKNOWLEDGEMENTS

Firstly, I would like to thank the Union Minister, the Ministry of Science and Technology for giving me the opportunity to study Ph.D. course by providing support that all me to perform the research in University of Computer Studies, Yangon.

Secondly, I am very deeply appreciating to Dr. Mie Mie Khin, Rector, the University of Computer Studies, Yangon, for authorizing me to implement this dissertation and provide me valuable tips in that time of my research.

I will be fond of to present many acknowledgements to Dr. Mie Mie Thet Thwin, former Rector, the University of Computer Studies, Yangon, for permitting me to do this thesis and providing recommendations and instruction through the time of my research.

I am really appreciate to the external examiner, Dr. Thinn Thinn Wai, Professor, Faculty of Computer Science, the University of Information Technology, for helpful advice.

I will also be fond of to express my respectful and lovely gratitude to Dr. Thin Lai Lai Thein, Professor, Course-coordinator of of the Ph.D. 10th batch, the University of Computer Studies, Yangon, for her kindness and excellence guidelines

I might also delight to show my immeasurable and honorable gratitude to my supervisor, Dr. Khin Mar Soe, Professor, Faculty of Computer Science, the University of Computer Studies, Yangon, tolerant supervision, gentleness and kindness, push and giving best ideas during the study of this research time.

I would like to express my immeasurable thanks to my former supervisor, Dr. Khin Thandar Nwet, Associate Professor, Faculty of Computer Science, the University of Information Technology, Yangon, for her tolerant monitoring, kindness, push and giving me with best plan in the time of research.

I would also like to express my admiration and gratitude to Dr. Khine Khine Oo, Professor and Former Dean of the Ph.D. 10th Batch at the University of Computer Studies in Yangon.

My gratitude goes to Dr. Win Pa Pa, Professor, UCSY, for her directions and valuable instructions.

I decide to present my admiring appreciation to ours beloved teachers for their motivation and guidance for research. To the reading board teachers, mainly Daw Aye Aye Khine, Associate Professor, Head of English Department, I am very grateful to her for helpful advice and modifying my research by viewing language direction.

Finally, I am very thankful to my family for every time trust to me, for their innumerable adore and encourage. They always make provision for me throughout the times of this Doctorate Course.

I also recognize to all my soulmates from the Ph.D.10th batch for their collaboration and friendly. I want to voice my gratitude to everyone who helped make this thesis successful, whether directly or indirectly.

ABSTRACT

As the development of internet technology is raising, the volume of information used for the internet users also increase in the web. Users can apply that information and give opinions for decision making system. Sentiment analysis also known as opinion mining is a task of text categorization methods that take opinion presented in a piece of text. An active research area is the sentiment analysis of text documents. The essential text resources found on social media, such as reviews, comments, tweets, posts, opinions, and articles, are available in a variety of languages. These could be analyzed to learn more about people's attitudes, beliefs, and feelings concerning various topics and products. With a focus on the Asian Language Treebank, news from Ministry of Information website(www.moi.gov.mm), and comments from social media webpage (www.facebook.myanmarcelebrity.com.mm), this paper aims to target news and comment of sentiment analysis in Myanmar social media. In order to categorize the sentiment polarity of each social media comment into "positive," "negative," or "neutral,", automated analyzer methods were proposed in this paper.

This system constructs corpus for news comments for Myanmar language. The datasets were then split into training and testing datasets, with the training dataset being randomly split in a non-overfitting way using the cross-validation approach. In order to improve the performance of the classifier, the case of imbalanced datasets was then considered. The hyperparameters were modified to improve the performance and outcomes of the classification. In addition, a number of information visualization techniques were used to display the results, indicate how effectively the classifiers performed, and highlight the key terms that had an impact on the classification process.

Feature weighting and selection are required in sentiment analysis to get more efficiency. The proposed system implements sentiment analysis system for Myanmar News and comments. TF-IDF and N-gram are used for feature weighting and extraction. Support vector machine (SVM) is a supervised learning methods that analyze data and recognize the patterns that are used for classification. Hyperparameter optimization is used to find the set of specific model configuration arguments that does in the best performance of the model. Random search is an algorithm in which random

combinations of hyperparameters are chosen and applied to train a model. The best random hyperparameter combinations are choosed. This system improves the Myanmar news sentiment analysis system using SVM with Random search optimization. This system also studies the machine learning algorithms for Myanmar sentiment analysis system. This system showed that the comparison results of Naïve Bayes, Linear SVC, and Linear SVC with random search optimization. Linear SVC with RandomizesearchCV has the highest performance.

This system shows the most significant terms that had an impact on the classification process as well as the classifiers' performance. The results were then presented, along with ideas for how to optimize them in the further and information on how well the suggested systems worked.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

The development of Web 2.0 technologies has increased and provided many chances for opinions mining, not only of the general public data but also of organization, entertainment, and diplomacy. Sentiment analysis has been more popular in recent years for automatic customer satisfactions analysis of online services such as blogging and social network as it can provide business insights by classifying public opinions on social data. It is widely used in data mining, text mining, web mining and information retrieval. Sentiment analysis is one of the text mining techniques. Forums, blogs, review sites etc. give the freedom to society to talk the sense with no limitations. News is important for real life with number of reasons in a society. News is mainly informed to the public about events that are take placed around world and may influence world. All text data is coming from by the opinions of the people that gives a new way to classify those group of data. Labeling those data depend on sentiments have a better, clearer and a strong insight to the user. Sentiment analysis also involved a main role in classifying and tagging important news and information from the more casual content [42]. Applications of sentiment analysis are widely applied to many possible areas such as commercial, political, education, and so many societies. News can also be used for entertainment to give an interference of information. News can provide people feeling related too. News is an important place of social gathering place too and can be taken by either online or physical place. There are many different sentiment analysis systems for different languages by using different algorithms, levels, and resources. So, Myanmar automatic sentiment analysis system is not widely applied. Therefore, developing sentiment analysis systems for Myanmar documents is a challenging task due to the scarcity of language resources like automatic tools for part of speech tagger, feature selection and stemming etc.

In this research, an automatic sentiment analysis system for Myanmar news is proposed. The proposed system is implemented by using supervised learning approach. For the training data set, news from Myanmar media websites are collected for data set. N-Gram and TF-IDF are used as a feature selection and extraction method and improved support vector machine is applied in implementing the text classifier.

1.1 Motivation

Recently, there is no widely used sentiment analysis system for Myanmar language. Sentiment analysis has been a considerable effort from industry and academia. Opinionated postings in social media have helpful to redevelop businesses, and sway public sentiments and emotions, which have more effected on social and political systems. Sentiment analysis applications have widely spread to almost every possible application. The practical applications and industrial interests have provided strong motivations for research in sentiment analysis. Therefore, improved sentiment analysis systems for Myanmar news are needed to implement.

1.2 Opinion Mining

Opinions are mainly to almost all human tasks and are essential impacts of their manners. Opinion and its associated entities such as sentiments, validations, attitudes, and emotions are the content of opinion mining field. Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions for products, services, organizations, individuals, issues, events, topics, and their attributes. Sentiment analysis applications have widely spread to almost many domains such as consumer products, services, healthcare, financial services, and social events and political elections. Sentiment analysis has generally in three level such as document level, sentence level, and aspect or word level. Sentiment analysis is done by three approaches such as machine learning approach, lexicon-based approach, and hybrid approach [36].

1.3 Machine Learning

Machine learning regularly defines to the variations in systems that do functions related with AI. Such functions contain recognition, diagnosis, planning, robot control, prediction, etc. There are many techniques in machine learning derive from the psychologist's decision to get more precise their theories of animal and human learning by using computational models. Generally, three machine learning methods such as supervised learning, unsupervised learning, and semi-supervised learning are used in today [65].

1.4 Problem Statement

With the explosive development of social media on the Internet, most people and societies are widely using the content in these media for decision making. Today, if one needs to buy a product, there are a lot of user reviews and suggestions in public forums on the Internet about the product. For societies, there are no longer needed to connect surveys, opinion polls, and focus groups in order to get public opinions due to abundance of such information publicly available. But, searching and computing opinion on the Web occurs difficulty task because of the spreading of many different sites. Each site typically consists of large number of opinions that is not usually easily defined in long blogs and forum postings. The reader will not have easily identifying related sites and extracting and analyzing the opinions in them. So, automated sentiment analysis systems are required and real-life applications are courses of why sentiment analysis is a popular research problem. It has a main challenge as a NLP research topic research problems related to the existing research NLP problem. It uses every aspect of NLP, e.g., anaphora resolution, negation handling, and word sense disambiguation, which are difficult to solve problems in NLP. However, it is applied to perceive that sentiment analysis is a mainly constricted NLP problem because the system does not require to fully interpret the semantics of each sentence or document but only requires understanding some aspects of it, i.e., positive or negative sentiments and their target topics. Therefore, sentiment analysis gives a significant platform for NLP researchers to make development on all appearances of NLP with the mainly huge practical impress.

1.5 Objectives of the Research

The aim of this research is to do sentiment analysis system for Myanmar language. This system uses machine learning method to model classifier of sentiment of news dataset. The primary goal of the research is to optimize support vector machine with Random search optimization algorithm. The framework will be applied to search sentiment orientation of news that could help in the environment to care and overcome difficult situations. The specific objective of this research area are shown in below:

1. To analyze large volume of specific news articles and find positive, neutral, and negative opinion

2. To explore various sentiment analysis/opinion mining approaches
3. To improve SVM classification with Random search optimization in sentiment analysis
4. To develop a sentiment analyzer for Myanmar language

1.6 Target of the Research

The research is concentrated on establishing an improved SVM to sentiment analysis methods. These target works involve the following:

(i) Determining the past opinion mining methods and their natures
(ii) Analyzing the Myanmar languages to get the highest expressing power for sentiment corpus
(iii) Studying the improved SVM's quality measurements
(iv) Introducing an improved SVM classification algorithm with N-gram and TF-IDF features to satisfy the SVM quality
(v) Evaluating the improved SVM performance measurement
(vi) Proving error analysis of an improved SVM

1.7 Contributions of the Research

The item of the research intended to provide in following:

- This system constructs sense annotated corpus for Myanmar language.
- This system proposes sentiment analyzer for Myanmar news by using Random search optimization

The research constructs news sentiment corpus and integrate existing news' data and applied machine learning methods such as naïve bayes, support vector machine (SVM), and improved support vector machine with random search optimization. As a result, randomize optimization method is defined to improve SVM's performance.

1. This paper presents and develops most appropriate methods for pre-processing, feature extraction, and classification of sentiment analysis. The proposed system will improve the analysis of sentiment information. Moreover, pre-processing will process in many steps such as word segmentation, tokenization, and stop words removing. And then, N-gram will also be applied for feature extraction. Combinations of N-gram are used in this proposed system. Combination of unigram and bigram

feature is giving better performance results for Myanmar text than other N-gram features [36].

2. This paper shows on the efficiency of pre-processing with different combinations of feature weighting schemes on Myanmar text and finding the benefit of using improved algorithms in Myanmar sentiment analysis.

1.8 Structure of the Research

This dissertation is structured with six chapters, involving starting of sentiment analysis, the important role sentiment analysis in several domains, problem issues, aims and contributions of the research. Chapter 2 studies and records the many resources to sentiment analysis and opinion mining methods that are concerning with the dissertation. The backgrounds theory of opinion mining and machine learning, the differences in resources are expressed in Chapter 3. The flow of design and developing the proposed system and algorithms for sentiment analysis are explained in Chapter 4. Chapter 5 shows the performance evaluation of the experiment by measuring with machine learning algorithms and improved support vector machine algorithm's quality measurement metrics and processing time. Finally, Chapter 6 expresses conclusion drawn from this research effort and outlines the future tasks to carry on with it.

CHAPTER 2
LITERATURE REVIEW

In this chapter, the review of related works of the sentiment analysis and opinion mining is described. The literature study and background investigation for sentiment analysis, Myanmar sentiment analysis, and Myanmar language are presented in this chapter. It also describes the many types of sentiment analysis as well as its stages, jobs, and difficulties. The chapter gives details on the feature selection process and offers information on news and social media sentiment analysis. The adopted sentiment analysis methods for news and social network data are then highlighted, along with the sentiment analysis topics covered in scholarly works. The chapter concludes with a discussion of Myanmar sentiment analyses in terms of linguistic peculiarities, difficulties, and scholarly contributions. There are three mainly different levels in task for sentiment analysis system such as document, sentiment, and aspect or entity level.

- **Document level:** In this level, sentiments are classified for the whole document is positive, negative, and neutral opinion. For example, a product or event review is determined to classify sentiment of overall opinion. This level can analyze every document or event have a single entity and not to evaluate or compare multiple entities.
- **Sentence level:** The level is responsible for a task to determine sentence is positive, negative, and neutral opinion. This task is related to subjectivity classification.
- **Entity and Aspect level:** Both the document level and the sentence level analysis do not find what exactly people liked and disliked. Aspect level that is also called feature level does finer-grained analysis. Opinion targets are presented by different entities and/ aspects. The aim of this analysis is to find sentiments on entities and/or their aspects [42].

There are three methods that have been adapted from machine learning and lexicon-based to deep learning approaches. Research mainly developed mainly emphasis on the use of corpus-based methods in this research. All of the related approaches and works are discussed in later. And then, finish with a consideration on the current research to this research that will find suggestion in relation to corpus-based sentiment analysis for Myanmar news.

2.1 Sentiment Analysis

The goal of sentiment analysis, is also called opinion mining, is to examine how people feel about various types of items, including people, themes, issues, services, organizations, services, products, events, and their characteristics. This is a large area of research. A number of names, including opinion extraction, subjectivity analysis, emotion analysis, opinion mining, sentiment mining, review mining, and affect analysis, can be used to describe sentiment analysis. However, all of the aforementioned phrases fall either within the sentiment analysis or opinion mining umbrella. The field of sentiment analysis is an important research area in applied linguistics. The significance of sentiment analysis is recognized in a variety of fields, including political science, education, and marketing. Furthermore, sentiment analysis or opinion mining extracts information by determining the data that indicate negative, positive, or neutral texts in given documents. This extraction could be accomplished using machine learning, natural language processing, and statistics, which aid in determining the polarity of a given record. The extracted critique, feedback, or comment may contain sentiments that can be used as a valuable indicator for a variety of purposes.

It is possible to classify a sentiment as either negative, positive, or neutral, or to use an n-point scale to assign it a rating such as "very awful," "bad," "satisfying," "good," or "very good." Each of the earlier classes creates an emotion. Corporations can use this approach to assess the market acceptance of their products and help them develop plans to improve the quality of their offerings. Additionally, legislators or policy makers may benefit from using sentiment analysis to assess public opinion on issues, services, or programs. Due to the widespread use of social media, it is vital for many marketers, other players in the social media sector, and outside agencies to analyze the feelings and emotions of social media users.

2.2 Machine Learning Method

Most of the research are developed on the use of machine learning and deep learning techniques in sentiment classification.

2.2.1 Supervised Machine Learning

The field of text classification well-motivated through supervised machine learning algorithms utilized with sentiment-labelled training data to analyze the sentiment of unlabeled test data. Initially, standard text pre-processing, feature selection and vector-space representation make use of the training and test data. Therefore, at the training phase, model is trained by using machine learning algorithm. And then, at the testing/prediction phase, documents are classified that are previously unseen by the model.

There are many different sentiment analysis systems for different languages by using different algorithms, levels, and resources. In previous work, sentiment analysis system by using ensemble classifier that used trip advisor review. The system applied bag of word model for feature selection and support vector machine, logistics regression, and naïve bayes are employed to categorize reviews while forming an ensemble classifier [34]. Sentiment analysis system for feedback marketplace at aspect level feature is developed by using support vector machine algorithm to classify consumer review of marketplace with N-gram and TF-IDF features [65].

News sentiment analysis system for business is developed by using Weka tool for data preprocessing and feature generation. And then, they used naïve bayes algorithm to classify news article [56]. Bollywood song lyrics corpus annotated with sentiment polarity that was created by three annotators. They classified with support vector machine, Multinomial naïve bayes, and Bernoulli naïve bayes [25]. Machine learning algorithms is implemented for sentiment analysis system. They analyzed three machine learning algorithms for opinion mining such as support vector machine, naïve bayes, and maximum entropy [35]. In the paper written by Bagarwal, V K Sharma, and N Mittal, Point-wise Mutual Information (PMI) based method is applied with extracted sentiment-rich phrases by using Part of Speech (POS) based rules and dependency relation in the document [33].

PranavWaykar, Kailash Wadhwani, Pooja More from Department of Computer Engineering, DYPIET, Pimprihas developed sentiment classification system using supervised approach. The unigram algorithm was applied to derive feature set from twitter dataset and Naïve Bayes classifier was then used on derived features for final categorization [42]. The algorithm is used for extracting bi gram feature. The research applied methods using R Studio software.

M.Kanakaraj and R.M.R.Guddeti presented the implementation of sentiment analysis to search the polarity value of opinions in text selected from e-commerce magazines and blogs in Arabic language. They implemented a small emotion converter and an elongated words checker. Features are represented n-gram range (unigrams, bigrams, tri-grams, mixture of unigrams and bigrams, mixture of bigrams and trigrams, mixture of unigrams, bigrams, and trigrams) for feature extraction. In standard corpus, the highest result value is obtained by unigram, compound of unigram and bigram, and compound of unigram, bigram, and trigram by using Arabic light stemming and Naïve Bayes classifier. In preprocess stemmed corpus, the best performance result is obtained by using support vector machine in the union of unigrams and bigrams, combination unigrams, bigrams, and trigrams[33]. Thititorn Seneewong Na Ayutthaya and Kitsuchart Pasupa implemented sentiment analysis system for Thai children's stories. They developed combination of part-of-speech and sentic features. They implemented SA system by combining Bidirectional Long Short-term Memory and Convolutional Neural Networks models. They used 40 Thai stories and got the highest performance result of f1 score 78.79 [25]. C. J. Varshney, A. Sharma and D. P. Yadav proposed sentiment analysis system using ensemble method. They used twitter dataset and used TFIDF vectorizer with n-gram methods for feature selection. They compared classification algorithms such as Random Forest, Support Vector Machine, Logistic Regression, Naïve Bayes, and SGD classifier and combined these algorithms with ensemble method[59]. Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar developed aspect-based sentiment analysis system of SemEval14. They used laptop and restaurant review dataset and. They used firstly supervise machine learning and then extend to latent semantics discovery (LDA) and the method established on sentiment vocabularies[57]. Georgios Paltoglou and Mike Thewall presented feature weighting scheme from information retrieval to increase performance of sentiment analysis. Authors applied movie reviews, multi-domain sentiment corpus and BLOGS06 corpus and test with different weighting method such as classic tf.idf weighting schemes, Delta tf.idf schemes, and SMART and BM25 tf.idf scheme by using SVM method [68].

