Bachelor of Science in Computer Science & Engineering

Classification of Bangla Sports News Using Machine Learning Techniques

by

Adrita Barua

ID: 1504015

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

Classification of Bangla Sports News Using Machine Learning Techniques

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Adrita Barua
ID: 1504015

Supervised by

Dr. Mohammed Moshiul Hoque
Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.
April, 2021

The thesis titled “**Classification of Bangla Sports News Using Machine Learning Techniques”** submitted by ID: 1504015, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Chairman

Dr. Mohammed Moshiul Hoque
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (Ex-Officio)

Dr. Asaduzzaman
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (External)

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Adrita Barua

Signature of the candidate

Date: 19-4-2021

Acknowledgements

The progress and result of this thesis involved a great deal of support and assistance from several people without whom it would not have been possible. I consider myself extremely fortunate to have received this during my thesis. First and foremost, I want to thank my thesis supervisor, Dr. Mohammed Moshiul Hoque, Professor, Department of Computer Science Engineering, Chittagong University of Engineering Technology (CUET), for his constant guidance, encouragement, and unwavering support during the preparation of this thesis work. I am thankful for his many crucial questions, his exalted support throughout the entire time, and motivating me to see things from diverse perspectives on Bengali language processing domains.

I owe my gratitude to Omar Sharif, Lecturer, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET), to take an interest in my work and guide me in providing the requisite expertise whenever needed.

I am incredibly thankful to Prof. Dr. Asaduzzaman, Head, Department of Computer Science Engineering, Chittagong University of Engineering Technology (CUET), for his outstanding support throughout my undergraduate education. I would also like to express my gratitude to all my teachers, data crawlers, annotators and supporting staffs of CSE Department for their helpful cooperation, and assistance, which contributed to the thesis's successful completion.

Finally, I want to thank my father and mother for their unconditional love, support, encouragement, and contribution throughout my life and academic career in every aspect over the years.

This research was conducted under the support of CUET NLP Lab, whose support is greatly appreciated.

Abstract

With the growing advancement of technology, categorization of the electronic text documents in an organized manner has become a very crucial task. Text categorization has a vast application in Natural Language Processing (NLP), starting from search engines to text mining. Hence, developing an efficient classifier model requires a thorough investigation of different feature spaces extracted from correctly labeled annotated corpus which has not yet been studied in the realm of Bengali language processing. This thesis explores the machine learning (ML) based approach to classify Bengali sports textual news into one of the four categories: cricket, football, athletics, and tennis. Due to the unavailability of benchmark dataset, this work develops a Bengali news corpus consisting a total of 43,306 documents 2,02,830 unique words to perform the sports news classification. This work also investigates the various features using TF-IDF technique for three different n-grams: unigram, unigram-bigram, unigram-bigram-trigram. To investigate the performance of Bengali sports news classification task, six ML models (such as LR, NB, SVC, RF, DT and KNN) are evaluated on the developed corpus with various feature combinations. The comparative analysis among classifiers show that SVC method with unigram+bigram+trigram features outperforms the other ML techniques with acquiring a highest Weighted F1-score of 97.60% on test dataset.

Keywords— **Natural language processing, Text categorization, Sports news classification, Bengali language processing, Machine learning, Feature extraction and Evaluation.**

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DT Decision Tree. 15, 26

KNN K-Nearest Neighbour. 15, 26

LR Logistic Regression. 26

ML Machine Learning. 15

NB Naive Bayes. 26

RF Random Forest. 26

SVC Support Vector Classifier. 26

TF-IDF Term Frequency-Inverse Document Frequency. 16, 23

Chapter 1

Introduction

1.1 Introduction

In the present century, the brisk enhancement of the World Wide Web has generated an explosive amount of electronic data, especially news documents. Like any other language, several Bengali online news portals publish a vast amount of news articles according to their different layout and categorization process. The variability of different formats and categorization methods makes it difficult for the users to find their preferable news articles for a specific category without separately scanning through individual news portals. Consequently, to make the news documents accessible to the users, it is inevitable to organize and label them into appropriate categories. This situation emphasizes the importance of text classification to automatically analyze and assign text documents into their respective categories based on their contents. Moreover, the hierarchical sub-categorization is now more important than ever to find a particular news article from the vast pile of scattered news documents. This theses proposed an ML-based sports news classification model that can classify sports news into four classes: sub-categories: cricket, football, tennis, and athletics. This Chapter provides an overview of general framework of news classification. Additionally, Chapter 1 discusses the motivation, applications, and contributions of the work. Finally, the organization of the thesis is presented at the end of this Chapter.