G.Paltoglou and M. Thelwall implemented sentiment analysis system that used TF-IDF weighting transformation and support vector machine classification method. And then they used twitter user comment database for dataset [47].

2.2.2 Unsupervised Machine Learning

Carlos Henríquez Miranda and Edgardo Buelvas implemented an unsupervised sentiment analysis system for Spanish language. They used restaurant opinion corpus and external libraries such as Freeling to process part of speech tagging (POS) and MCR to extract aspect-based feature that use attenuation and negation. The system is intended to fit any language and domain. Point-wise mutual information (PMI) is used to calculate sentiment polarity [42]. M. Fernández-Gavilanes and T. Álvarez-López and Jonathan Juncal-Martínez and E. Costa-Montenegro and Francisco, and J. González-Castaño propose unsupervised sentiment analysis for online textual messages such as tweets and reviews. This system used classification algorithms that holds ways to process natural language and different types of emotion characteristics originally obtained from sentiment lexicons. And then, use dependency parsing to evaluate tweet polarity and sentiment words to consider special structure and linguistics structure of message. Results from the Cornell Movie Review Obama-McCain Debate and SemEval-2015 datasets demonstrate the system's competitive performance and robustness[22]. Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu investigate whether the signals can provide an integrated leverages to sentiment analysis model that have two essential section of emotion sign and emotion interrelation These signals are combined into an unsupervised learning for sentiment analysis. And then, trained on two Twitter datasets to search the existence of emotion [59].

X. Hu, J. Tang, H. Gao, and H. Liu developed a sentiment analysis method based on K-means and online transfer learning. The source domain data and a small number of target domain data are preprocessed for text segmentation and stop words deletion. And then, map the text to word vector by using Word2Vec model. Authors first used the K-means clustering algorithm to generate data from one or multiple source domains and selected the data similar to target domain data to model the classifier [31]. S. Wu , Y. Liu1, J. Wang and Q. Li implemented a model for searching sentiment class in reviews by using unsupervised method. This model uses a generalized method to learn multi-word class and a set of rules is used to take number of the effect of an opinion word on detecting the class. A new measurement based on mutual information and aspect frequency is used to calculate result of algorithm [60]. Toma´s Hercig , Toma´s Brychc , Luka´s Svoboda , Michal Konkol, and Josef Steinberger participated employing only unsupervised or weakly-supervised approaches for SemEval-2015 task 12. They required minimum annotated or hand-crafted content and use Word2Vec to force in-domain semantic matches of words for many of the involved subtasks. SemEval 2014. use labeled and unlabeled corpora within the restaurant's domain for two languages: Czech and English to show that their models improve the (aspect-based sentiment analysis) ABSA performance and demonstrate the value of our strategy. They created word clusters and used as feature. And then, used unsupervised stemming algorithm (HPS) and show that GloVe and CBOW model seem to be the best [57]. Xinxin Guan, Yeli Li, Hechen Gong , Huayan Sun, and Chufeng Zhou develop improved sentiment analysis system for book review using SVM and Bayes algorithms[31]. They used 4000 review data set for book Douban reading and implement construct sentiment dictionary for book review. Jaspreet Singh, Gurvinder Singh & Rajinder Singh improved sentiment analysis system using machine learning classifier. Three review dataset such two Amazon product and one IMDB movie review are used with Naïve Bayes, J48, BFTree and OneR for optimization of sentiment analysis [55]. Himani Khullar and Amritpal Singh proposed a sentiment analysis and sarcasm detection by using bagged gradient boosting with particle swarm optimization for feature selection and noisy data removing [55].

2.3 Lexicon-Based Method

The lexicon-based approach relies on large set of known and pre-trained domain sentiment words. It is subdivided into dictionary-based and corpus-based approaches applied to search the polarity score from huge corpus. The dictionary-based approach finds the opinion word and searches its synonyms and antonyms in the source list to calculate the polarity from a dictionary. The corpus-based approach is an approach which used an individual domain developed by individual. So, the sentiment words in the corpus are specified by context. The lexicon-based scheme is used to find sentiments from subjectivity lexicons to develop a corpus-based dictionary by remodeling the insights of the co-occurrence and conjunction method. These lexicons contain sentiment words that are also called opinion words listed with their polarity and strength. Khin Zezawar Aung and Nyein Nyein Myo generated opinion of student feedback comments. They used English sentiment lexicon for student feedback related to teacher. They represented polarity score of opinion and polarity level such as strongly negative, negative, weakly negative, strongly positive, positive, weakly positive, and neutral. This lexicon defined polarity score value range from -3 to +3 and sentiment are calculated by heuristic method [56].

T. Al-Moslmi, M. Albared, A. A l-Shabi, N. Omar, and S. Abdullah developed Arabic sentiment lexicon that contain 3880 positive and negative sense tagged with their part of speech, polarity score, direct senses, and inflected form. Their system described the corpus that contain 8860 positive and negative tagged review. Naïve bayes, KNN, SVM, logistics regression, and Neunet algorithms are trained with both sentiment lexicon and corpus. Results show that lexicon-based approach is better performance than corpus-based approaches [33].

2.4 Deep Learning

W. L. K. Khine, N. T. T Aung implemented a convolutional neural network (CNN) model with the gating control mechanism for aspect level sentiment analysis. By using gating control technique, system can be more accurate and efficient in aspect filtering. They used different domains such as laptop, product, restaurant, and hotel and corpus such as IMDB, amazon, yelp, etc...[36]. M.T.A.Bangsa, S. Priyanta, Y. Suyanto implement sentiment analysis of product reviews. They classified aspect level for sentiment from reviews on Indonesian site www.bukalapak.com. Convolutional neural network (CNN) method is used and Word2vec is used for feature extraction [35]. Makoto Okada and Hidekazu Yanagimoto, and Kiyota Hashimoto developed sentiment analysis system for customer review of Amazon Product Review dataset and TripAdvisor Japanese review dataset. Gated convolutional neural networks model is used and then compared the best performance of gCNN faster than RNN in different datasets [45].

K. Devipriya, D. Prabha, V. Pirya, and S. Sudhakar developed deep learning model to classify sentiment online text. They trained with various deep learning algorithm and Word2vec feature and crate recommender system for various social applications [61]. Maha Heikal, Marwan Torki, Nagwa El-Makky implemented sentiment analysis system for Arabic tweets that use ensemble, CNN, and LSTM deep learning model. They used pretrain word2vector representation. This research showed that how DL algorithms can enhance the performance of sentiment analysis than the traditional machine learning algorithms for text-based analysis[29].

2.5 Feature Selection

Data can be represented as feature with numeric vector. Many different techniques are existed to transform data into feature that will be numeric form. Feature can be identified by scores and the highest scores value is selected for using in training model. A feature is important if it is greatly related with the relevant variable. The importance of feature score is to provide information for extract feature. The features of data will directly affect the learning models and the results may be better.

Feature engineering is the transformation process of data into features to get better performance of system model. Feature extraction is a process of dimension reduction of feature that can be used in modelling. Feature selection identify the problems by choosing an entity that are most suitable to the problem [33].

Maria Mihaela Trușcă proposed support vector machine models using both TF-IDF approach and Word2Vec and Doc2Vec neural networks for text data representation. In non-linear SVM, their results show that Word2Vec is outperformed than other Doc2Vec and TF-IDF. In linear case, TF-IDF is more efficient than other. Reuters 21578 dataset is used in the system. [57]. T. Georgieva-Trifonova, and M. Duraku developed text classification by using N-gram for feature selection. They used Reuters-21578 and Customer_feedback_bg dataset by using and feature selection in performed by using Relief algorithm, Chi-squared, Information gain, and Gini index. K-NN, Decision tree, Deep Learning, the rule-based classifiers RIPPER (JRip), Ridor, PART algorithms are applied in this system for classification[47]. A. Yang, J. Zhang, L. Pan, and Y. Xiang developed enhancement of sentiment analysis for twitter by utilizing feature selection and combination. They combined sentiment lexicon and unigram of high information gain and classified by using six machine learning algorithms and show that multinomial naïve bayes is better than other in efficiency[61].

M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S Manandhar developed sentiment classification model that use relative term frequency and inverse document frequency. Senti-TFIDF is evaluated on movie data.

2.6 Summary

This chapter discusses the several different methods to sentiment analysis with theory, methodologies and application tools. Finding out the state-of-the-art opinion mining or sentiment analysis become one of the popular research areas for researchers to support sentiment analysis system to organization, product, and events. Those methods widely use the machine learning algorithms to classify the different opinion. And then, the machine learning will be learned by many researchers to develop prediction system for organization, products, events, and so on. According to the classifier's measurement method, many metric calculations should be considered in different processes. In this chapter, the methods and aspect of sentiment analysis and various levels of sentiment analysis have been described. Generally, the semantic orientation and machine learning approaches for sentiment analysis have been described thoroughly. Sentiment analysis for Myanmar news has done with Naïve Bayes and Support Vector Machine. This paper used TF-IDF and N-gram for feature extraction with 3000 news data which contain 2000 positive news and 1000 negative news [62]. Sentiment analyzer for Myanmar news with Support Vector Machine, Naïve Bayes, and Logistics Regression algorithms is implemented with TF-IDF and Countvectorizr features. There are totally 3000 news that contain 2000 positive news and 1000 negative news from Asian Language Tree bank [63]. Comparing performance results of Support Vector Machine and K Nearest Neighbors are presented with TF-IDF feature. Totally, 3000 news data that contain 2000 positive news and 1000 negative news is used for classification [64] .By comparing both approaches, the advantages and limitations of each approach are also considered, so several research problems have been defined. To resolve these conditions, a framework to analyze sentiments by using machine learning, optimizing algorithms with features is developed in this research. Optimizing support vector machine algorithm for Myanmar language did not apply in sentiment analysis system. In reality, SVM has many parameters for tuning to optimize classifier’s accuracy. The common parameter needs to be considered in SVM optimization which is explained in next chapter.

CHAPTER 3
BACKGROUND THEORY

In this chapter, descriptions about concepts and characters of word segmentation tools, feature extraction method, and machine learning algorithms with their common and differences among them are described. Moreover, optimizing approach for support vector machine is also presented.

3.1 Segmentation

The process of breaking down written text into understandable components, such as syllable, words, sentences, or subjects, is known as text segmentation.

3.1.1 Syllable Segmentation

Myanmar script uses no space between syllable and words units for segmentation that represents a significant process in many NLP tasks such as text classification, machine translation, information retrieval and so on. Burmese characters have s round shape and script structure is arranged from left to right. There is no space between words but spaces are usually applied to split words and phrases. Syllables are basic component of words and syllable segmentation is important role in the language processing of Myanmar script. A Myanmar syllable contains one initial consonant, zero or more medial, zero or more vowels and dependent various signs. Independent vowels, independent various signs and digits can be considered as stand-alone syllables. As defined in the Unicode standard, the consonants are saved before vowels [41]. Examples of syllables segmentation are shown in following table.

Table 3.1 Segmentation Result

Myanmar Texts	Segmented Results
က္လစာသယ	|က္လ|စာ|သ|ယ
ေကရာဇ်	|ေ|က|ရာဇ်
ဝကန္တဉာဏ်	|ဝကန္တ|ဉာဏ်|

3.1.2 Word Segmentation

Word segmentation is the limitation of words without word divisor in orthography. Word segmentation is the very important task in language processing tasks [39]. Almost NLP tasks are necessary to segment given sentences into single bounded words prior to other tasks. Dictionary-based and machine learning approaches are existed to split the compound words. In English and many other languages using some form of the Latin alphabet, space is a better measurement of a word splitter (word delimiter), but this idea has boundaries because of the variability with which languages respect apposition and compounds [41]. For example, in text classification, words should be firstly segmented into a range of terms and then extracted features from it and classify to get target class. For Asian languages, most research on this task has focused on the segmentation and morphological analysis of Chinese, Japanese, and Korean, for which the standard, state-of-the-art technique using conditional random fields has achieved satisfactory performance [66]. In Myanmar language, word segmentation is fundamental for language processing task as it does not consider white space to define the words like other Asian languages. This research uses Pyidaungsu python library that used CRF model for word segmentation [51].

3.2 N-Gram

N-gram is used as a language model that belong to bag of word model. N-gram is a series of word with size of length n and is feature selection process in speech and language processing. N-gram with size n=1 known as unigram and size n = 2 termed as bigram and then size n = 3 also defined as trigram. Performance of text analysis is related to text description [36]. The system performances with text representation are more enhancing. N-gram range for example after removing stop words is described in table.

3.3 Feature Transformation

In feature transformation, we apply a mathematical formula to a specific column (feature) and alter the values to make them more relevant for our further analysis. It is a method by which we can improve the performance of our models. It is often referred to as "feature engineering," and it involves constructing new features out of preexisting ones in order to enhance the performance of the model.

3.3.1 CountVectorizer

The Countvectorizer is a simplest method for both text document vector and construct knowns word vocabulary. It also coded document into vector which contains vocabulary. These encoded vectors have vocabulary's length and numbers of each word involved in the document. This vector is changed to array that can be easily applied by machine learning algorithm.

3.3.2 TF-IDF

TF-IDF is a feature weighting and extraction method that is used to determine importance of text in a document. TF-IDF is also type of the bag-of-words (BoW) model. So, it does not need to obtain text position, meaning, circumstance in different documents, etc. [36].

TF is term frequency that describe number of the word showed in individual document in text corpus. Its value increases once the frequency of word inside the document increase. Formula of TF is as follows:

$$TF(t) = \frac{(Number\ of\ times\ term\ t\ appears\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)} \tag{3.1}$$

IDF is inverse document frequency that identified the weight of words in all documents in the corpus. The words that show hardly in the corpus raise in IDF value. Formula of IDF is as follows:

$$IDF(t) = \log\_e \frac{(Total\ number\ of\ documents)}{(Number\ of documents\ with\ term\ t\ in\ it)} \tag{3.2}$$

Then, TF-IDF is calculated as follows:

$$TF - IDF = TF * IDF \tag{3.3}$$

3.4 Machine Learning Algorithms

The key component in making such advances efficient and reliable on the market is the accuracy of machine learning (ML) models. When employed in real-world situations, a model that is more accurate will produce precise results in a variety of cases, enhancing the customer experience.

3.4.1 Support Vector Machine

Support Vector Machine (SVM) is one of the most widely used supervised machine learning algorithms which can be used for both classification and regression challenges. Each data item is put as a point in n-dimensional space (n is number of features) in which each feature has the value of a specific coordinate. Classification is done by finding the hyper-plane that classified the two classes very well. SVM defines the hyper plane that has the highest margin points [45, 46, 48].

The decision function is as follow:

$$f(\chi)=\omega\chi+b \tag{3.4}$$

The hyper plane function is as follow:

$$f(\chi)=\omega\chi+b=0 \tag{3.5}$$

For positive case:

$$f(\chi)=\omega\chi+b>0 \tag{3.6}$$

For negative case:

$$f(\chi)=\omega\chi+b<0 \tag{3.7}$$

Where, x is input vector, w is weight, and b is bias.

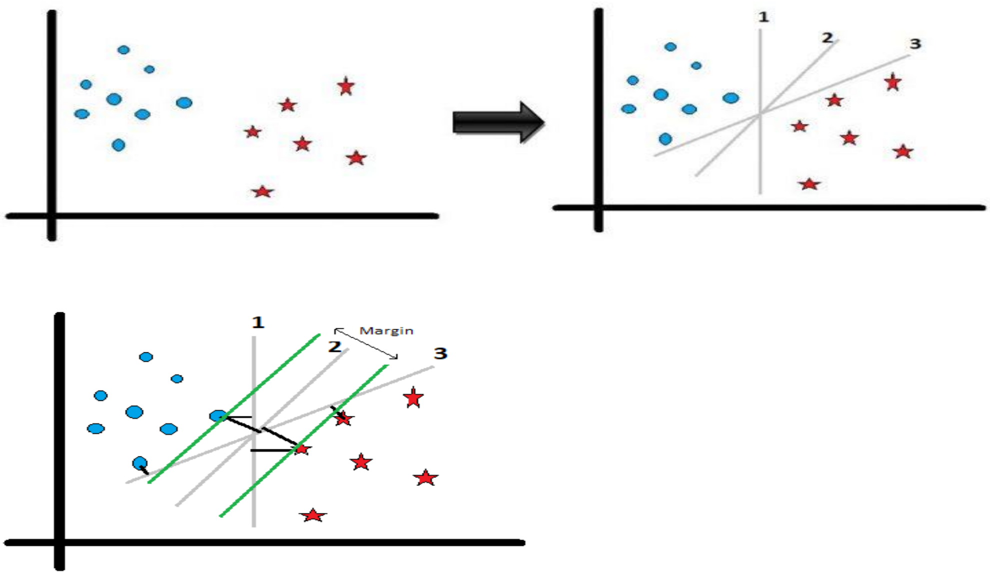

Figure 3.1 Support Vectors, Margin, and Hyperplane

Kernels are similarity functions that return inner products between data points. Kernels can often be computed efficiently even for very high dimensional spaces. Types of kernel functions are as follows:

1. Linear

$K(\vec{x}_i, x) = \vec{x}_i \vec{x}$ (3.8)

$f(x) = \Sigma\alpha_i y_i \ K(\vec{x}_i, \vec{x}) + b = 0$ (3.9)

2. Polynomial

$K(\vec{x}_i, x) = (\vec{x}_i . \vec{x} + 1)^h$ (3.10)

$f(x) = \Sigma\alpha_i y_i \ K(\vec{x}_i, \vec{x}) + b = 0$ (3.11)

3. Radial basis function (RBF)

$K(\vec{x}_i, x) = e^{-||x_i - x||2/2\sigma 2}$ (3.12)

$f(x) = \Sigma\alpha_i y_i \ K(\vec{x}_i, \vec{x}) + b = 0$ (3.13)

4. Sigmoid etc

$K(\vec{x}_i, x) = tanh(Kx_i.x - \sigma)$ (3.14)

$f(x) = \Sigma\alpha_i y_i \ K(\vec{x}_i, \vec{x}) + b = 0$ (3.15)

1. *Linear:* $K(\vec{x}_i, x) = \vec{x}_i \vec{x}$ (3.16)

$f(x) = \Sigma\alpha_i y_i \ K(\vec{x}_i, \vec{x}) + b = 0$

$f(x) = \Sigma\alpha_i y_i \ (\vec{x}_i \vec{x}) + b \quad = 0$

$f(x) = (\Sigma\alpha_i y_i \ \vec{x}_i)\vec{x} + b \quad = 0$

$\Sigma\alpha_i y_i \vec{x}_i = \vec{w}$

Hyper plane: $F(x) = w \vec{x} \vec{\ } + b = 0$

For positive case: $F(x) = w \vec{x} + b = 1$

For negative case: $F(x) = w \vec{x} + b = -1$

$K(\vec{x}_i, \vec{x})$ is kernel function

$\vec{x}$: support vector data

α_i is largange multiplier and y_i is the label of class.

In this research, Linear kernel is used for classification because Linear kernel has less training time than other kernels. Linear SVC is almost used for text classification.

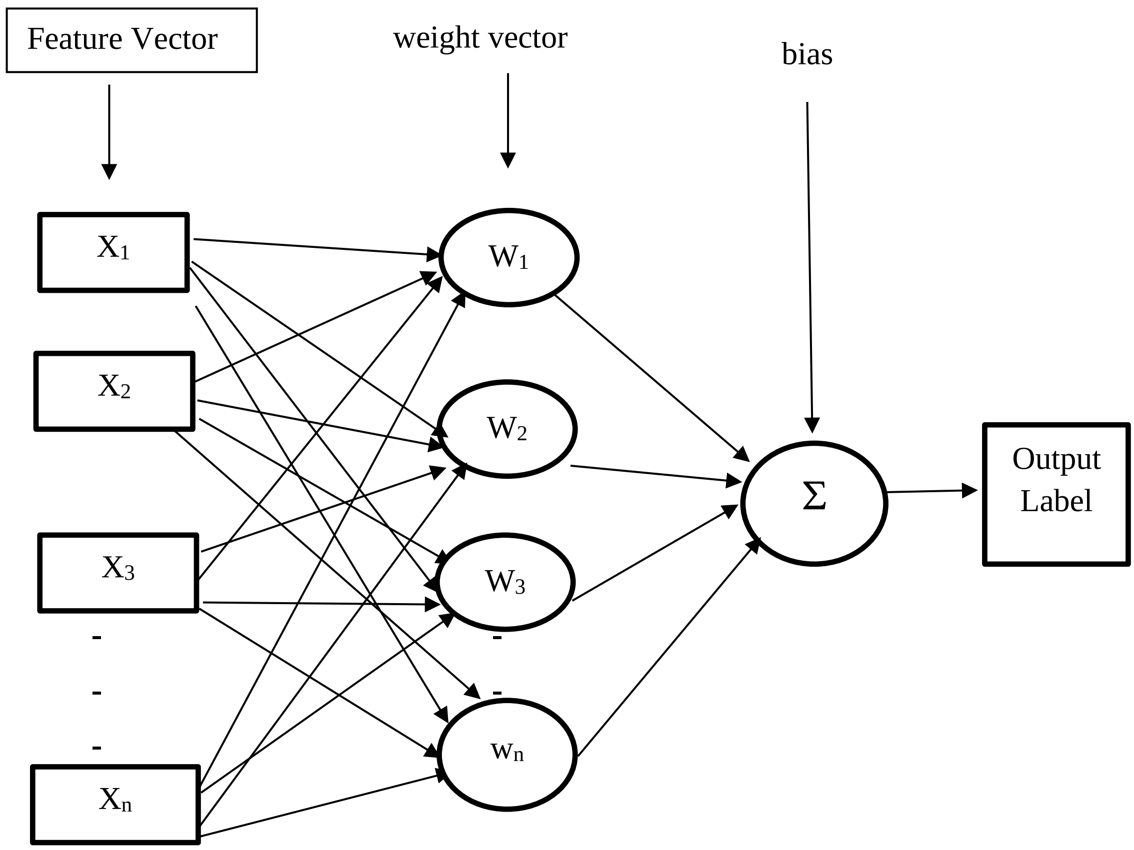

Figure 3.2 Support Vector Machine

Advantages of SVM are as follows:

- When there is a large gap between classes, SVM performs comparatively well.
- In large dimensional spaces, SVM performs better.
- If there are more dimensions than samples, SVM works well in certain situations.
- SVM uses relatively little memory

3.4.2 Random Search Optimization

Optimization is at the heart of machine learning. Rastrigin developed random search that is known for random optimization or random sampling. Identify a search space as a bounded domain of hyperparameter values and randomly sample points in that domain. It is a technique in which random combinations of the hyperparameters are applied to find the best solution for the model. Before the hyperparameter optimization process starts, the number of evaluations in random search must be set in prior. The random search algorithm randomly makes an effort several predefined combinations, and then the hyperparameters are estimate, then the best results are extracted. Random search is proficient and can operate data with large volume well. A random search algorithm is a way that utilizes some kind of randomness or probability and is also called a stochastic algorithm in literature. Random search algorithms are applied for irregular optimization situation, where the goal can also be nonconvex, discontinuous, discrete, or mixed continuous-discrete domain. An optimization process with continuous variables contains several local optima. It is a methodology in which random combinations of the hyperparameters are applied to search the best solution for the model under examination. For example, instead of iterating through all 100,000 samples, only 1000 random samples of hyperparameter sets are countered. The number of calculations in random search must be define in the beginning, earlier than the hyperparameter optimization operation come into existence, and, the complexity of Random Search running n evaluations is O(n)[34],[11],[66]. The algorithms of random search are shown in Figure 3.3 .

Step 1. Initialize algorithm parameters and iteration index

Step 2. Generate a collection of candidate points according to a specific generator and associated sampling distribution.

Step 3. Update iteration sequences based on the candidate points, previous iterates and algorithmic parameters. Also update algorithm parameters

Step 4. If a stopping criterion is met, stop. Otherwise increment iteration and return to Step 1

Figure 3.3 Random Search Algorithm

Process of randomized search with LinearSVC are presented in Figure 3.4.

1.Initiating the number of iterations of the parameter combination 2. Initializing all values of the parameter 3. Iterating random combinations of parameter values based on the number of iterations 4. Conducting training using LinearSVC on training data 5. Evaluating the resulting classifications with test data. 6. Storing the best value from the classification result and the best parameter value

Figure 3.4 Random Search-Linear SVC

Randomized search is the most widely used method for hyper-parameter optimization like Grid search. Advantages of random search algorithm are as follows:

- A greatly broad class of optimization issues
- In spite of being an asymptotic guarantee, provide integration
- Is the best parameter search technique when there are a smaller number of dimensions
- Reduced chance of overfitting and much faster than grid search

Hyperparameters are the parameters that are well defined by the user to manage the process in learning model. To improve the machine learning model, and hyperparameter values are defined prior to learning process. Softening or maximizing of the margin is controlled by a regularization hyperparameter known as the soft-margin parameter, lambda, or capital-C ("C").C is the regularization parameter that controls the trade-off between maximizing the separation margin between classes and minimizing the number of misclassified samples. A value of C indicates a hard margin and no tolerance for violations of the margin. Small positive values identify some violation, whereas large integer values, such as 1, 10, and 100 identify for a much softer margin. Learning algorithm will favor the majority class, as concentrating on it will lead to a better trade-off between classification error and margin maximization. The Linear SVC provide the class_weight argument that can be defined as a model hyperparameter. The class_weight is a dictionary that defines each class label (e.g. 0 and 1) and the weighting to apply to the C value in the calculation of the soft margin. A best practice for using the class weighting is to use the inverse of the class distribution present in the training dataset.

3.4.3 Logistics Regression

Logistic regression is like a discriminant technique for analyzing categorized data. Logistic regression analysis (LRA) is an extended version of multiple regression technique that has categorical outcome. It describes on the regression formulation, odds ratios, confidence limits, likelihood, and deviance. It makes an extensive continuing analysis containing diagnostic extra reports and setup. It can do an independent variable choosing process that going-over for the finest regression model with the lowest independent variables [49, 50, 51].

The statistical formula for logistic regression is

$$\log\left(\frac{p}{1} - p\right) = b_0 + b_1 \tag{3.17}$$

Where, p = binomial proportion, x = explanatory variable $b0$ and $b1$= parameter.

There are many benefits as shown below:

- The training of logistic regression is very effective and easier to implement and analyze.
- It doesn't make any assumptions about how classes are distributed in feature space.
- It is simple to add several categories (multinomial regression) and a natural statistical approach on class predictions.
- It gives an indication of a predictor's suitability (coefficient size), as well as the direction of relationship (positive or negative).
- It classifies unfamiliar records fairly quickly.
- It performs well when the dataset can be linearly separated and has good accuracy for a variety of simple data sets.
- It can use model coefficients to determine the significance of a characteristic.
- Although it is less likely to do so, high-dimensional datasets can cause overfitting in logistic regression.

- To prevent over-fitting in these cases, one may want to take into account regularization (L1 and L2) approaches.

3.4.4 Naïve Bayes

The Bayesian algorithm is applied as one of the probability models for classification. Naive Bayes not depended on features that the inclusion (or exclusion) of a particular feature is not correspond to the existence (or inexistence) of any other [13]. The probability of sentiment class given document is as follow:

$$P(c/d) = P(c)P(d/c)/P(d) \tag{3.18}$$

Where, *c* is class and *d*=document

P(d) has the equal value, so *p(d)* can be cut out. Document has many features, function can be as follows:

$$P\left(\frac{c}{d}\right) = arg\max P(c) \prod_{f \in F} p\left(\frac{f}{c}\right) \tag{3.19}$$

Where, *F*= features vector, *c*= sentiment class, *d*=document

Advantages of working with NB algorithm are:

- Need to have a small number of training data to learn the parameters
- Can be trained relatively fast compared to sophisticated models
- The main disadvantage of NB Algorithm is:
- It is a decent classifier but a bad estimator
- It works well with discrete values but won't work with continuous values (can't be used in a regression) [53]

Advantages of naïve bayes have the following:

- This algorithm is efficient and can greatly reduce processing time.
- Naive Bayes works well for multi-class prediction issues.
- It can outperform other models and needs a lot less training data if its assumption about the independence of characteristics is correct.
- For categorical input variables as opposed to numerical variables, Naive Bayes is more appropriate.

Table 3.2 Sentiment Example

Input Text	မထက်အားပေးနေပါတယ်
Feature	'မ' 'ထက်' 'အား' 'ပေး' 'နေ'
Output	Positive sense

3.4.5 K Nearest Neighbor (KNN)

KNN algorithm is also an idle learning method because the process for the predictions is suspended up to classification. The idea is to train the training dataset and then to test the new instance on the closest neighbors in the training dataset. The algorithm is processed on the rule that minimum distances from the test data to the training samples is chose. After defining rule, a simple measurement is applied to test dataset. Number of distances between the test data and the training samples are calculated by any standard means such as Euclidean distance. The distance of the training samples to the test samples must be less than or equal to Kth smallest distance. The performance of the predictions be determined by the distance measure [45].

$$\text{Euclidean Distance, } \sqrt{\sum_{i=1}^{k}(x_i - y_{i)}{}^2} \tag{3.20}$$

Advantages of working with K Nearest Neighbor algorithm are:

- More Understandable
- No limitation in data
- Can be used in classification and regression
- Performed with multi-class problems

Disadvantages of working with K Nearest Neighbor algorithm are:

- More memory expensive
- Sensitivity of data
- Don’t well on scare variable [38]

3.4.6 MLP

The multilayer perceptron is one of the most frequently used type of neural network. MLP can be applied for classification of linearly inseparable text and for function approximation. The basic features of multilayer perceptron's are as follows:

- Each neural network consists of a nonlinear animating function that has transmission.
- One or more hidden layers are provided.
- High connectivity degrees of network are existed for considering network's weight.

$$u_k = \sum w_{ki} x_i \tag{3.21}$$

$$y_k = \varphi(u_k + b_k \tag{3.22}$$

$w_{ki} = weights, x_i = input, u_k = linear\ combiner, b_k = bias, y_k = output$

The training proceeds in two phases: In the forward phase, the network's weights are determined and the input signal is passed through the network layers. Thus, in this phase, changes are defined for renewing essential and provides outputs results.

In the backward phase, an error signal is emitted by differentiating the network output with designated feedback. The transmitting error signal is transferred through the network layer, but the propagation is done in the backward direction. In this second phase, efficient adjustments are done to the network weight. Adjustments process for the output layer is simple, but it has more challenges for the hidden layers [53, 54]. Advantages are shown in follows:

- It can be applied to resolve challenging nonlinear issues.
- It effectively manages vast volumes of input data.
- Following training, quickly makes predictions.
- Even with less samples, the same accuracy ratio is still possible.

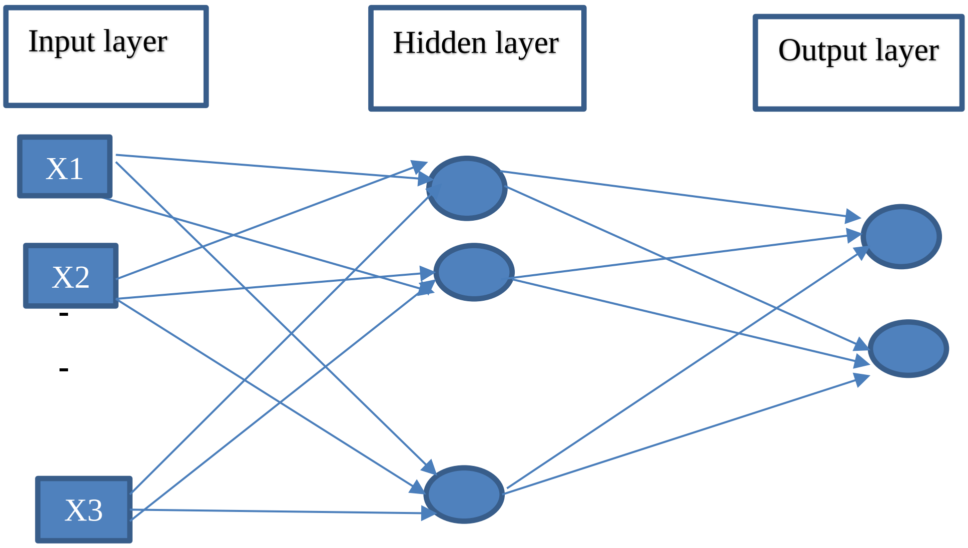

Figure 3.5 Multi-Layer Perceptron

3.5 Summary

After studying the nature and concepts of word segmentation, N-gram, TF-IDF, machine learning algorithms, and optimizing algorithms are learned. Feature are importance for text classification. Feature selection with N-gram is done in this system. Feature to vector transformation is input to classification algorithm. TF-IDF feature weighting and transformation is used for this paper. Machine Learning algorithm are used for classification. Training is firstly processed and then testing is done to predict result from training. Naïve Bayes, Linear SVC use in this paper. Optimization is the problem of searching the set of input parameter to a target model for maximizing performance and minimizing error. Optimizing Linear SVC algorithm has done with Random search algorithm. to have more accuracy to classify sentiment.

CHAPTER 4
SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is an emerging technology that helps in reshapes an organization, product, event, and so on. Opinion mining is also one of the text mining techniques that analyze non-objectivity information held in document. The sentiment is defined as a class subject that has three orientations such as positive negative, and neutral. With the progress of internet technology, social media's data has become a very popular due to the fast outgrowth in information science. Applications of sentiment analysis are mostly utilized in many domains, from finance, and education to organization. News can be taken from many materials and get precious and important knowledge for people.

4.1 Data Collection and Corpus Building

There are two data set in this system. Dataset 1 has fewer data (text sentence) than Dataset2. But, each sentence in Dataset1 has more words or feature than Dataset 2. Dataset 1 is better performance in F1-score than Dataset 2. Dataset 2 is better in accuracy than Dataset 1. Firstly, news data are collected from Asian language treebank [67], www.moi.gov.mm, 7 days newsletters and eleven media. And then, manually tagged sentiment label as positive, negative, and neutral.

News that feel happy, delighted, wonderful are defined as positive news.နားလည်မှု , ဝမ်းမြောက် etc words are used as positive words .

- မြန်မာနိုင်ငံ၏နှစ်၇၀ပြည့်လွတ်လပ်ရေးနေ့အထိမ်းအမှတ်အဖြစ်သမ္မတဦးထင်ကျော် ထံ ကမ္ဘာ့ခေါင်းဆောင်များက ဝမ်းမြောက်ကြောင်း သဝဏ်လွှာ ပေးပို့ ကြသည်

News that feels sad, nervous is defined as negative news. ဘေးအန္တရာယ်, ငလျင်လှုပ် , word etc are used as negative words

- ငလျှင် လှုပ်ခက်မှုကြောင့် အီရန်မြို့တော် တီဟီရန် တွင် ပြည်သူများ လမ်းမများ ပေါ် ပြေးထွက် ခဲ့ရသည်

News that feels no emotion is identified as neutral sense. ယှဉ်ပြိုင် , ပြောပြ , သီချင် words are neutral words.