1.2 General Framework of News Classification

Text classification is a highly popular and active research area within the field of Natural Language Processing. The classifier models are generated using two main approaches: Supervised and Unsupervised. In Supervised method the classifier models are trained using annotated corpus, but in Unsupervised method the corpus is not annotated. In this work, supervised classifier models are used for the sub-categorization problem as it is the most widely used method for text classification. Again in the classification process the classifier models are required to be trained with effective feature space generated using different feature values from the training dataset.

For Sports news sub-categorization , the TF-IDF values of combined n-grams are used as the feature extraction technique which is described in the proposed framework. The framework comprises the following main steps –

- **Pre-processing** The raw text documents are pre-processed to remove unwanted noise/symbols and stop wards to reduce the size of irrelevant features.
- **Feature Extraction** During the feature extraction process, the textual documents are represented numerically using a combination of three N-grams: Unigram-Bigram-Trigram. The TF-IDF values of each N-gram is used for the feature vector generation.
- **Classifier Model Generation** Finally, the classifier models are generated using the supervised approach, where the models are trained on the training dataset's feature space. The learned model is deployed to predict unknown test documents and gives the predicted label of a Sports document as the output of the model.

The Block diagram of Bangla Sports News Sub-classification framework is shown below:

1.3 Difficulties

Despite the wide use and popularity of the Bangla language, there are only a few recent advances in Bengali document sub-classification compared with other popular languages such as English. Though very few research activities have been

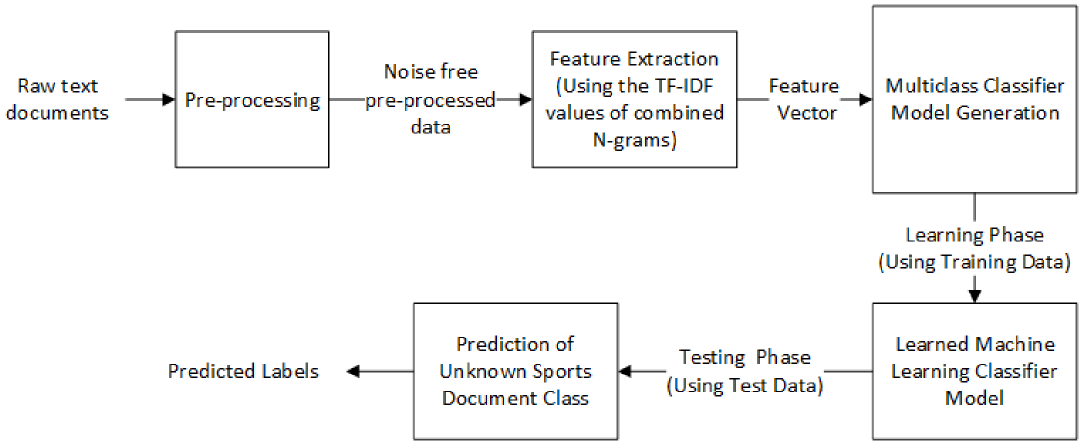

Raw text documents
Pre-processing
Noise free pre-processed data
Feature Extraction (Using the TF-IDF values of combined N-grams)
Feature Vector
Multiclass Classifier Model Generation
Learning Phase (Using Training Data)
Learned Machine Learning Classifier Model
Testing Phase (Using Test Data)
Prediction of Unknown Sports Document Class
Predicted Labels