- ၂၀၁၀ ခုနှစ် အထွေထွေရွေးကောက်ပွဲ တွင် ဝင်ရောက် ယှဉ်ပြိုင် ခဲ့သည် ။
- အခု ဆို နေတဲ့ သီချင်း နာမည် လေး ပြောပြ ပါ
- Rachel အသက် သီချင် ပါတယ်

Table 4.1 Dataset 1

News	Sentence
Positive	3797
Negative	3577
Neutral	1143
Total	8517

Secondly, Myanmar News comments are collected from Myanmar celebrity.com website(www.facebook.myanmarcelebrity.com.mm) for training and testing data. This corpus is constructed by Haymar Su Aung and Win Pa Pa[6] and segmented by also using Pyidaungsu python library[67].

News that feels happy, delighted, wonderful are defined as positive news.

* မထက်အားပေးနေပါတယ်
* ထက်ထက်မိုးဦးကြိုက်တယ်

News that feels sad, nervous is defined as negative news.

* မင်းသားကမထက်ထက်တောင်ခြောက်နေသလားလို့😁

* ထက်ထက်နဲ့မနိုင်းနဲ့အဘွဲ့အညာသူမကခုထိအညာစရိုက်ကမပျောက်ဘူး

ထက်ထက်နဲ့မနိုင်းပါနဲ့မထက်လောက်အရည်အချင်းမရှိပါ

Table 4.2 Dataset 2

News	**Sentence**
Positive news	17974
Negative news	8088
Neutral news	5721
Total	31783

4.2 Preprocessing

Preprocessing is the firstly process in so many natural languages process. Three preprocessing tasks are used in the system.

• Word Segmentation is the basic process in natural language processing task that identified edges of word. Myanmar word segmentation is the task of setting spaces into text by excluding other replacing function. Examples of sentences are as shown in Table 4.3:

Table 4.3 Word Segmentation Example

Input Text	**Segmented Text**
မထက်အားပေးနေပါတယ်	မ ထက် အား ပေး နေ ပါ တယ်
ထက်ထက်မိုးဦးကြိုက်တယ်	ထက် ထက်မိုး ဦး ကြိုက်တယ်

မင်းသားကမထက်ထက်တောင်ခြောက်နေသလားလို့😁	မင်းသား က မ ထက် ထက် တောင် ခြောက် နေ သလား လို့ 😁
ထက်ထက်နဲ့မနှိုင်းနဲ့အဘွဲ့အညာသူမကခုထိအညာစရိုက်ကမပျောက် ဘူး	ထက် ထက် နဲ့ မနှိုင်း နဲ့ အဘွဲ့ အညာ သူမ က ခု ထိ အညာ စရိုက် က မ ပျောက် ဘူး
ထက်ထက်နဲ့မနှိုင်းပါနဲ့မထက်လောက်အရည်အချင်းမရှိပါ	ထက် ထက် နဲ့ မနှိုင်း ပါ နဲ့ မ ထက် လောက် အရည်အချင်း မ ရှိ ပါ

• Tokenization is the separating process for a sequence of text to extract words, phrases, keywords and other elements. Tokens are identified by using white space, punctuation marks or line breaks.

• Stop words are typically used words that are determined to bypass for searching, retrieving, classification, and other tasks. Stop word removal is also a fundamental preprocessing step to provide more effective performance results. Examples of Myanmar stop words are shown in below.

Table 4.4 Example of Myanmar Stop Words

တွင်	ကျွန်တော်	ပါတယ်
အပေါ်	ကျွန်မ	တယ်
အနက်	မှာ	ငါ
အမြဲတမ်း	ဪ	ဤ
အတွင်းတွင်	နော်	ဧ
မကြာမီ	တာ	၌
မတိုင်မီ	ပါ	ဝယ်
ဒါ့အပြင်	လို့	က

After Stop word removing is done, sentences may be as follow:
မ ထက် အား ပေး နေ
ထက် ထက်မိုး ဦး ကြိုက်တယ်
မင်းသား မ ထက် ထက် တောင် ခြောက် နေ သလား 😁
ထက် ထက် မနှိုင်း အာ့ အညာ အညာ စရိုက် မ ပျောက် ဘူး
ထက် ထက် မနှိုင်း မ ထက် လောက် အရည်အချင်း မ ရှိ

4.3 Training

In training phase, the sentiment corpus is used to provide the training data. N-grams range with TF-IDF are used in Linear SVC, Linear SVC with random search optimization and Naive Bayes (NB) classifiers. During this step, the machine learning classifier was used to train and to classify each news in the dataset as carrying either a positive news or a negative news or neural news. In the training stage, the classifier algorithm learns from the labelled news data. As in testing stage, the classifiers have to classify new, non-labelled news

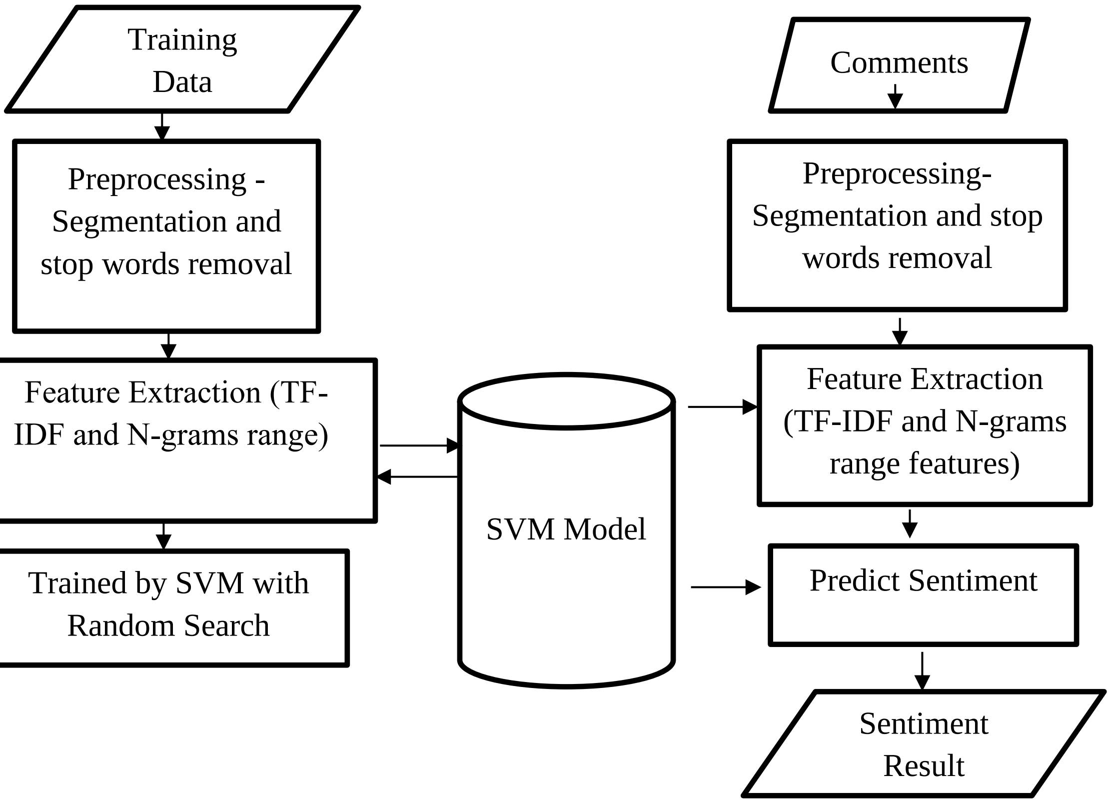

Figure 4.1. System Architecture

4.3.1 Myanmar Word with N-Gram

N-gram is the arrangement of words into the range of size. Due to nature of language, combination of words can increase performance of classifier. For example, combination of မ၊ and လွယ်ပါဘူး words can be more accurate than single word. There are six levels of n-gram in the proposed research work as Table 4.5:

Table 4.5 N-gram Example

Input Text	မ ထက် အား ပေး နေ ထက် ထက်မိုး ဦး ကြိုက်တယ် မင်းသား မ ထက် ထက် တောင် ခြောက် နေ သလား 😁 ထက် ထက် မနိူင်း အာ့ အညာ အညာ စရိုက် မ ပျောက် ဘူး

	ထက် ထက် မနှိုင်း မ ထက် လောက် အရည်အချင်း မ ရှိ
Unigram	['မ' 'ထက်' 'အား' 'ပေး' 'နေ'] [‘ထက်’ ‘ထက်မိုး’ ‘ဦး’ ‘ကြိုက်တယ်’] ['မင်းသား''မ''ထက်''ထက်' 'တောင်' 'ခြောက်' 'နေ' 'သလား' '😄'] ['ထက်''ထက်' 'မနှိုင်း' 'အာ့' 'အညာ''အညာ''စရိုက် 'မ''ပျောက်' 'ဘူး'] ['ထက်''ထက်' 'မနှိုင်း' 'မ''ထက်' 'လောက်''အရည်အချင်း''မ''ရှိ']
Bigram	['မ ထက်''ထက် အား' 'အား ပေး' 'ပေး နေ'] ['ထက် ထက်မိုး' 'ထက်မိုး ဦး' 'ဦး ကြိုက်တယ်'] ['မင်းသား မ'မ ထက်''ထက် ထက်''ထက် တောင်' 'တောင် ခြောက်' 'ခြောက် နေ''နေ သလား' 'သလား 😄'] ['ထက် ထက်''ထက် မနှိုင်း''မနှိုင်း အာ့' 'အာ့ အညာ' 'အညာ အညာ' 'အညာ စရိုက်'''စရိုက် မ' 'မ ပျောက်' 'ပျောက် ဘူး']

	['ထက် ထက်' 'ထက် မနိုင်း''မနိုင်း မ''မ ထက်' 'ထက် လောက်' 'လောက် အရည်အချင်း''အရည်အချင်း မ' 'မ ရှိ']
Trigram	['မ ထက် အား' 'ထက် အား ပေး''အား ပေး နေ'] ['ထက် ထက်မိုး ဦး' 'ထက်မိုး ဦး ကြိုက်တယ်'] ['မင်းသား မ ထက်' 'မ ထက် ထက်' 'ထက် ထက် တောင်''ထက် တောင် ခြောက်' 'တောင် ခြောက် နေ' 'ခြောက် နေ သလား' 'နေ သလား 😁'] ['ထက် ထက် မနိုင်း' 'ထက် မနိုင်း အဲ့' 'မနိုင်း အဲ့ အညာ' 'အဲ့ အညာ အညာ''အညာ အညာ စရိုက်''အညာ စရိုက် မ''စရိုက် မ ပျောက်' 'မ ပျောက် ဘူး'] ['ထက် ထက် မနိုင်း' 'ထက် မနိုင်း မ' 'မနိုင်း မ ထက်''မ ထက် လောက်' 'ထက် လောက် အရည်အချင်း' 'လောက် အရည်အချင်း မ' 'အရည်အချင်း မ ရှိ']
Unigram+Bigram	['မ' 'မ ထက်''ထက်' 'ထက် အား' 'အား' 'အား ပေး' 'ပေး' 'ပေး နေ''နေ'] ['ထက်' 'ထက် ထက်မိုး' 'ထက်မိုး' 'ထက်မိုး ဦး' 'ဦး' 'ဦး ကြိုက်တယ်''ကြိုက်တယ်']

	['မင်းသား' 'မင်းသား မ''မ' 'မ ထက်''ထက်''ထက် ထက်''ထက်' 'ထက် တောင်' 'တောင်' 'တောင် ခြောက်' 'ခြောက်' 'ခြောက် နေ' 'နေ' 'နေ သလား' 'သလား' 'သလား 😄' '😄'] ['ထက်' 'ထက် ထက်''ထက်' 'ထက် မနိုင်း''မနိုင်း' 'မနိုင်း အဘွ''အဘွ' 'အဘွ အညာ''အညာ' 'အညာ အညာ''အညာ' 'အညာ စရိုက်''စရိုက်' 'စရိုက် မ' 'မ' 'မ ပျောက်' ' 'ပျောက်' 'ပျောက် ဘူး' 'ဘူး'] ['ထက်' 'ထက် ထက်' 'ထက်''ထက် မနိုင်း' 'မနိုင်း' 'မနိုင်း မ''မ' 'မ ထက်' 'ထက်' 'ထက် လောက်' 'လောက်' 'လောက် အရည်အချင်း' 'အရည်အချင်း' 'အရည်အချင်း မ' 'မ''မ ရှိ' 'ရှိ']
Bigram+Trigram	['မ ထက်' 'မ ထက် အား''ထက် အား' 'ထက် အား ပေး' 'အား ပေး' 'အား ပေး နေ' 'ပေး နေ'] ['ထက် ထက်မိုး' 'ထက် ထက်မိုး ဦး' 'ထက်မိုး ဦး' 'ထက်မိုး ဦး ကြိုက်တယ်' 'ဦး ကြိုက်တယ်'] ['မင်းသား မ' 'မင်းသား မ ထက်''မ ထက်' 'မ ထက် ထက်' 'ထက် ထက်' 'ထက် ထက် တောင်''ထက် တောင်'

	'ထက် တောင် ခြောက်''တောင် ခြောက်' 'တောင် ခြောက် နေ' 'ခြောက် နေ' 'ခြောက် နေ သလား' 'နေ သလား' 'နေ သလား 😁' 'သလား 😁'] ['ထက် ထက်' 'ထက် ထက် မနိုင်း''ထက် မနိုင်း' 'ထက် မနိုင်း အ့ာ' 'မနိုင်း အ့ာ' 'မနိုင်း အ့ာ အညာ' 'အ့ာ အညာ' 'အ့ာ အညာ အညာ''အညာ အညာ' 'အညာ အညာ စရိုက်''အညာ စရိုက်' 'အညာ စရိုက် မ' 'စရိုက် မ' 'စရိုက် မ ပျောက်' 'မ ပျောက်' 'မ ပျောက် ဘူး' 'ပျောက် ဘူး'] ['ထက် ထက်' 'ထက် ထက် မနိုင်း' 'ထက် မနိုင်း' 'ထက် မနိုင်း မ''မနိုင်း မ' 'မနိုင်း မ ထက်' 'မ ထက်' 'မ ထက် လောက်''ထက် လောက်' 'ထက် လောက် အရည်အချင်း' 'လောက် အရည်အချင်း' 'လောက် အရည်အချင်း မ''အရည်အချင်း မ' 'အရည်အချင်း မ ရှိ' 'မ ရှိ']
Unigram+Bigram+Trigram	['မ' 'မ ထက်' 'မ ထက် အား''ထက်' 'ထက် အား' 'ထက် အား ပေး' 'အား' 'အား ပေး' 'အား ပေး နေ' 'ပေး' 'ပေး နေ ''နေ'] ['ထက်' 'ထက် ထက်မိုး' 'ထက် ထက်မိုး ဦး' 'ထက်မိုး' 'ထက်မိုး ဦး' 'ထက်မိုး ဦး ကြိုက်တယ်'' 'ဦး' 'ဦး ကြိုက်တယ်'' ကြိုက်တယ်]

	['မင်းသား' 'မင်းသား မ' 'မင်းသား မ ထက်' 'မ' 'မ ထက်' 'မ ထက် ထက်' 'ထက်''ထက် ထက်' 'ထက် ထက် တောင်' 'ထက်' 'ထက် တောင်' 'ထက် တောင် ခြောက်' 'တောင်' 'တောင် ခြောက်' 'တောင် ခြောက် နေ' 'ခြောက်' 'ခြောက် နေ' 'ခြောက် နေ သလား' 'နေ' 'နေ သလား' 'နေ သလား 😁' 'သလား' 'သလား 😁' '😁'] ['ထက်' 'ထက် ထက်' 'ထက် ထက် မနှိုင်း''ထက်''ထက် မနှိုင်း' 'ထက် မနှိုင်း အာ့''မနှိုင်း' 'မနှိုင်း အာ့' 'မနှိုင်း အာ့ အညာ' 'အာ့' 'အာ့ အညာ' 'အာ့ အညာ အညာ''အညာ''အညာ အညာ' 'အညာ အညာ စရိုက်'အညာ' 'အညာ စရိုက်' 'အညာ စရိုက် မ''စရိုက်' 'စရိုက် မ' 'စရိုက် မ ပျောက်' 'မ' 'မ ပျောက်' 'မ ပျောက် ဘူး' 'ပျောက်' 'ပျောက် ဘူး' 'ဘူး'] ['ထက်' 'ထက် ထက်' 'ထက် ထက် မနှိုင်း' 'ထက် မနှိုင်း' ' ထက် မနှိုင်း မ' 'မနှိုင်း' 'မနှိုင်း မ' 'မနှိုင်း မ ထက်''မ' 'မ ထက်' 'မ ထက် လောက်''ထက်' 'ထက် လောက်' 'ထက် လောက် အရည်အချင်း''လောက်' 'လောက် အရည်အချင်း' 'လောက် အရည်အချင်း မ ' 'အရည်အချင်း' 'အရည်အချင်း မ' 'အရည်အချင်း မ ရှိ''မ' 'မ ရှိ' 'ရှိ']

4.3.2. TFIDF

The required different parameters setting for TF-IDF are presented in the following Table [4.6].

Table 4.6 TF-IDF Parameters

N-gram range	(1,1), (2,2), (3,3), (1,2), (2,3), (1,3)
Max _features	5000
use_idf	True
analyzer	'word'
lowercase	False
stop_words	stopwords
tokenizer	tokenize
min_df	1
sublinear_tf	True

In Table 4.7, TF-IDF values are described with array. Example of unigram feature are shown with vector form as follow:

['ကြိုက်တယ်' 'ခြောက်' 'စရိုက်' 'တောင်' 'ထက်' 'ထက်မိုး' 'နေ' 'ပေး' 'ပျောက်'

'ဘူး' 'မ' 'မင်းသား' 'မနှိုင်း' 'ရှိ' 'လောက်' 'သလား' 'အညာ' 'အရည်အချင်း'

'အဝ့' 'အား' 'ဦး' '😄']

Table 4.7 TF-IDF Values for Unigram

(နေ)0.4513377882617187 (ပေး)0.5594215551143539 (အား) 0.5594215551143539	(ဘူး) 0.3181980160895788 (ပျောက်) 0.3181980160895788 (စရိုက်) 0.3181980160895788

(ထက်)0.2665673684153409	(အညာ) 0.6363960321791576
(မ) 0.31516834601529353	(အခု) 0.3181980160895788
(ကြိုက်တယ်) 0.5566685141652766	(မနှိုင်း) 0.25672015584344143
(ဦး)0.5566685141652766	(ထက်) 0.30324611916908584
(ထက်မိုး) 0.5566685141652766	(မ) 0.17926721185385933
(ထက်)0.26525552965220073	(ရှိ) 0.3789392784872758
(😀) 0.3813421777365623	(အရည်အချင်း) 0.3789392784872758
(သလား) 0.3813421777365623	(လောက်) 0.3789392784872758
(ခြောက်) 0.3813421777365623	(မနှိုင်း) 0.305725823887815
(တောင်) 0.3813421777365623	(ထက်) 0.5416997897135692
(မင်းသား) 0.3813421777365623	(မ) 0.42697555912607005
(နေ) 0.3076644678722554	
(ထက်) 0.36342318187661254	
(မ) 0.21484153108568452	

4.3.3 Classification

This task in the sentiment analysis relies on the algorithm applied for classification. Random search optimization with Linear SVC and Naive Bayes (NB) classifiers are used in this research. In this step, the machine learning algorithms was used to train and to classify each news in the dataset as bring either a positive news or a negative news or neural news. In the training stage, the classifier algorithm learns from the labelled news data. As in testing stage, the classifier has to classify new, non-labelled news. The input sentence is မထက်ပြောတာလေးကချစ်ဖို့ကောင်းတယ်. This testing sentence is segmented into following and considered as an input into the system.

'မ', 'ထက်','ပြောတာ', 'လေး', 'က', 'ချစ်', 'ဖို့', 'ကောင်းတယ်'

Figure 4.2. Input Text

This sentence is changed into feature vector by using TFIDF method. As in the beginning process, predetermined words are collected and applied as the beginning population.

Table 4.8 Predefined Sense with Collected Sentences

Segmented sentences	Sense
မ ထက် အား ပေး နေ ပါ တယ်	Positive
ထက် ထက်မိုး ဦး ကြိုက်တယ်	Positive
မင်းသား က မ ထက် ထက် တောင် ခြောက် နေ သလား လို့ 😁	Negative
ထက် ထက် နဲ့ မနှိုင်း နဲ့ အာ့ အညာ သူမ က ခု ထိ အညာ စရိုက် က မ ပျောက် ဘူး	Negative

These sentences are trained as the form of feature vectors by using TFIDF transformation. These feature vectors are also trained to classify with the machine learning algorithm- Naïve Bayes, Linear SVC, and Linear SVC with Random search and to store the classification models for the future classification task.

The required parameter setting for Linear SVC with Random search are described in Table 4.13.

Table 4.9 Parameter Values for Random Search Optimization

Estimator	Linear SVC
param_distributions	‘C’: 0.01, 0.1, 1, 10, 100, 1000

4.4 Summary

This chapter shows the development of the sentiment analysis system and proposed algorithms which are applied in the processing step. The feature extraction and selection algorithms generate feature that are transformed into vector for used in classification. Classification algorithms are used to classify sentiment. Optimization algorithms are applied to improve proposed algorithm. The quality of system is measured by hold-out validation methods and cross validation, performance results are discussed in chapter 5.

CHAPTER 5
EXPERIMENT AND EVALUATION

This chapter shows experimentations and evaluations of the sentiment analysis process discussed in Chapters 3 and 4. These consist of the evaluation of sentiment analysis and the contribution of each strategy combined into the system. It also learns the performance of the optimized approach compared with baseline Linear SVC and Naïve Bayes machine learning approaches. Finally, the performance of optimized method was investigated that applied features extracted from feature selection and transformation methods is more effective than other methods.

5.1 Experimental Study

The proposed sentiment analysis method considers the essential factors to generate sentiment orientation by analyzing performance quality factors and measurement metrics. It uses Myanmar news in order to test and train data. The proposed research work uses Scikitlearn python library[54] on Anaconda jupyter notebook.

5.2 Datasets

The above machine learning algorithms were tested on the Myanmar news dataset. There are totally 2 news datasets in Table 5.1. The 80 % of dataset is used for training and 20% of dataset is used for testing.

Table 5.1 Sentiment Datasets

Sense	Dataset 1	Dataset 2
Positive	17974	3797
Negative	8088	3577
Neutral	5721	1143
Total	31783	8517

5.3 Evaluation Metrics

The cross-validation and hold out method is applied in the system. The evaluation metrics such as CV score, precision, recall, and F1 measure are calculated using equation 5-8.

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (5.1)$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \quad (5.2)$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \quad (5.3)$$

$$\text{F-Measure}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \quad (5.4)$$

TP =true positive

FP=false positive

TN =true negative

FN =false negative

CV-Score is one of the major issues concerning testing in learning models. It is a method for evaluation and verification the accuracy of a machine learning. It involves setting aside a particular sample from a dataset on which the model has not been trained. Later, this sample is used to test the model and assess it.

5.4 Experimental Results

The proposed random search with Linear SVC model has the superior result rather than baseline Linear SVC and Multinomial NB. It is shown that optimize Linear SVC models overcome baseline Linear SVC and Multinomial NB model because optimized Linear SVC performs parameters tuning process with random search to get better accuracy.

5.4.1 10-fold Cross Validation

Cross validation is an effective evaluation technique that is mostly suggested to determine model performance. It is an empowering strategy since it is clear and provides test results that is more accurate. The dataset is shuffled in either case and divided into 10 equal lengths for 10-fold cross validation. One of the ten clusters is retained for testing validation, while the other nine are recycled for training. To obtain a single result, the data from each fold are then averaged. Table 5.2 and Table 5.4

represent the results of 10-fold cross validation in training, with Table 5.2 having the highest score value (75.72) in randomized search with Linear SVC in combination of unigram, bigram, and trigram features, and Table 5.4 having the highest value (77.11) in unigram features with randomized search and Linear SVC. Tenfold cross validation scores for testing are displayed in Tables 5.3 and 5.5. The highest value (68.32) in Table 5.3 was discovered using a randomize search with Linear SVC in conjunction with a bigram and a unigram.

Table 5.2 Accuracy Score in Training for Dataset 1(10 fold)

TFIDF	Multi NB	Linear SVC	Random-Linear SVC
Unigram	71.18	73.94	**75.35**
Bigram	70.19	71.95	**73.25**
Trigram	64.79	65.72	**67.11**
Unigram+Bigram	71.65	75.04	**75.48**
Bigram+Trigram	69.32	71.12	**71.88**
Unigram+Bigram+Trigram	71.31	74.42	**75.72**

In Table 5.2, training accuracy for dataset1 with 10 fold cross validation is shown. Combination of unigram and bigram with Random -Linear SVC is the highest accuracy score value (75.48). Trigram with naïve bayes is the lowest accuracy value (64.79). Each feature increased accuracy value in Random -Linear SVC.

Table 5.3 Accuracy Score in Testing for Dataset 1(10 fold)

TFIDF	Multinomial NB	Linear SVC	Random-Linear SVC
Unigram	64.13	67.19	**67.24**
Bigram	62.27	60.67	**62.01**

Trigram	56.21	54.81	**55.44**
Unigram+Bigram	63.98	67.39	**68.32**
Bigram+Trigram	62.48	60.09	**61.85**
Unigram+Bigram+Trigram	64.29	66.82	**67.96**

In the experiment of testing Dataset 1, combination of Unigram and Bigram increased accuracy value in Random-Linear SVC(68.32) shown in Table 3.3.

Table 5.4 Accuracy Score in Training for Dataset 2(10-fold)

TFIDF Feature	**Multinomial NB**	**Linear SVC**	**Random-Linear SVC**
Unigram	75.87	77.00	**77.11**
Bigram	70.05	69.49	**69.80**
Trigram	63.31	63.50	**63.52**
Unigram+Bigram	76.09	76.48	**77.07**
Bigram+Trigram	69.73	69.02	**69.47**
Unigram+Bigram+Trigram	76.05	76.46	**76.98**

In training stage in Dataset 2, Unigram feature has the greatest accuracy score in Random-Linear SVC. Trigram feature has the lowest accuracy value (63.31) in Naïve Bayes as shown in Table 5.4.

Table 5.5 Accuracy Score in Testing for Dataset 2(10 fold)

TFIDF	**Multinomial NB**	**Linear SVC**	**Random-Linear SVC**
Unigram	70.28	73.09	**73.67**
Bigram	66.09	66.81	**69.80**
Trigram	60.41	62.10	**63.52**
Unigram+Bigram	72.19	73.02	**74.42**
Bigram+Trigram	66.32	66.49	**67.12**
Unigram+Bigram+Trigram	72.16	73.05	**74.09**

In testing experiment in Dataset 2, integration of Unigram and Bigram has greatest accuracy value in Random-Linear SVC. In Random-Linear SVC, accuracy increased to 1% in all features in Table 5.5.

5.4.2 5-fold Cross Validation

Cross validation is an evaluation method and is mainly employed to estimate execution of model. It is a popular technique due to understandable and give more accurate for test data. In 5-fold cross validation, the dataset is shuffle in anyway and divide it into 5 equal size length. Of the 5 clusters, one cluster is kept for validation in testing, and the remaining 4 clusters are recycled for training. The results from each fold are then be average to get a single result. The results from 5-fold cross validation are presented in Table 5.6, Table 5.7, Table 5.8, and Table 5.9. In table 5.6, combination of unigram and bigram in Linear SVC with random search is the highest score value(70.18). In table 5.7, Linear SVC optimized with random search has greatest score value in combination of unigram and bigram(74.12). Optimized with Linear SVC in combination of unigram and bigram is the greatest score(76.85) in table 5.8. In table 5.9, combination of unigram and bigram has the highest score(74.12) in Linear SVC with random search.

Table 5.6 Accuracy Score for Training in Dataset 1

TFDF	**Multinomial NB**	**Linear SVC**	**Randomize Linear SVC**
Unigram	70.64	73.85	**74.43**
Bigram	69.35	71.36	**72.13**
Trigram	64.27	65.45	**65.44**
Unigram+Bigram	71.05	74.52	**75.18**
Bigram+Trigram	68.59	70.37	**70.85**
Unigram+Bigram+Trigram	70.69	74.08	**74.69**

In training stage with Dataset1, compound of Unigram and Bigram gives highest accuracy(75.18) in Random-Linear SVC. In each feature, performance increased to 0.5 % compared to Linear SVC and about 1-4 % compared to Naïve bayes.

Table 5.7 Accuracy Score for Testing in Dataset 1

TFIDF	**Multinomial NB**	**Linear SVC**	**Randomize Linear SVC**
Unigram	63.67	72.42	**73.37**

Bigram	61.49	66.12	**67.08**
Trigram	56.57	61.73	**61.81**
Unigram+Bigram	63.56	72.74	**74.12**
Bigram+Trigram	62.11	65.79	**66.64**
Unigram+Bigram+Trigram	63.25	72.79	**74. 00**

Table 5.8 Accuracy Score for Training in Dataset 2

TFIDF	**Multinomial NB**	**Linear SVC**	**Randomize Linear SVC**
Unigram	75.49	76.68	**76.74**
Bigram	69.83	69.08	**69.57**
Trigram	63.13	63.22	**63.29**
Unigram+Bigram	75.66	76.33	**76.85**
Bigram+Trigram	69.64	68.88	**69.26**
Unigram+Bigram+Trigram	75.69	76.33	**76.79**

In training experiment, compound of Unigram, Bigram, and Trigram provides greatest accuracy (76.79%) in Random-Linear SVC. Trigram has worse accuracy in Naïve Bayes in Table 5.8.

Table 5.9 Accuracy Score for Testing in Dataset 2(5 fold)

TF-IDF	**Multinomial NB**	**Linear SVC**	**Randomize Linear SVC**

Unigram	69.40	72.42	**73.37**
Bigram	65.69	66.12	**67.08**
Trigram	59.97	61.73	**61.81**
Unigram+Bigram	71.53	72.74	**74.12**
Bigram+Trigram	65.94	65.79	**66.64**
Unigram+Bigram+Trigram	71.89	72.79	**74.00**

In testing phase, addition of Unigram and Bigram feature gives the best accuracy score(74.12) in Random-Linear SVC in Table 5.9.

5.4.3 Confusion Matrix

A confusion matrix is a table that is generally used to show the performance of a classification model on a set of tests set for which the actual values are seen. A confusion matrix views and summarizes the performance of a classification model. Computation a confusion matrix provides a better way of what model is taking right and what types of errors it is done. The better way to evaluate the performance of a classifier is to scan at the confusion matrix. It is a table which shows different combinations of predicted and true values. The accuracy of classification systems could be assessed using a calculative statistical evaluator known as the confusion matrix, which is made up of false positives, true positives, false negatives, and true negatives. Table5.10, Table 5.11, Table 5.12, 5.13, 5.14, and Table 5.15 show result of confusion matrix for baseline naïve bayes, Linear SVC and Linear SVC optimize with random search in TFIDF features in real dataset. In all tables, positive sense has more correct senses than other because positive news has greatest data.

Table 5.10. Confusion Matrix of Multinomial NB with TFIDF Vectorizer for Dataset 1

Features	Class	Negative	Neutral	Positive
Unigram	Negative	615	10	85
	Neutral	124	196	137
	Positive	140	32	593
Bigram	Negative	594	32	84
	Neutral	129	221	107
	Positive	165	62	538
Trigram	Negative	467	26	217
	Neutral	114	197	146
	Positive	131	35	599
Unigram + Bigram	Negative	621	23	66
	Neutral	103	217	104
	Positive	148	52	565
Bigram+ Trigram	Negative	577	41	92
	Neutral	139	209	109
	Positive	167	60	538
Unigram +Bigram+ Trigram	Negative	616	30	64
	Neutral	138	218	101
	Positive	154	60	551

Table 5.11. Confusion Matrix of Linear SVC with TFIDF Vectorizer for Dataset 1

Features	Class	Negative	Neutral	Positive
Unigram	Negative	568	34	108
	Neutral	49	324	84

	Positive	91	89	585
Bigram	Negative	540	43	127
	Neutral	53	312	92
	Positive	140	84	541
Trigram	Negative	416	56	238
	Neutral	53	270	134
	Positive	101	77	587
Unigram + Bigram	Negative	564	42	104
	Neutral	57	322	78
	Positive	101	94	507
Bigram+ Trigram	Negative	532	46	132
	Neutral	55	310	92
	Positive	133	88	544
Unigram +Bigram+ Trigram	Negative	561	43	106
	Neutral	53	322	82
	Positive	100	90	575

Table 5.12. Confusion Matrix of Random search-LinearSVC with TFIDF Vectorizer for Dataset 1

Features	Class	Negative	Neutral	Positive
Unigram	Negative	565	9	109
	Neutral	41	352	64
	Positive	89	110	566
Bigram	Negative	532	56	122
	Neutral	42	345	70
	Positive	141	100	524
Trigram	Negative	401	65	244
	Neutral	35	331	91
	Positive	96	95	574

Unigram+Bigram	Negative	563	48	99
	Neutral	42	355	60
	Positive	101	104	560
Bigram+ Trigram	Negative	526	60	124
	Neutral	33	372	52
	Positive	131	109	525
Unigram +Bigram+ Trigram	Negative	550	49	111
	Neutral	26	384	47
	Positive	93	102	570

Table 5.13. Confusion Matrix of Multinomial Naïve Bayes with TFIDF Vectorizer for Dataset 2

Features	Class	Negative	Neutral	Positive
Unigram	Negative	1175	45	335
	Neutral	359	305	496
	Positive	253	41	3348
Bigram	Negative	905	79	572
	Neutral	233	204	723
	Positive	222	56	3364
Trigram	Negative	480	55	1020
	Neutral	97	106	957
	Positive	86	27	3529
Unigram + Bigram	Negative	1204	72	279
	Neutral	355	350	455
	Positive	289	53	3300
Bigram+ Trigram	Negative	891	96	568
	Neutral	234	14	712
	Positive	244	61	3337
Unigram +Bigram+ Trigram	Negative	1200	77	278
	Neutral	347	364	449

	Positive	293	59	3290

Table 5.14. Confusion Matrix of Linear SVC with TFIDF Vectorizer for Dataset 2

Features	Class	Negative	Neutral	Positive
Unigram	Negative	1100	197	258
	Neutral	264	522	374
	Positive	235	151	3256
Bigram	Negative	852	140	563
	Neutral	198	312	650
	Positive	235	113	3294
Trigram	Negative	508	67	980
	Neutral	107	117	936
	Positive	103	39	3500
Unigram + Bigram	Negative	1100	199	256
	Neutral	148	259	161
	Positive	236	148	3258
Bigram+ Trigram	Negative	829	147	579
	Neutral	204	291	665
	Positive	230	110	3302
Unigram +Bigram+ Trigram	Negative	1084	202	269
	Neutral	282	519	359
	Positive	242	152	3248

Table 5.15. Confusion Matrix of Random Search-Linear SVC with TFIDF Vectorizer for Dataset 2

Features	Class	Negative	Neutral	Positive
Unigram	Negative	1158	202	195
	Neutral	282	597	281
	Positive	307	207	3128
Bigram	Negative	898	159	498
	Neutral	208	342	610
	Positive	249	152	3241
Trigram	Negative	508	95	952
	Neutral	103	135	922
	Positive	107	51	3484
Unigram + Bigram	Negative	1161	199	195
	Neutral	289	582	289
	Positive	302	207	3133
Bigram+ Trigram	Negative	885	161	509
	Neutral	212	320	628
	Positive	256	150	3236
Unigram +Bigram+ Trigram	Negative	1153	205	197
	Neutral	288	582	290
	Positive	316	206	3120

5.4.4 F1-Score

A Classification report is utilized to measure the quality of classification algorithm. How many predictions are True and how many are False? True Positives, False Positives, True negatives and False Negatives are used to determine the metrics of a classification report. The classification report displays the precision, recall, F1, and support scores for the model. Precision is the fraction of predicted instances among the actual instances. Recall is the fraction of predicted instances that have been actually over the total number of instances. The F1 score is a balanced harmonic division of precision and recall. F1 scores are smaller than accuracy scores as precision and recall are embed into their computation. When data is unbalanced, a common performance metric for classification is the F1 score. It mainly used to compare the effectiveness of two classifiers. Table 5.16 and Table 5.17 show performance with f1-score for Multinomial NB, Linear SVC , and Linear SVC optimized with random search with TFIDF features. In each algorithm, positive news has largest data and highest precision, recall, and f1-score. In table 5.16, unigram feature in Linear SVC optimized with random search is greatest f1-score value for data1. In table 5.17, Linear SVC optimized with random search has highest performance value in unigram for data2. It is better results for unigrams than other because of higher order n-grams have a data sparsity problem.