Figure 1.1: Block Diagram of Sports News Sub-classification Framework

conducted till to date in Bengali news articles classification (e.g., sports, entertainment) [1] investigation the categorization process of main news categories. However, none of them have given satisfactory performance due to lack of a standard corpus and poor feature selection. In the proposed work, a new corpus is developed to perform the sports news sub-categorization task. Due to the unavailability of any well structured corpus or any standard feature extracting tools we had to go through some theoretical, scientific and practical difficulties. The major challenges observed during to conduct this work are stated in the following:

1. **Unavailability of properly annotated Bangla corpus:** A well-annotated corpus is required to generate the supervised machine learning classifier models. However, unlike other languages, there is no standard Bangla corpus available, which contains a diverse range of Bangla text documents. As a result, developing a properly labeled Bangla corpus is a challenging task.

2. **Co-related features in a Multi-class Classification problem :** Feature extraction is a fundamental step in many data classification tasks because it helps define which features in the dataset are relevant and not. Labels in a single classification are assumed to be mutually exclusive, which is not entirely accurate in multiclass classification. As a result, there is an urgent need to use appropriate feature extraction methods specifically designed to handle multi-label data. It would be even better if these feature extraction methods take into account label correlations.

3. **Unpredictable Nature of the News Documents:** The news is frequently about individuals, places, and organizations. In contrast to other types of textual data, these entities often enter the news vocabulary suddenly; for example, ”Jahanara Alam” is recently featured in the Bangladesh Women Squad Women’s Twenty20 Asia Cup she gained tremendous popularity. While it is often sufficient to say ”Jahanara Alam” for the human brain to immediately register the news as ”Cricket”, classifiers analyzing months or years of historical data are proving less effective at making that determination, as the term is highly relevant, but it is in a very small sample size.

4. **Generation of narrow Feature Space:** The majority of significant classification algorithms uses word frequency (or term frequency) for their computational analyses. Though the total number of unique words/terms is high in a news corpus but the distribution of the terms is highly imbalanced. The vast majority of content is composed of a relatively small vocabulary of common words. The large variety of words is substantially made up of proper nouns and other types of entities, given that news content is frequently written for a diverse audience with varying language skills and reading levels. As a result, it becomes difficult to extract a diverse set of critical features necessary for creating an efficient classifier model.

1.4 Applications

News document classification has a wide area of application in this present era of information technology. It facilitates the analysis and organization of the news documents in a convenient way. The real world application of Sports news subcategorization is stated here :

- A good classifier model can be used effectively in organizing the vast pile of unstructured data and can have a wide range of use in data mining and other data analysis process.

- Sports news holds a wide range of readers, and it endorses many business advertisements and sponsors. A proper classification method will help the different organizations to carry out practical data analysis contributing to their particular business interest.

- As most news portals publish news of different categories based on their individual taxonomy, it is difficult for the readers to find out certain news. A generalized classification method will help the readers to find out information according to their desired interest.
- A good classifier dedicated for the Sports news documents will help any writer, blogger or publication organization to publish the biographies or documentations on a particular Sports person or any specific sports category.
- It will help to generate and optimize search engines, and any data analysis process will be beneficial using this classification model.

1.5 Motivation

Automation is required to compete with other countries' technological advancements. Additionally, in the last decade, Bangladesh has seen a huge increase in the use of electronic media to generate digital data. There are hundreds of news portals that generate a large volume of news documents daily. As Sports news attracts a sizable portion of the audience, developing an efficient Bangla Classifier model that can handle news documents will enable us to retrieve any news information automatically. It will aid in the optimization of search engine results and any data analysis model. Numerous organizations and commercial enterprises will benefit from our categorization model in the course of their data mining processes. All of these factors contributed to the development of an efficient categorization model for Sports news documents.

1.6 Contribution of the thesis

This thesis focuses on utilizing the diverse feature space to develop a good functional classifier model using a dedicated corpus of sports news documents. The following are the primary contributions of this thesis:

1. Develop a annotated corpus containing a total of 43,306 sports news documents with 2,02,830 unique words labelled into four distinct categories: Cricket, Football, Tennis and Athletics.
2. Investigate the various features suitable for Bengali sports news classification by exploring n-gram.
3. Develop ML-based framework using Tf-Idf feature extraction and SVC classifier to classify textual sports news with various parameters selection.
4. Investigate and compare the performance of the proposed model with other ML baselines and existing techniques.