Table 5.16 F1-Score for Dataset 1

TFIDF Feature	Multinomial NB	Linear SVC	Random-Linear SVC
Unigram	71.51	76.44	**77.94**
Bigram	69.38	72.07	**72.50**
Trigram	64.64	65.70	**67.02**
Unigram+Bigram	71.80	75.35	**76.92**
Bigram+Trigram	67.84	71.72	**73.61**
Unigram+Bigram+Trigram	70.93	75.46	**77.86**

The experiment of Random-Linear SVC has greatest performance in Unigram(77.94%). Trigram with Naïve Bayes is the worst performance (64.64).

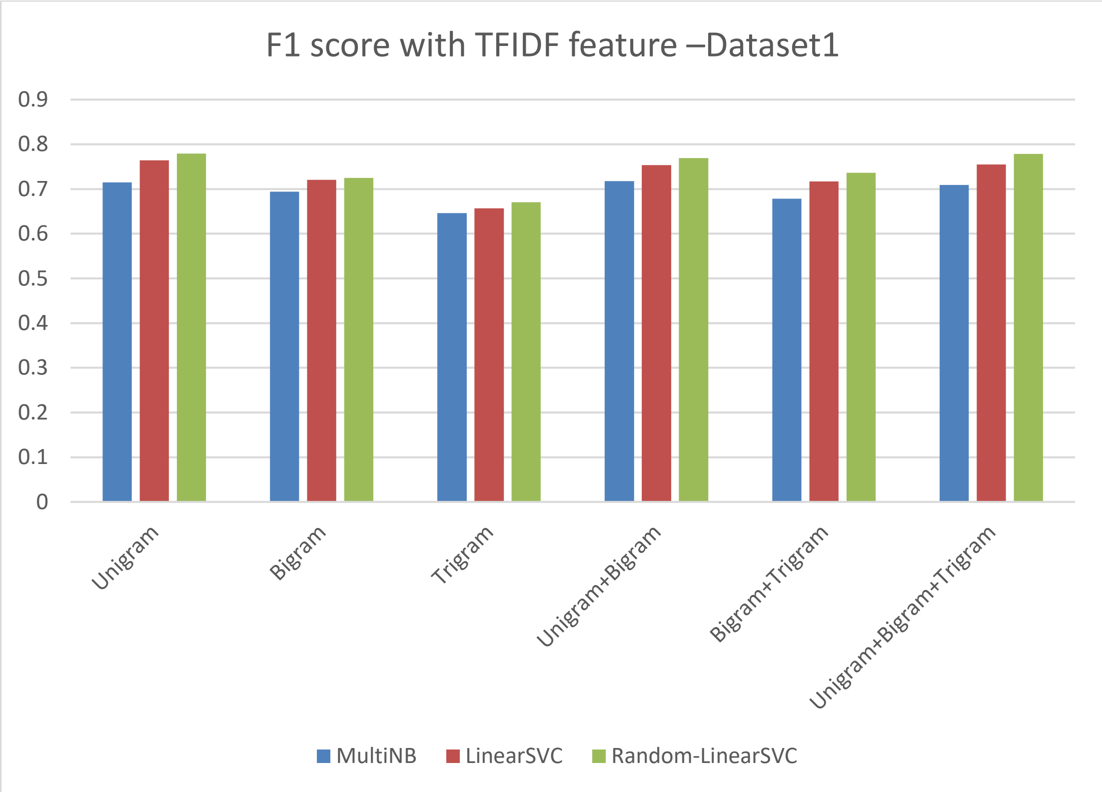

Figure 5.1 F1 Score for Data1

Table 5.17 F1 score for Dataset 2

TFIDF Feature	Multinomial NB	Linear SVC	Random-Linear SVC
Unigram	73.43	75.99	**76.68**
Bigram	66.58	67.59	68.45
Trigram	57.66	58.30	58.71
Unigram+Bigram	74.38	76.23	76.52
Bigram+Trigram	66.31	66.82	67.67
Unigram+Bigram+Trigram	74.51	75.59	76.21

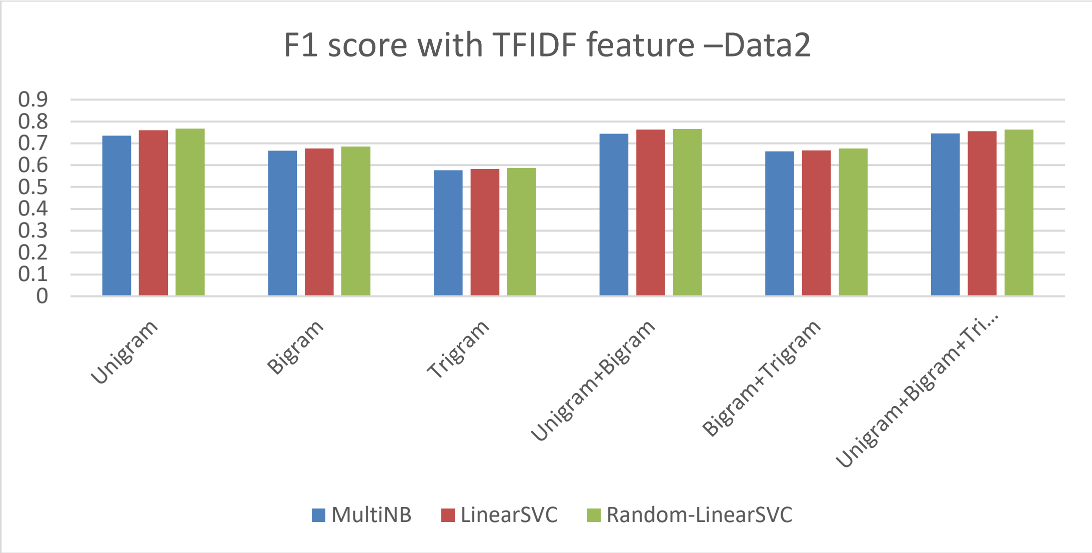

Figure 5.2 F1 Score for Dataset 2

5.4.5 Performance Analysis

In the following sentences, optimized LinearSVC is more accurate in testing result.

- တော်လိုက်တော့မတင်ညွှန့်အမျိုးမျိုးပြောနေ is positive in LinearSVC and it is actually in negative in optimized LinearSVC.
- ကိုလွင်မိုးခင်များကိုယ်တိုင်တည့်တည့်မပြောရဲပါလား is neutral in LinearSVC and it is correctly negative in optimized LinearSVC.
- လှိုင်သာယာအိုးအိမ်တရားစွဲသူတွေအဖမ်းခံရ။ is positive in LinearSVC and actually predicted negative in optimized LinearSVC.
- **NYDC**ကုမ္ပဏီသည်ရန်ကုန်မြို့သစ်အကောင်အထည်ဖော်ရာတွင်ရင်းနှီးမြှုပ်နှံသူဖိတ်ခေါ်ရန်နှင့်ရင်းနှီးမြှုပ်နှံသူများကိုကြီးကြပ်ရန်ဖြစ်ကြောင်းတိုင်းအစိုးရကကြေညာထားသည်။ is wrongly positive in baseline LinearSVC and actually predicted as neutral in optimized LinearSVC.
- ရှမ်းတိုင်းရင်းသားများဒီမိုကရက်တစ်ပါတီသည်မြန်မာနိုင်ငံရှိနိုင်ငံရေးပါတီတစ်ခုဖြစ်ပြီးကျားဖြူပါတီဟုလည်းခေါ်ဆိုကြသည်။ is incorrectly predicted negative in LinearSVC and correctly neutral news in optimized LinearSVC.
- အတွေ့အကြုံရှိပြီးသားမလား is wrongly negative news in LinearSVC,but it is really predicted as neutral in optimized LinearSVC.
- အမှန်ကိုသာပြောဆိုခြင်း။ is wrongly negative news in LinearSVC,but it is really predicted as positive news in optimized LinearSVC.

- သူမိန်းမကအညာသူလေကောင်းမှာပေါ့ is wrongly neutral news in LinearSVC,but it is really predicted as positive in optimized LinearSVC.
- မကြီးစန်းကပိုကျက်သရေရှိနေတယ်မြင်မိတယ် is wrongly predicted as negative news in LinearSVC, but it is really predicted as positive in optimized LinearSVC.

5.4.6 Error Analysis

Error analysis is showed in following related examples.

- In neutral sense, following sentences are falsely predicted :
- ကိုမြတ်သူဘာတွေလုပ် as negative sense
- လူတွေကတော့ပြောကြမှာပါ ကိုယ့်ဘဝအတွက် ကိုယ်လုပ်သင့်တာ လုပ်ပါ as negative sense.
- In negative sense, incorrect sense are predicted in below sentences.
- ညီမလေးနဲ့ခလေးကံမကောင်းတာပါဘဝပေးကံဆိုးလွန်းတယ် 😢 😢 😢 as positive sense.
- အနာဂတ်လူငယ်လေးတွေ အတွက်ရင်လေးစရာပဲ as positive sense.
- ထပ်တူခံစားရတယ်ခံစားဘူးတယ် ဒါပေမယ့် တစ်နေ့ကြုံတွေ့ရမယ် ဆိုတာကို အကိုပြောသိလိုနားလည်းပါတယ် 😥 ။ as negative news.
- စိတ်မကောင်းပါဘူးကိုယ်ချင်းစာပါတယ် ။ as positive news.
- In positive case, following sentences are wrongly predicted:
- ချစ်တယ်ဘယ်တော့မှမမေ့ပါဘူး ။ as negative news
- အခုလို.Topfan ဖြစ်အောင် အဖက်ဖက်ကကူညီပီး ဘေလ်ရောင်းပေးတဲ့ ကိုယ်စားလှယ်များကိုလည်း ကျေးဇူးတင်ရှိပါသည် as neutral news.
- ဘာပဲဖြစ်ဖြချစ်တစ်ကွာယ်များများ စားကွာအစာအိမ်တော့မဖြစ်ခံနဲ့ ချစ်တုံး as negative news.

5.5 Performance Results with Stop word

In this research, classification of sentiment is also performed with stop word. Most of the results without stop word are better than the results with stop word because unnecessary words or common words are removed. In Table 5.18, testing accuracy for 10-fold cross validation shown for dataset 1. Random-Linear SVC gives greatest value. Table 5.19 is shown accuracy for dataset 2 with 10-fold cross validation in which Unigram feature has highest value. Table 5.20 describes accuracy for 5-fold cross validation with dataset1 which contains greatest value in Unigram feature. Table 21 shown accuracy for dataset 2 with 5-fold cross validation in which combination of Unigram and Bigram is the highest performance value. Table 5.22, Table 5.23, Table 5.24, Table 5.25, Table 5.26, and Table 5.27 are described confusion matrix for both datasets. F1-Score is shown in Table 5.28 for dataset1 that has highest score (68.32) in Random-Linear SVC with Unigram feature and Table 5.29 shown F1-Score for dataset 2 which has greatest score in Random-Linear SVC in combination of Unigram and Bigram.

Table 5.18 Accuracy Score with Stop Words for Dataset 1(10 fold)

TFIDF	**Multinomial NB**	**Linear SVC**	**Random-Linear SVC**
Unigram	63.77	67.45	**68.32**
Bigram	62.06	62.48	**61.65**
Trigram	59.47	58.49	**57.56**
Unigram+Bigram	63.41	66.56	**67.50**
Bigram+Trigram	62.48	62.47	**62.84**
Unigram+Bigram+Trigram	64.44	67.08	**66.67**

Table 5.19 Accuracy Score with Stop Words in Dataset 2 (10 fold)

TFIDF	**Multi NB**	**Linear SVC**	**Random-Linear SVC**
Unigram	63.78	67.45	**68.33**
Bigram	62.06	62.48	**61.657**
Trigram	59.47	58.49	**57.56**
Unigram+Bigram	63.40	66.56	**67. 50**
Bigram+Trigram	62.48	62.47	**62.84**
Unigram+Bigram+Trigram	64.44	67.08	**66.67**

Table 5.20 Accuracy Score with Stop Words in Dataset 1(5 fold)

TFDF	Multinomial NB	Linear SVC	Randomize Linear SVC
Unigram	63.77	66.15	**66.72**
Bigram	61.23	63.20	**62.42**
Trigram	59.52	58.28	**59.06**
Unigram+Bigram	62.78	66.15	**66.20**
Bigram+Trigram	61.75	62.52	**61.96**
Unigram+Bigram+Trigram	63.04	66.20	**66.25**

Table 5.21 Accuracy Score with Stop Word in Dataset 2 (5fold)

TFIDF	Multinomial NB	Linear SVC	Randomize Linear SVC
Unigram	69.61	66.14	**73.63**
Bigram	67.39	63.20	**68.68**
Trigram	61.37	58.28	**62.99**
Unigram+Bigram	72.57	66.15	**74.34**
Bigram+Trigram	67.77	62.52.	**73.95**
Unigram+Bigram+Trigram	72.63	66.20	**68.35**

Table 5.22. Confusion Matrix of Multinomial NB with TFIDF Vectorizer for Dataset 1 with Stop Words

Features	Class	Negative	Neutral	Positive
Unigram	Negative	625	9	76
	Neutral	137	178	142
	Positive	142	29	594
Bigram	Negative	605	24	81
	Neutral	148	205	104
	Positive	170	47	548
Trigram	Negative	545	32	133
	Neutral	153	189	115
	Positive	161	52	552

Unigram + Bigram	Negative	625	28	57
	Neutral	153	212	92
	Positive	152	52	561
Bigram+ Trigram	Negative	591	38	81
	Neutral	160	189	108
	Positive	169	55	541
Unigram +Bigram+ Trigram	Negative	609	36	65
	Neutral	151	205	101
	Positive	154	52	559

Table 5.23. Confusion Matrix of Linear SVC with TFIDF Vectorizer for Dataset 1 with Stop Words

Features	Class	Negative	Neutral	Positive
Unigram	Negative	571	31	108
	Neutral	54	324	79
	Positive	92	92	581
Bigram	Negative	555	47	108
	Neutral	63	297	97
	Positive	118	86	561
Trigram	Negative	480	61	169
	Neutral	64	287	106
	Positive	119	90	556
Unigram + Bigram	Negative	568	42	100
	Neutral	49	320	100
	Positive	104	90	571
Bigram+ Trigram	Negative	551	53	106
	Neutral	68	290	99
	Positive	123	86	556
	Negative	572	48	90

Unigram +Bigram+ Trigram	Neutral	51	323	83
	Positive	106	87	572

Table 5.24. Confusion Matrix of Random Search-LinearSVC with TFIDF Vectorizer for Dataset 1 with Stop Words

Features	Class	Negative	Neutral	Positive
Unigram	Negative	569	39	102
	Neutral	41	358	58
	Positive	90	105	570
Bigram	Negative	552	57	101
	Neutral	50	335	72
	Positive	117	98	550
Trigram	Negative	478	68	164
	Neutral	51	322	84
	Positive	121	98	546
Unigram+Bigram	Negative	559	49	102
	Neutral	22	388	47
	Positive	105	103	557
Bigram+ Trigram	Negative	538	51	121
	Neutral	32	372	53
	Positive	124	100	541
Unigram +Bigram+ Trigram	Negative	565	57	88
	Neutral	43	349	65
	Positive	103	10	559

Table 5.25. Confusion Matrix of Multinomial Naïve Bayes with TFIDF Vectorizer for Dataset 2 with Stop Words

Features	Class	Negative	Neutral	Positive
Unigram	Negative	1192	43	320
	Neutral	378	286	496
	Positive	262	33	3347
Bigram	Negative	1007	103	445
	Neutral	276	228	656
	Positive	238	60	3344
Trigram	Negative	621	79	855
	Neutral	147	128	885
	Positive	159	45	3438
Unigram + Bigram Confusion Matrix of MultiNB:	Negative	1221	82	252
	Neutral	373	348	439
	Positive	321	67	3254
Bigram+ Trigram	Negative	1000	113	442
	Neutral	280	242	638
	Positive	251	76	3315
Unigram +Bigram+ Trigram	Negative	1212	93	250
	Neutral	384	362	414
	Positive	353	67	3222

Table 5.26. Confusion Matrix of Linear SVC with TFIDF Vectorizer for Dataset 2 with Stop Words

Features	Class	Negative	Neutral	Positive
Unigram	Negative	1112	192	251
	Neutral	274	531	355
	Positive	248	149	3245

Bigram	Negative	944	166	445
	Neutral	231	334	595
	Positive	246	124	3272
Trigram	Negative	612	107	836
	Neutral	137	166	857
	Positive	182	66	3394
Unigram + Bigram	Negative	1095	207	253
	Neutral	274	530	356
	Positive	242	150	350
Bigram+ Trigram	Negative	912	612	481
	Neutral	231	327	602
	Positive	234	123	3285
Unigram +Bigram+ Trigram	Negative	1102	182	271
	Neutral	272	531	357
	Positive	247	161	3234

Table 5.27. Confusion Matrix of Random Search-Linear SVC with TFIDF Vectorizer for Dataset 2 with Stop Words

Features	Class	Negative	Neutral	Positive
Unigram	Negative	1165	204	186
	Neutral	274	606	280
	Positive	325	200	3117
Bigram	Negative	996	181	378
	Neutral	234	388	538
	Positive	284	165	3193
Trigram	Negative	654	113	788
	Neutral	143	183	834
	Positive	194	90	3358
Unigram + Bigram	Negative	1150	217	188
	Neutral	291	599	270
	Positive	318	200	3124
Bigram+ Trigram	Negative	970	189	396
	Neutral	249	368	543
	Positive	281	169	3192
Unigram +Bigram+ Trigram	Negative	1153	210	192
	Neutral	295	587	278
	Positive	326	207	3109

Table 5.28 F1-Score for Dataset1 with Stop Words

TFIDF Feature	**Multinomial NB**	**Linear SVC**	**Random-Linear SVC**
Unigram	70.80	76.38	**77.51**

Bigram	69.43	73.07	**74.40**
Trigram	65.71	68.45	**69.69**
Unigram+Bigram	71.52	75.50	**77.83**
Bigram+Trigram	67.41	72.23	**75.07**
Unigram+Bigram+Trigram	70.20	75.91	**76.29**

Table 5.29 F1-Score for Dataset 2 with Stop Words

TFIDF Feature	**Multinomial NB**	**Linear SVC**	**Random-Linear SVC**
Unigram	73.21	76.22	**76. 80**
Bigram	68.76	69.49	**70.56**
Trigram	60.21	60.64	**61.51**
Unigram+Bigram	74.02	76.02	**76.57**
Bigram+Trigram	68.67	68.94	**69.71**

Unigram+Bigram+Trigram	73.82	75.88	**76.18**

5.6 Summary

According to experimental results, different parameter and features are used in the system. Accuracy and F1-score are applied for performance evaluation of research. Unigram feature in TF-IDF with random search has highest F-1 score (77.94) for data 1 and (76.68) for data 2 in word level segmentation. Sentiment analysis without stop words is better performance than sentiment analysis with stop word.