1.7 Thesis Organization

The remaining of the thesis is organized as follow:

- Chapter 2 summarizes previous research in the field of text classification using various Machine Learning approaches, including their contributions and limitations.
- Chapter 3 details the proposed methodology for categorizing Sports news documents into their appropriate sub-categories. In the proposed framework, a combination of three n-grams are used to generate the supervised classifier models.
- Chapter 4 details the process of creating a working dataset and analyzing performance metrics for the proposed framework.
- Chapter 5 summarizes the thesis work, its limitations, and future recommendations.

1.8 Conclusion

This chapter provides an overview of our research. A summary of the proposed framework has been included, as well as the motivation for this thesis and its applications. Finally, we discuss our research contributions and the difficulties we encountered. The following chapter will provide background information, literature reviews, and an assessment of the current state of the problem.

Chapter 2

Literature Review

2.1 Introduction

The Sub-categorization process of the text documents requires a detailed analysis of different feature space using various classification models. We must provide a thorough understanding of previous research works to understand the text classification system fully. Multiple feature extraction techniques and classification methods used by different researchers will be discussed in this chapter, along with their performance on different aspects of the system. Chapter 2 explains a few basic terminology and terms related to sports news classification. Moreover, this Chapter discuss the various textual news classification techniques that is closely related to the proposed work. Few implementation challenges also highlights at the end of this Chapter.

2.2 Important Terms and Terminology

There are several critical terms and terminology related to the text-based emotion classification task which is described in the following:

- **Text Classification:** The process of assigning a document(D) to a particular class label(C_i) among other available class labels in the given set of classes(C) is known as Text Classification [2]. It can be done using two approaches: Statistical and Machine Learning. Here the ML approaches can be further divided into Supervised, Unsupervised and Semi-supervised methods.

- **News Classification:** News classification is the process where news documents are classified into the main news categories(e.g Sports, Entertainment etc.) to ensure the easier navigation of the news articles [3]. The process of News classification is mainly performed in two ways : Content based and Headline based classification.

- **News Sub-classification:** The process where the classification model can identify not only the topic(Sports), but also the category(Cricket) of a news article is known as News Sub-classification.
- **News Corpus**: Text classification process is solely dependent on the nature of the corpus used. So, the understanding of the corpus nature is a vital issue for the classification process. The corpus that provides a collection of different news documents labeled with their respective genres, outlets and other information is known as a news corpus [4].

2.3 Feature Extraction Methods for News Classification

Extracting important features is a key step in any classification task. Several contextual and non-contextual techniques have been used to extract relevant syntactic/semantic features from text expression. Some of the most popular non-contextual feature extraction techniques are described in the following:

2.3.1 N-gram

N-gram is used as a feature to represent n-items that occurs in-order in the given text sequence. The most common representation of textual documents is to convert it into a vector space where each term is indexed with a numeric value(TF) by which it has occurred in the document. Here the terms can be consists of words or sequence of words. If each word(1-gram) in the document is considered as a term in the vocabulary it is named as unigram. Unigram approach does not represent the syntactic relation between the words as the words loses their order. But if the whole document is sliced into consecutive sets of n-words (n=2 for bigram, n=3 for trigram and so on) it will provide better information for the classifier model. A combination of two n-grams can also be considered.

In a combination of unigram and bigram, both the single words and the sets of two words will be considered separate terms in the vector space of the document.

2.3.2 TF-IDF

During the TF-IDF calculation, Term frequency(TF) gives higher values to the most frequent terms occurring in the document. However, some words are generally common to all the document sets having no significance in document categorization. In this case, the high TF value of a term is adjusted using its Inverse Document Frequency(IDF). The *idf* value of each term is represented using the following equation,

$$idf = log_e(\frac{D+1}{df_i+1}) + 1 \tag{2.1}$$

Where, D = Total no. of documents in the corpus, and df_i = No. of documents containing the term i.

Finally, the $\tau f - idf$ value is calculated and normalized using the following formula,

$$\tau f - idf = \frac{(\tau f_i * idf_i)}{\sqrt{\sum_{i=1}^{n}(\tau f_i * idf_i)^2}} \tag{2.2}$$

Where, τf_i is the number of times the term i has occurred in the document and idf_i is the inverse document frequency of that term.

2.4 ML-based Methods for News Classification

Nowadays, various machine learning (ML) techniques have been employed for the classification task. These techniques also most commonly used in analysis non-textual news classification in many languages. Some popular ML techniques are explained in the following subsections.

2.4.1 SVC

SVC is a powerful supervised classifier model that is used widely used to solve classification problems. It generates a hyperplane that separates the data points of different classes with maximum margin, which means having the maximum distance from the closest data points on both sides of the hyperplane. These closest data points are known as support vectors used as the deciding factors to find the "predicted" class of new test documents. The main goal of a SVC model is to generate the most optimal hyperplane iteratively using labeled training data. For a given set of N labeled training examples $(x, y)_i$, where the input samples $x_i \in R^n$ belonging to two different classes $y \in \{-1, 1\}$, a hyperplane is constructed using the following formula,

$$y = \mathbf{w}\mathbf{x} + b \tag{2.3}$$

with normal vector $\mathbf{w} \in \mathbf{R}^n$ and bias b. The margin of the hyperplane (2.3) is maximized by solving the following constrained optimization problem:

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1 \tag{2.4}$$

This optimization problem is solved by employing the Lagrange theory, leading to the maximum-margin hyperplane normal vector,

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \tag{2.5}$$

with α_i being the Lagrange coefficients. This hyperplane is generated to separate only two classes, for any multi-class classification the problem space is divided into k binary classifiers. It is implemented in practice using a kernel that transforms the inseparable data space into the separable data points by converting it into a higher dimensional space.

2.4.2 Logistic Regression

Logistic Regression(LR) is one of the probabilistic methods used in the filed of Machine Learning to perform classification problems. Here, a logistic function is used to transform the real valued output of a linear model into a value between 0 and 1 to indicate to probability of a default class. If the probability value is greater than .5 then the given document belongs to the default class otherwise it does not. In text classification, the logistic function's output p indicates the probability of a document D belonging to a default class C. For a input variable X we can write a simple linear function,

$$y = b0 + b1 * X \tag{2.6}$$

Here,the output y is transformed into the probability value p using the following logistic function,

$$p = \frac{1}{1 + e^{-y}} \tag{2.7}$$

We can rewrite the equation (2.7) as,

$$ln(\frac{p}{1-p}) = b0 + b1 * X \tag{2.8}$$

The coefficients (Beta values b) of the logistic regression algorithm is estimated from the training data using maximum-likelihood estimation.The mechanism of maximum-likelihood for logistic regression is to find the values of the coefficients(Beta values) that minimize the error in the probabilities predicted by the model. For our text classification model, the LR model predicts the class of a document by calculating the coefficients $\{b1, b2, ...bn\}$ for given set of input variables $\{X1, X2...Xn\}$ using the training data-set. This LR model normally predicts the binary categories, but for multi-class classification problem multinomial LR is used where the whole classification problem is divided into separate binary classifiers.

2.4.3 Naive Bayes

Naive Bayes(NB) is a classifier model based on a simple probabilistic method. It calculates the probability of a document d belonging to a class c_i using the Bayes' theorem,

$$\rho(c_i/d) = \frac{\rho(d/c_i)\rho(c_i)}{\rho(d)} \tag{2.9}$$

Here, $\rho(c_i/d)$ is known as the posterior probability. The class having the highest posterior probability is assigned to the document d as follows,

$$C_{max} = \operatorname*{argmax}_{c_i \in C} \rho(d/c_i)\rho(c_i) \tag{2.10}$$

Here, as the document probability $\rho(d)$ is common to all the classes it can be ignored. The probability of a class $\rho(c_i)$ can be calculated from the number of documents in that category divided by documents number in all categories. $\rho(d/c_i)$ represents the probability of document d for a given class c_i which can be calculated by representing the document as a set of independent features $\{X_1, X_2, X_3, ...X_n\}$. So the final equation is given by ,

$$C_{max} = \underset{c_i \in C}{\operatorname{argmax}} \, \rho(X_1/c_i) \times \rho(X_2/c_i) \times \rho(X_3/c_i)...\rho(X_n/c_i)\rho(c_i) \tag{2.11}$$