CHAPTER 6
CONCLUSION AND FUTURE DIRECTION

The significance of the research work involves construction of Myanmar news corpus which contain positive tagged news, negative tagged news, neutral tagged news. And then, optimized the model that classified opinion of Myanmar news. Various parameter tuning of support vector machine and TF-IDF are performed to get more performance and effective.

6.1 Dissertation Summary

This dissertation develops automatic sentiment analysis system in Myanmar news. News is very important in all human activities and a key influencer of daily life. Myanmar news sentiment analysis system can be seemed to be useful and applicable in natural language processing research for Myanmar language. N-gram and TFIDF are applied as feature extraction process to get more accuracy. Support vector machine and naïve bayes are used in this system and then support vector machine is optimized by using RandomizesearchCV to get better performance. This system shows performance results of those algorithms with TFIDF feature. In TFIDF vector transformation, LinearSVC with unigram feature has highest has highest F-1 score (77.94) for data 1 and (76.68) for data 2 in word level segmentation. This system is also implemented for sentiment analysis with stop words.

6.2 Advantages and Limitation of the Proposed System

The proposed system provides opinion of Myanmar news to the reader. News can help in society to get more effective, efficient, and safety environment. This research can search sentiment of sentence and document level and then use news corpus. Feature selections (TF-IDF) that use various N-gram range give more accuracy to the system. And then, optimizing system model can help to produce more accurate performance.

However, the proposed research has some limitation because it does not search sentiment of aspect level of news. Word2Vec feature can maintain semantic sense of word and does not waste context information. This proposed work does not use

Word2Vec feature. Deep learning contains own feature and can search feature over data feature and learning is quickly done. This system cannot be modelled by using deep learning.

6.3 Future Directions

The proposed system used TF-IDF method with n-gram range to get feature. Moreover, machine learning, optimizing support vector machine with RandomizesearchCV is performed to classify sentiment. This research work can extend with different optimizing technique, lexicon based, Word2Vec feature, POS tagged feature, and deep learning approach.

6.4 Conclusion

To conclude the dissertation, three contributions are processed in this research. Those contributions meet with objectives presented in Chapter 1. To build News sentiment corpus, senses are manually tagged. Optimizing Linear SVC with Random search can give better performance for this research. Stop words removing in feature almost provide more accurate than performance with stop words.

AUTHOR'S PUBLICATIONS

1. Thein Yu and Khin Thandar Nwet, "Annotation and Sentiment Analysis for Myanmar News", Proceeding of the 16th International Conference on Computer Applications (ICCA 2018), Yangon, Myanmar, Feb 27-28 2019. Pg 160-163
2. Thein Yu and Khin Thandar Nwet, "Automatic Sentiment Analysis for Myanmar News", Proceeding of the 17th International Conference on Computer Applications (ICCA 2019), Yangon, Myanmar, Feb 26-27 2019. Pg 160-164
3. Thein Yu and Khin Thandar Nwet, "Annotation and Sentiment Analysis System for Myanmar News using Naïve Bayes", Myanmar University Research Conference 2019 (MURC), Yangon, Myanmar, May 24-25,2019. Pg 21
4. Thein Yu and Khin Thandar Nwet, "Sentiment Analysis System for Myanmar News Using Logistics Regression Algorithm", 2019 Joint International Conference on Science, Technology and Innovation, Mandalay by IEEE, (ICSTI, Mandalay), Mandalay Myanmar,September 15-16, 2019. Pg 83-89.
5. Thein Yu and Khin Thandar Nwet, " Sentiment Analysis System for Myanmar News using Support Vector Machine and Naïve Bayes", 13^{rd} International Conference on Genetic and Evolutionary Computing 2019, (ICGEC), Springer, AISC 1107, pp. 551–557, 2020.,Qingdao, China, Nov 1-3,2019.
6. Thein Yu and Khin Thandar Nwet, "Myanmar News Sentiment Analyzer Using Support Vector Machine Algorithm", International Journal of Advanced Trends in Computer Science and Engineering, 20^{th}-21^{th} December, 2019. Pg 114-118.
7. Thein Yu, Khin Thandar Nwet, "Sentiment Analysis System for Myanmar News Using Multi-Layer Perceptron (MLP)", ICSE 2019) Yangon, Myanmar, 7^{th} -8^{th} December, 2013.
8. Thein Yu, Khin Thandar Nwet, "Comparing SVM and KNN Algorithms for Myanmar News Sentiment Analysis System", 6^{th} International Conference on Computing and Data Engineering (ICCDE 2020), ACM ICPS volume, ACM ISBN 978-1-4503-7673-0/20/01, Sanya, China, 4^{th}-6^{th} January, 2020.
9. Thein Yu, Khin Thandar Nwet, " Sentiment Analysis System for Myanmar News using K Nearest Neighbor (KNN) and Naïve Bayes", 12 nd International

Conference on Future Computer and Communication (ICFCC 2020), Yangon, Myanmar, 27-28 February, 2020.

BIBLIOGRAPHY

[1] B. Agarwal, V.K.Sharma, and N.Mittal. "“Sentiment Classification of Review Documents using Phrase Patterns”, International Conference on Advances in Computing, Communications and Informatics (ICCAI) 2013,

[2] A. Aizawa,. “An information-theoretic perspective of tf–idf measures”, Information Processing and Management, United of Kingdom: Elsevier; 2003, pp.45–65

[3] T. Al-Moslmi, M. Albared, A. A l-Shabi, N. Omar, and S. Abdullah, “Arabic senti-lexicon: Constructing publicly available language resources for Arabic Sentiment Analysis”, Journal of Information Science, 1–19, 2017, DOI: 10.1177/0165551516683908

[4] D. A. Anggoro, and S. S. Mukti, “Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure”, International Journal of Intelligent Engineering and Systems, Vol.14, No.6,pp.198-208, 2021, DOI: 10.22266/ijies2021.1231.19

[5] H. L. Ardi, E. Sediyono., R. Kusumaningrum, “Support Vector Machine Classifier for Sentiment analysis of Feedback Marketplace with A Comparison Features at Aspect Level”, International Journal of Innovative Research in Advance Engineering, 2017; 11:7-12.

[6] H. M. S. Aung and W. P. Pa, "Analysis of Word Vector Representation Techniques with Machine-Learning Classifiers for Sentiment Analysis of Public Facebook Page’s Comments in Myanmar Text”, 2020 IEEE Conference on Computer Applications(ICCA), 2020, pp. 1-7, doi: 10.1109/ICCA49400.2020.9022842

[7] K .Z. Aung and N.N.Myo “Lexicon Based Sentiment Analysis of Students’ Feedback Comments”, ICCA 2017, pp 116-122Feb , 2017.

[8] T. S. N. Ayutthaya and K. Pasupa, "Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features," 2018 International Joint Symposium on Artificial Intelligence and Natural Language

Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-6, doi: 10.1109/iSAI-NLP.2018.8692836.

[9] A. Bagheri , M. Saraee, and F. d. Jong, “An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews”, NLDB 2013, LNCS 7934, pp. 140–151, 2013. © Springer-Verlag Berlin Heidelberg 2013

[10] M.T.A. Bangsa, S. Priyanta, Y. Suyanto, “Aspect-Based Sentiment Analysis of Online Marketplace Reviews Using Convolutional Neural Network”, JCCS (Indonesian Journal of Computing and Cybernetics Systems)Vol.14, No.2, April 2020,pp. 123~134ISSN(print): 1978-1520, ISSN (online): 2460-7258DOI: 10.22146/ijccs.51646

[11] J. Bergstra, and Y. Bengio, “Random Search for Hyper-Parameter Optimization”, Journal of Machine Learning Research 13 (2012), pp. 281-305

[12] D. Bhalla, “K NEAREST NEIGHBOR : STEP BY STEP TUTORIAL”, https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

[13] J. Brownlee, “How to Encode Text Data for Machine Learning with scikit-learn”,https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

[14] J. Brownlee, “Discover Feature Engineering, How to Engineer Features and How to Get Good at It”, https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

[15] T. Brychcin, M. Konko, and J. Steinberger, “UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis”, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 817–822, Dublin, Ireland, August 23-24, 2014

[16] D. Bzdok, M. Krzywinski and N. Altman, “Machine learning: Supervised methods, Nature Methods”, United Kingdom: Nature Publishing Group; 2018. pp.1-6

[17] A. G, M. Cuadros, and G. Rigau, “V3: Unsupervised Aspect Based Sentiment Analysis for SemEval-2015 Task 1”, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 714–718, Denver, Colorado, June 4-5, 2015. c 2015 Association for Computational Linguistics

[18] D. C. Daniel, L. Shyamala, “An Insight on Sentiment Analysis Research from Text using Deep Learning Methods”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10, August 2019, DOI: 10.35940/ijitee.J9316.0881019

[19] K. Devipriya, D. Prabha, V. Pirya, and S. Sudhakar, “Deep Learning Sentiment Analysis For Recommendations In Social Applications”, IJSTR, Volume 9, Issue 01, January 2020 ISSN 2277-8616 3812 IJSTR©2020

[20] T. Fletcher, “Support Vector Machines Explained”, [Online], Available:https://cling.csd.uwo.ca/cs860/papers/SVM_Explained.pdf, 2008

[21] M. F. Gavilanes , T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, F. J. González-Castaño ,”Unsupervised method for sentiment analysis in online texts”, Expert System with Application 58(2016) 57-75.

[22] M.F. Gavilanes, T. A.López , J. J Martínez , E.C. Montenegro and Francisco, “Unsupervised method for sentiment analysis in online texts”, Expert Systems with Application, Vol. 17 - No. 1 / Enero - Junio de 2019 / 87 - 95

[23] T. Georgieva-Trifonova, and M. Duraku, “Research on N-grams feature selection methods for text classification”, IOP Conf. Series: Materials Science and Engineering 1031 (2021) 012048

[24] N. Godbole, M. Srinivasaiah, and S.Skiena, “Large-Scale Sentiment Analysis for News and Blogs”, Google Inc., New York NY, USA ,Stony Brook University, ICWSM’2007 Boulder, Colorado, USA pp.1-4

[25] D. A. Gopikrishnan., R. Mamidi, “BolLy: Annotation of Sentiment Polarity in Bollywood Lyrics Dataset”, 15th International Conference of the Pacific Association for Computational Linguistics, Yangon, Myanmar, September, 2017.

[26] R. Grosse, “Lecture 5: Multilayer Perceptrons”, [online], Available: https://www.cs.toronto.edu/~rgrosse/courses/csc311_f21/readings/ Multilayer%20Perceptrons.pdf

[27] X. Guan , Y. Li , H. Gong , H. Sun and C. Zhou, “An Improved SVM for Book Review Sentiment Polarity Analysis”, International Conference on Transportation & Logistics, information & Communication, Smart City

(TLICSC 2018), Advances in Intelligent Systems Research, volume 161, pp.139-143, Atlantis press

[28] S. Wu , Y. Liu1, J. Wang and Q. Li, “Sentiment Analysis Method Based on S. Haykin.,” Neural Networks and Learning Machines”, Third Edition, Canada; Pearson; 2018.

[29] M. Heikal, M. Torki, and N. El-Makky, “Sentiment Analysis of Arabic Tweet using Deep Learning”,Procedia, Computer Science, 142 (2018) pp-114–122,

[30] T . Hercig, T . Brychc, L . Svoboda, M. Konko, J. Steinberger, “Unsupervised Methods to Improve Aspect-Based Sentiment Analysis in Czech”, Computación y Sistemas, Vol. 20, No. 3, 2016, pp. 365–375 doi: 10.13053/CyS-20-3-2469

[31] T . Hercig, T . Brychc, L . Svoboda, M. Konko, J. Steinberger, “Unsupervised X. Hu, J. Tang, H. Gao, and H. Liu, “Unsupervised Sentiment Analysis with Emotional Signals”, International World Wide Web Conference Committee (IW3C2). 2013 Rio de Janeiro, Brazil ACM 978-1-4503-2035-1/13/05.

[32] D. Jurafsky ”Speech and Language Processing”, Stanford University, https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

[33] M. Kanakaraj and R.M.R.Guddeti, “NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers”, 3rd International Conference on Signal Processing, Communication and Networking(ICSCN) 2015

[34] H. Kaushik, “Sentiment Analysis using Ensemble Classifier”, International Journal of Innovative Research in Computer and Communication Engineering, 2017; 5:15125-15130.

[35] J. Khairnar, and M. Kinikar, “Machine Learning Algorithms for Opinion Mining and Sentiment Classification”, International Journal of Scientific and Research Publications, 2013; volume 3, pp.1-6.

[36] W. L. K. Khine, N. T. T Aung, “Aspect Level Sentiment Analysis for Different Domain using Deep Learning Approach”, MURC 2019, May 24-25, 2019, pp.39-42

[37] H. Khullar, and A. Singh, “A Proposed Approach for Sentiment Analysis and Sarcasm Detection on Textual Data”, International Journal of Recent

Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019

[38] W. Koehrsen, “A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning”, https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f

[39] Y. Lee, K.Papineni, and S. Roukos, “Language Model Based Arabic Word Segmentation”, ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1July 2003 Pages 399–406https://doi.org/10.3115/1075096.1075147

[40] C. Ma and J. Yang, "Burmese Word Segmentation Method and Implementation Based on CRF," 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 340-343, doi: 10.1109/IALP.2018.8629163.

[41] Z. M. Maung and Y. M. Y. Mikami, “A Rule-based Syllable Segmentation of Myanmar Text”, Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 51–58, Hyderabad, India, January 2008

[42] C. Ma and J. Yang, "Burmese Word Segmentation Method and Implementation Based on CRF," 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 340-343, doi: 10.1109/IALP.2018.8629163.

[43] F. Neumanna, and I. Wegenerb, “Randomized local search, evolutionary algorithms, and the minimum spanning tree problem”, Theoretical Computer Science 378 (2007), pp. 32–40

[44] M. Okada, H. Yanagimoto and K. Hashimoto, "Sentiment Classification with Gated CNN for Customer Reviews", 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2018, pp. 1-5, doi: 10.1109/iSAI-NLP.2018.8692959.

[45, W.P. Pa, “Myanmar Word Segmentation”, http://www.nlpresearch-ucsy.edu.mm/wordsegmentation.html

[46] "W. P. Pa, and N. Thein,” Myanmar Word Segmentation Using a Combined Model”, ecase 2009, University of Computer Studies, Yangon,Myanmar

[47] G. Paltoglou and M. Thelwall, “A study of Information Retrieval weighting schemes for sentiment analysis”, Proceedings of the 48th Annual Meeting of

the Association for Computational Linguistics, pages 1386–1395, Uppsala, Sweden, 11-16 July 2010.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn", https://scikit-learn.org/

[49] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S Manandhar. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

[50] S. Ray, "Learn How to Use Support Vector Machines (SVM) for Data Science", https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[51] K.H. San, "Pyidaungsu 0.1.4", https://pypi.org/project/pyidaungsu/

[52] M. A. Sghaier and M. Zrigui, "Sentiment analysis for Arabic e-commerce websites," 2016 International Conference on Engineering & MIS (ICEMIS), 2016, pp. 1-7, doi: 10.1109/ICEMIS.2016.7745323.

[53] M. Sharma, "Sentiment Analysis: An Introduction to Naive Bayes Algorithm", https://towardsdatascience.com/sentiment-analysis-introduction-to-naive-bayes-algorithm-96831d77ac91

[54] S. S. Shwartz, and S. Ben-David, "Understanding Machine Learning. United State of America:", Cambridge University Press, 2014.

[55] J. Singh, G. Singh, and R. Singh, "Optimization of sentiment analysis using machine learning classifiers", Hum. Centric. Computing and. Information. Science. (2017) pp. 7-32 DOI 10.1186/s13673-017-0116-3

[56] C. Swati, S.Pragati, "Sentiment Analysis of News Articles Using Machine Learning Approach", International Journal of Advances in Electronics and Computer Science, 2015; 2:114-116.

[57] A. Thanda, "What is Logistic Regression? A Beginner's Guide", https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

[58] M. M.Trușcă, "Efficiency of SVM classifier with Word2Vec and Doc2Vec models", Proceedings of the 13th International Conference on Applied Statistics 2019, pp 496-503, ISSN 2668-6309

[59] C. J. Varshney, A. Sharma and D. P. Yadav, "Sentiment Analysis using Ensemble Classification Technique," 2020 IEEE Students Conference on Engineering & Systems (SCES), 2020, pp. 1-6,doi: 10.1109/ SCES50439. 2020.9236754.