2.4.4 Decision Tree

Decision Tree is a tree-based classifier model that uses a tree-like structure where the nodes represent the features. The edge represents a decision rule to choose the feature value and the leafs correspond to categories.The topmost root node learns to partition based on the feature value. It partitions the tree using recursive partitioning. The best feature value is selected using Attribute Selection Measures (ASM) to split the records. The most common ASM is Information gain. It calculates how much information a feature provides about a class. The information gain of a attribute A can be calculated as,

$$Gain(A) = Entropy(D) - [(WeightedAvg) * \sum_{a_i \in A} Entropy(a_i)] \tag{2.12}$$

Where, Entropy is a metric to measure the impurity in a given attribute. If, D is total number of documents in the training set, the attribute A with the highest information gain, $Gain(A)$, is chosen as the splitting attribute.

It recursively makes new decision trees by splitting the nodes based on attribute values until it finally reaches a leaf node with a chosen category, thus assigning the documents into their respective categories.

2.4.5 Random Forest

Random Forest is one of the most popular test classification methods based on ensemble learning using decision trees as the estimators. It creates separate decision trees on randomly selected data points from the training data, gets prediction from each tree and selects the best solution with the highest number of votes among all the trees. It can generate a model with a very fast execution time, which can handle a large amount of data points as all the trees are independent of each other. It provides better performance as the majority vote's prediction value among several estimators can end up being better than the performance of any individual estimator.

2.4.6 KNN

KNN uses a very basic approach to classify documents. For a given test document d, the model finds the k nearest neighbours among the training documents based on their similarity score using Euclidean distance. This score is used to weight the class of the neighbouring documents for the test document d. When multiple k nearest documents belong to the same class, each neighbour's similarity scores are added to weight that class. Finally, the test document is assigned to the candidate class having the highest score. The decision rule to find the class of a test document d is given by,

$$s(d, c_i) = \sum_{d_j \in KNN} Sim(d, d_j)\gamma(d_j, c_i) \tag{2.13}$$

Here, $s(d, c_i)$ is the score of a candidate class c_i with respect to the test document d where $Sim(d, d_j)$ is the cosine similarity score of a document d_j that belongs to the set of K-nearest neighbours KNN. $\gamma(d_j, c_i)$ gives 1 or 0 values depending on whether the document d_j belongs to class c_i or not.

Finally the class c_i having the highest score is assigned to the document d. In practice the number of k is needed to be defined by the programmer to generate the KNN model. So it is difficult to find the optimum number of K-neighbours for the given corpus.

2.5 Related Literature Review

Although the textual news articles classification has achieved an enormous progress concerning the highly resources languages (i.e., English, Chinise, Arabic, and Spanish), there is a very few research activities have been conducted till to date in Bengali language. Based on the survey of previous literature, we presents the reviews on sports news classification concerning non-Bengali language-based and Bengali language based news classification.

2.5.1 Non-Bengali Language based News Classification

In the field of NLP, text classification holds an important and large area of implementation [5]. A huge amount of works have been done on text classification in various languages that use different ML techniques [6] following some basic outlines [7]. However, different studies are still going on to evaluate more efficient approaches. The performance of text classifier models largely depends on the dataset that has been used to train those models. To generate an effective news document classifier, [8] Zakzouk et al. generated three separate supervised binary classifiers-, C4.5(DT) and Naive Bayes by training them on a small corpus using only 332 English documents of different sports news articles. But here, the models could only perform binary classification to label - Cricket and non-Cricket documents, as there were not enough documents of other sports categories to train the models. A comparative analysis of different ML models on BBC news documents is addressed by Shah et al. [9] where different performance metrics are used to evaluate the model's performance. To maximize the effectiveness of Test classifier models, Bidi et al. [10] used a generic approach for optimum feature selection and evaluated its performance using three ML classifiers-SVM, Naive Bayes and KNN on two separate corpus sets. Other than English, some studies have been done for Arabic document classification [11] using a dataset collected from the Aljazeera news portal where the Maximum Entropy method was used. Another work [12] investigated supervised ML classifiers for Urdu news documents. Suleymanov et al. [13] used Azerbaijani news articles to develop three classifier models-SVM, NB and Artificial Neural Networks, where stemming and other feature reduction methods were used to improve classifier performance. Here an optimum number of categories were generated using K-means clustering. Other than these supervised Machine Learning Approaches, some semi-supervised [14] and Deep Leaning methods [15] are also being used to facilitate the text classification process.

2.5.2 Bengali Language based News Classification

For the recent few years, some of the works are being done for Bangla text classification, mainly based on supervised Machine Learning approaches. Islam et al. [16] generated an SVM based Bangla document classifier using TF-IDF feature extraction method for twelve main news categories and showed unigram feature works best for their system. Al Mostakim et al. [17] developed a corpus of 5870 Bangla news documents and used six supervised machine learning classifiers to classify multi-class documents. They concluded that the Logistic Regression model performed better. However, no explanation was provided in this work regarding the validation of the evaluation metric used for a very small imbalanced corpus size. Another study on Bangla Document Classification has been done by Mandal et al. [18] using a corpus of 1000 news documents where four supervised learning methods - DT, KNN, NB and SVM were used to classify five document classes. They showed that SVM works better on large and noisy text documents. Here, the work lacks in performance with a low f1-score of around 89%. Chy et al. [19] proposed a Naive Bayes based document classifier where stemming, stop words removal, and several feature extraction methods were used to improve the classifier performance. Other than this Quadery et al. [20] proposed a keywords based Bengali document classifier where they simply stated the relation of the classifier performance based on the frequency of the keywords in each category. Dhar et al. [21] proposed an ML-based news classification method using the headlines of the news articles, which performed poorly due to the lacking of adequate feature space. Hasan et al. [22] proposed a Bi-LSTM based method to classify Bengali news based on news headlines which achieved an accuracy of 85.14 % for eight classes.

In table 2.1 a summary of the previous works in Bangla that worked with similar

Table 2.1: Summary of Previous Works in Bengali News Classification

Method	**News Class**	**Dataset**	**Limitation**
Unigram+SVM[16]	12	31908	They've worked only with main news classes and didn't mention the parameters used during the classification model generation.
Chi-square+LR [17]	5	5870	Used a very small dataset without mentioning any exact evaluation metric.
Tf-Idf + SVM[18]	5	1000	Used a very small imbalanced dataset with only 89.24 % f1 score.
Tf-Idf+ NB [21]	3	474	Only used the headlines with a poor performance of 74% f1 score.

classification problem is discussed. The limitations of these models are also mentioned so that we can address them in our proposed work. In our work we tried to built a larger dataset with proper description and validation as most of the works lack in performance due to the absence of an adequate dataset. A diverse dataset is generated to train the models for better performance. As most of the time imbalanced dataset is used we tried to provide an standard evaluation metrics using the Weighted f1-score for the performance assessment of the models.

2.5.3 Implementation Challenges

Due to the complex nature of Bangla text processing, we had to go through several challenges during the implementation of the work.The main challenges encountered during the implementation of the proposed framework is given below:

- A well-annotated corpus is a prerequisite for any supervised ML model. The creation of a dataset with proper annotation and enough document classes is quite challenging. As there is no standard source of the Bangla dataset, the data was collected and annotated manually, which is time-consuming and tedious.
- Again to process this huge amount of data, it requires high-speed hardware requirements to run the models.
- Furthermore, there is no standard library for pre-processing Bangla text

documents. So, all the processing was done using raw coding, which also needed particular expertise.