[60] S. Wu , Y. Liu1, J. Wang and Q. Li, "Sentiment Analysis Method Based on Kmeans and Online Transfer Learning", Computers, Materials & Continua CMC, vol.60, no.3, pp.1207-1222, 2019

[61] A. Yang, J. Zhang, L. Pan, and Y. Xiang, "Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination", 2015 International Symposium on Security and Privacy in Social Networks and Big Data, pp.52-57

[62] T. Yu, and K. T. Nwet, "Sentiment Analysis System for Myanmar News Using Support Vector Machine and Naïve Bayes", 13rd International Conference on Genetic and Evolutionary Computing 2019, (ICGEC), Springer, AISC 1107, pp. 551–557, 2020.,Qingdao, China, Nov 1-3,2019.

[63] T. Yu and K. T. Nwet,"Myanmar News Sentiment Analyzer Using Support Vector Machine Algorithm", International Journal of Advanced Trends in Computer Science and Engineering, 20th-21th December, 2019. Pg 114-118.

[64] T. Yu, K. T. Nwet, "Comparing SVM and KNN Algorithms for Myanmar News Sentiment Analysis System", 6th International Conference on Computing and Data Engineering (ICCDE 2020), ACM ICPS volume, ACM ISBN 978-1-4503-7673-0/20/01, Sanya, China, 4th-6th January, 2020.

[65]

https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf, "Speech and Language

[66] Processing 2019", Stanford University, September 23, 2018.

"Understanding-support-vector-machine-example-code/Text classification and Naive Bayes", [online] Available: http://nlp.stanford.edu/IR-book/html/html

[67] edition / naive-bayes-text-classification-1.html.2008.

[68] Asian Language Treebank (ALT) project: http://www2.nict.go.jp/astrecatt/member/mutiyama/ALT/
"Logistic Regression", Chapter 321, [Online], Available: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/ df/Procedures/ NCSS/ Logistic_Regression.pdf

LISTS OF ACRONYMS

NLP	Natural Language Processing
TFIDF	Term Frequency Inverse Document Frequency
ML	Machine Learning
SVM	Support Vector Machine
BO	Bayesian Optimization
LRA	Logistic regression analysis
KNN	K Nearest Neighbor
MLP	Multilayer Perceptron
ALT	Asian Language Treebank
NLTK	Natural Language Toolkit
SA	Sentiment Analysis
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

APPENDICES

Appendix: Example of Feature Vector with TFIDF Values

Feature vector and TFIDF value are shown in which first index (corpus inde x) and second index (feature vector index) and values are shown.

1. **Example of bigram feature vector are as follow:**

['ခြောက် နေ' 'စရိုက် မ' 'တောင် ခြောက်' 'ထက် တောင်' 'ထက် ထက်' 'ထက် ထက်မိုး' 'ထက် မနိုင်း' 'ထက် လောက်' 'ထက် အား' 'ထက်မိုး ဦး' 'နေ သလား' 'ပေး နေ' 'ပျောက် ဘူးး' 'မ ထက်' 'မ ပျောက်' 'မ ရှိ' 'မင်းသား မ' 'မနိုင်း မ' 'မနိုင်း အာ့' 'လောက် အရည်အချင်း' 'သလား 😄' 'အညာ စရိုက်' 'အညာ အညာ' 'အရည်အချင်း မ' 'အာ့ အညာ' 'အား ပေး' 'ဦး ကြိုက်တယ်']

TF-IDF Values for bigram

(0, 11)	0.5384979101064753	(3, 12)	0.3513765526966694
(0, 25)	0.5384979101064753	(3, 14)	0.3513765526966694
(0, 8)	0.5384979101064753	(3, 1)	0.3513765526966694
(0, 13)	0.3606383263504801	(3, 21)	0.3513765526966694
(1, 26)	0.5773502691896258	(3, 22)	0.3513765526966694
(1, 9)	0.5773502691896258	(3, 24)	0.3513765526966694
(1, 5)	0.5773502691896258	(3, 18)	0.3513765526966694
(2, 20)	0.38077552389107255	(3, 6)	0.2834883902689878
(2, 10)	0.38077552389107255	(3, 4)	0.23532097247744616
(2, 0)	0.38077552389107255	(4, 15)	0.39079368662348096
(2, 2)	0.38077552389107255	(4, 23)	0.39079368662348096
(2, 3)	0.38077552389107255	(4, 19)	0.39079368662348096
(2, 4)	0.2550098061181917	(4, 7)	0.39079368662348096
(2, 16)	0.38077552389107255	(4, 17)	0.39079368662348096
(2, 13)	0.2550098061181917	(4, 6)	0.3152898857306821
		(4, 4)	0.26171908645728925
		(4, 13)	0.26171908645728925

2. **Example of trigram feature vector are as follow:**

['ခြောက် နေ သလား' 'စရိုက် မ ပျောက်' 'တောင် ခြောက် နေ' 'ထက် တောင် ခြောက်'

'ထက် ထက် တောင်' 'ထက် ထက် မနှိုင်း' 'ထက် ထက်မိုး ဦး' 'ထက် မနှိုင်း မ'

'ထက် မနှိုင်း အာ့' 'ထက် လောက် အရည်အချင်း' 'ထက် အား ပေး'

'ထက်မိုး ဦး ကြိုက်တယ်' 'နေ သလား 😄' 'မ ထက် ထက်' 'မ ထက် လောက်' 'မ ထက်

အား' 'မ ပျောက် ဘူး' 'မင်းသား မ ထက်' 'မနှိုင်း မ ထက်' 'မနှိုင်း အာ့ အညာ'

'လောက် အရည်အချင်း မ' 'အညာ စရိုက် မ' 'အညာ အညာ စရိုက်' 'အရည်အချင်း မ ရှိ'

'အာ့ အညာ အညာ' 'အား ပေး နေ']

Table 4.9 TF-IDF Values for Trigram

(0, 25)	0.5773502691896258	(3, 1)	0.36152911730069653
(0, 10)	0.5773502691896258	(3, 21)	0.36152911730069653
(0, 15)	0.5773502691896258	(3, 22)	0.36152911730069653
(1, 11)	0.7071067811865475	(3, 24)	0.36152911730069653
(1, 6)	0.7071067811865475	(3, 19)	0.36152911730069653
(2, 12)	0.3779644730092272	(3, 8)	0.36152911730069653
(2, 0)	0.3779644730092272	(3, 5)	0.2916794154657719
(2, 2)	0.3779644730092272	(4, 23)	0.38775666010579296
(2, 3)	0.3779644730092272	(4, 20)	0.38775666010579296
(2, 4)	0.3779644730092272	(4, 9)	0.38775666010579296
(2, 13)	0.3779644730092272	(4, 14)	0.38775666010579296
(2, 17)	0.3779644730092272	(4, 18)	0.38775666010579296
(3, 16)	0.36152911730069653	(4, 7)	0.38775666010579296
		(4, 5)	0.3128396318588854

3. **Example of combination of unigram and bigram feature vector are as foll ow:**

['ကြိုက်တယ်' 'ခြောက်' 'ခြောက် နေ' 'စရိုက်' 'စရိုက် မ' 'တောင်' 'တောင် ခြောက်' 'ထက်' 'ထက် တောင်' 'ထက် ထက်' 'ထက် ထက်မိုး' 'ထက် မနှိုင်း' 'ထက် လောက်' 'ထက် အား' 'ထက်မိုး' 'ထက်မိုး ဦး' 'နေ' 'နေ သလား' 'ပေး' 'ပေး နေ' 'ပျောက်' 'ပျောက် ဘူး' 'ဘူး' 'မ' 'မ ထက်' 'မ ပျောက်' 'မ ရှိ' 'မင်းသား' 'မင်းသား မ' 'မနှိုင်း' 'မနှိုင်း မ' 'မနှိုင်း အဲ့' 'ရှိ' 'လောက်' 'လောက် အရည်အချင်း' 'သလား' 'သလား 😄' 'အညာ' 'အညာ စရိုက်' 'အညာ အညာ''အရည်အချင်း' 'အရည်အချင်း မ' 'အဲ့' 'အဲ့ အညာ' 'အား' 'အား ပေး' 'ဦး' 'ဦး ကြိုက်တယ်' '😄']

TF-IDF Values for Unigram + Bigram

(0, 19)	0.38796172348814745	(3, 31)	0.23585971530637553
(0, 45)	0.38796172348814745	(3, 11)	0.19029013321565302
(0, 13)	0.38796172348814745	(3, 22)	0.23585971530637553
(0, 24)	0.2598224877402933	(3, 20)	0.23585971530637553
(0, 16)	0.3130050757045849	(3, 3)	0.23585971530637553
(0, 18)	0.38796172348814745	(3, 37)	0.47171943061275107
(0, 44)	0.38796172348814745	(3, 42)	0.23585971530637553
(0, 7)	0.18486583995673084	(3, 29)	0.19029013321565302
(0, 23)	0.21857086769566408	(3, 9)	0.15795800018011838
(1, 47)	0.4007361920444453	(3, 7)	0.2247768361787918
(1, 15)	0.4007361920444453	(3, 23)	0.1328792494410644
(1, 10)	0.4007361920444453	(4, 26)	0.27204489286950356
(1, 0)	0.4007361920444453	(4, 41)	0.27204489286950356
(1, 46)	0.4007361920444453	(4, 34)	0.27204489286950356
(1, 14)	0.4007361920444453	(4, 12)	0.27204489286950356
(1, 7)	0.19095294266992674	(4, 30)	0.27204489286950356
(2, 36)	0.26944907400220286	(4, 32)	0.27204489286950356
(2, 17)	0.26944907400220286	(4, 40)	0.27204489286950356

(2, 2) 0.26944907400220286	(4, 33) 0.27204489286950356
(2, 6) 0.26944907400220286	(4, 11) 0.2194841066331795
(2, 8) 0.26944907400220286	(4, 29) 0.2194841066331795
(2, 9) 0.18045318516765887	(4, 9) 0.18219163531619745
(2, 28) 0.26944907400220286	(4, 24) 0.18219163531619745
(2, 48) 0.26944907400220286	(4, 7) 0.3888925472396203
(2, 35) 0.26944907400220286	(4, 23) 0.30653069458527626
: :	

4. **Example of combination of bigram and trigram feature vector are as follo w:**

['ခြောက် နေ' 'ခြောက် နေ သလား' 'စရိုက် မ' 'စရိုက် မ ပျောက်' 'တောင် ခြောက်'
'တောင် ခြောက် နေ' 'ထက် တောင်' 'ထက် တောင် ခြောက်' 'ထက် ထက်'
'ထက် ထက် တောင်' 'ထက် ထက် မနိုင်း' 'ထက် ထက်မိုး' 'ထက် ထက်မိုး ဦး'
'ထက် မနိုင်း' 'ထက် မနိုင်း မ' 'ထက် မနိုင်း အာ့' 'ထက် လောက်'
'ထက် လောက် အရည်အချင်း' 'ထက် အား' 'ထက် အား ပေး' 'ထက်မိုး ဦး'
'ထက်မိုး ဦး ကြိုက်တယ်' 'နေ သလား' 'နေ သလား 😀' 'ပေး နေ' 'ပျောက် ဘူး'
'မ ထက်' 'မ ထက် ထက်' 'မ ထက် လောက်' 'မ ထက် အား' 'မ ပျောက်' 'မ ပျောက် ဘူး'
'မ ရှိ' 'မင်းသား မ' 'မင်းသား မ ထက်' 'မနိုင်း မ' 'မနိုင်း မ ထက်'
'မနိုင်း အာ့' 'မနိုင်း အာ့ အညာ' 'လောက် အရည်အချင်း' 'လောက် အရည်အချင်း မ'
'သလား 😀' 'အညာ စရိုက်' 'အညာ စရိုက် မ' 'အညာ အညာ' 'အညာ အညာ စရိုက်'
'အရည်အချင်း မ' 'အရည်အချင်း မ ရှိ' 'အာ့ အညာ' 'အာ့ အညာ အညာ' 'အား ပေး'
'အား ပေး နေ' 'ဦး ကြိုက်တယ်']

TF-IDF Values for Bigram+Trigram

(0, 51)	0.39379498998448487	(3, 10)	0.20329067258779052
(0, 19)	0.39379498998448487	(3, 25)	0.2519735487633456
(0, 29)	0.39379498998448487	(3, 30)	0.2519735487633456
(0, 24)	0.39379498998448487	(3, 2)	0.2519735487633456
(0, 50)	0.39379498998448487	(3, 42)	0.2519735487633456
(0, 18)	0.39379498998448487	(3, 44)	0.2519735487633456
(0, 26)	0.26372909429700125	(3, 48)	0.2519735487633456
(1, 21)	0.4472135954999579	(3, 37)	0.2519735487633456
(1, 12)	0.4472135954999579	(3, 13)	0.20329067258779052
(1, 52)	0.4472135954999579	(3, 8)	0.1687496222457695
(1, 20)	0.4472135954999579	(4, 47)	0.2752528320388974
(1, 11)	0.4472135954999579	(4, 40)	0.2752528320388974
(2, 23)	0.2682495753279348	(4, 17)	0.2752528320388974
(2, 1)	0.2682495753279348	(4, 28)	0.2752528320388974
(2, 5)	0.2682495753279348	(4, 36)	0.2752528320388974
(2, 7)	0.2682495753279348	(4, 14)	0.2752528320388974
(2, 9)	0.2682495753279348	(4, 32)	0.2752528320388974
(2, 27)	0.2682495753279348	(4, 46)	0.2752528320388974
(2, 34)	0.2682495753279348	(4, 39)	0.2752528320388974
(2, 41)	0.2682495753279348	(4, 16)	0.2752528320388974
(2, 22)	0.2682495753279348	(4, 35)	0.2752528320388974
(2, 0)	0.2682495753279348	(4, 10)	0.22207225175621895
(2, 4)	0.2682495753279348	(4, 13)	0.22207225175621895
(2, 6)	0.2682495753279348	(4, 8)	0.18434002956503603
(2, 8)	0.17964986692588128	(4, 26)	0.18434002956503603
:	:		

5. **Example of combination of unigram ,bigram and trigram features are as follow:**

['ကြိုက်တယ်' 'ခြောက်' 'ခြောက် နေ' 'ခြောက် နေ သလား' 'စရိုက်' 'စရိုက် မ'
'စရိုက် မ ပျောက်' 'တောင်' 'တောင် ခြောက်' 'တောင် ခြောက် နေ' 'ထက်'
'ထက် တောင်' 'ထက် တောင် ခြောက်' 'ထက် ထက်' 'ထက် ထက် တောင်'
'ထက် ထက် မနှိုင်း' 'ထက် ထက်မိုး' 'ထက် ထက်မိုး ဦး' 'ထက် မနှိုင်း'
'ထက် မနှိုင်း မ' 'ထက် မနှိုင်း အာ့' 'ထက် လောက်' 'ထက် လောက် အရည်အချင်း'
'ထက် အား' 'ထက် အား ပေး' 'ထက်မိုး' 'ထက်မိုး ဦး' 'ထက်မိုး ဦး ကြိုက်တယ်'
'နေ' 'နေ သလား' 'နေ သလား 😄' 'ပေး' 'ပေး နေ' 'ပျောက်' 'ပျောက် ဘူး' 'ဘူး'
'မ' 'မ ထက်' 'မ ထက် ထက်' 'မ ထက် လောက်' 'မ ထက် အား' 'မ ပျောက်'
'မ ပျောက် ဘူး' 'မ ရှိ' 'မင်းသား' 'မင်းသား မ' 'မင်းသား မ ထက်' 'မနှိုင်း'
'မနှိုင်း မ' 'မနှိုင်း မ ထက်' 'မနှိုင်း အာ့' 'မနှိုင်း အာ့ အညာ' 'ရှိ'
'လောက်' 'လောက် အရည်အချင်း' 'လောက် အရည်အချင်း မ' 'သလား' 'သလား 😄'
'အညာ' 'အညာ စရိုက်' 'အညာ စရိုက် မ' 'အညာ အညာ' 'အညာ အညာ စရိုက်'
'အရည်အချင်း' 'အရည်အချင်း မ' 'အရည်အချင်း မ ရှိ' 'အာ့' 'အာ့ အညာ' 'အာ့ အညာ အညာ'
'အား' 'အား ပေး' 'အား ပေး နေ' 'ဦး' 'ဦး ကြိုက်တယ်' '😄']

TF-IDF Values for Unigram+Bigram+Trigram

(0, 71) 0.32201339860697215	(3, 47) 0.1593729226606077
(0, 24) 0.32201339860697215	(3, 13) 0.1322939225535182
(0, 40) 0.32201339860697215	(3, 10) 0.18825643097122963
(0, 32) 0.32201339860697215	(3, 36) 0.11128981827118893
(0, 70) 0.32201339860697215	(4, 65) 0.22270173156025774
(0, 23) 0.32201339860697215	(4, 55) 0.22270173156025774
(0, 37) 0.21565612596915448	(4, 22) 0.22270173156025774
(0, 28) 0.2597983824348725	(4, 39) 0.22270173156025774
(0, 31) 0.32201339860697215	(4, 49) 0.22270173156025774
(0, 69) 0.32201339860697215	(4, 19) 0.22270173156025774
(0, 10) 0.15344110979705491	(4, 43) 0.22270173156025774
(0, 36) 0.18141673181144607	(4, 64) 0.22270173156025774
(1, 27) 0.34864042110528976	(4, 54) 0.22270173156025774
(1, 17) 0.34864042110528976	(4, 21) 0.22270173156025774
(1, 73) 0.34864042110528976	(4, 48) 0.22270173156025774
(1, 26) 0.34864042110528976	(4, 52) 0.22270173156025774
(1, 16) 0.34864042110528976	(4, 63) 0.22270173156025774
(1, 0) 0.34864042110528976	(4, 53) 0.22270173156025774
(1, 72) 0.34864042110528976	(4, 15) 0.1796743547786879
(1, 25) 0.34864042110528976	(4, 18) 0.1796743547786879
(1, 10) 0.1661290286861683	(4, 47) 0.1796743547786879
(2, 30) 0.21940392963625183	(4, 13) 0.1491459451149307
(2, 3) 0.21940392963625183	(4, 37) 0.1491459451149307
(2, 9) 0.21940392963625183	(4, 10) 0.3183557050000827
(2, 12) 0.21940392963625183	(4, 36) 0.25093254183328983
: :	