2.6 Conclusion

In this chapter, a detailed literature review of related works has been discussed.The previous works of both Bangla and other languages are motioned in separate section. From the above section it was evident that, there is still a wide area of research that has yet to be validated as most previous Bengali works have suffered from a lack of performance due to the unavailability of an adequately annotated corpus. Furthermore, in most previously mentioned work, several datasets are used without clearly describing or validating the creation and development process. Again, all previous work has been done for the purpose of classifying major news classes. None of them has worked with the sub-categorization method of a specific news class into fine-grained sub-categories. Since the sub-categories have some common features, it's difficult to create a model that can solve this problem with respect to a diverse feature space. Addressing all these aspects, our proposed methodology for Bangla Sports News Sub-classification is discussed in details in the next chapter.

Chapter 3

Methodology

3.1 Introduction

Categorizing Bangla sports news into their respective subcategories requires a detailed analysis of various feature spaces. The majority of sports news contains a few common keywords, resulting in similar feature spaces. As a result, it is difficult to distinguish between different types of sports data. This chapter presents the proposed methodology for Sports news classification in Bengali and explains in details with its constituents. The corpus development procedure and its statistics also presented in this Chapter. The details description of the training model preparation, feature extraction and selecting optimum hyperparameters also included in Chapter 3.

3.2 Corpus Development

Due to the lack of a large and well-structured Bangla corpus, Chittagong University of Engineering & Technology has taken the initiative to develop a corpus(called BeNC) containing diverse Bengali document sets under the auspices of the CUET NLP lab.

3.2.1 Data Accumulation

In this work we used a subset of the BeNC corpus containing only sports news documents. To collect data automatically, distinct Python scripts were created based on the layouts of various news portals. A total of 43,322 sports news documents were collected along with their metadata.The entire data collection process took place over a year, from 2019 to 2020, with data publication dates ranging from 2015 to 2020.

3.2.2 Data Pre-processing

After creating the corpus it is a significant step to pre-process the data-set. Data pre-processing facilitates noise and dimension reduction as it removes the unimportant words and symbols from the feature space. It will help the classifier models to train over the relevant feature sets and improve its performance[23]. The pre-processing steps taken in this work is described here.

3.2.2.1 Tokenization

Tokenization is the first step of data pre-processing which identifies each terms or words in the document-set separated by white spaces or new lines and represents them into tokens for further processing.

3.2.2.2 Noise Removal

Any special symbols or digits ([,০-৯,0-9,a-z,@,!,#,$,%,&,.„ ,|,:,],+ etc.) are removed from the data set as they do no add any significance in the text categorization process.

3.2.2.3 Stop Words Removal

Some words that appear most frequently in a data-set(i.e. pronouns and conjunctions) but do not have any relevance with the document category are considered as stop words. Removing stop words would help the classifier models to extract important features as most of the feature extraction methods gives priority to the word frequency [24]. In this step these words are removed from the data set using a standard list of Bangla stop-words.

An example of the documents after pre-processing is given below.

Sample Documents:

D_1 = দুই ম্যাচেই টস জিতে ব্যাটিং করেছে বাংলাদেশ।

D_2 = ব্যাটিং করেছে একই ছন্দে এবং গতিতে।

Pre-processed Documents:

D_1 = ম্যাচেই টস জিতে ব্যাটিং বাংলাদেশ

D_2 = ব্যাটিং ছন্দে গতিতে

3.2.3 Data Annotation and Quality assessment

Throughout the process, it was discovered that individual news portals had their own distinct sub-categories within Sports news. The entire corpus was annotated with four common sub-categories: Cricket, Football, Tennis, and Athletics, all of which were covered by the majority of news portals. This annotation process was carried out automatically where the sub-categories are already defined by the news portals. But a portion of it(around 20%) is done manually by five postgraduate students working on BLP, where no sub-categories were defined by the news portals. To choose the initial label majority voting technique is applied [25]. Initial labels are scrutinized by a BLP researcher with many years of experience. If some of the original annotations were incorrect, the expert corrected the labeling. 16 document files were discarded as they didn't belong to any of the predefined classes.

3.2.4 Data Statistics

The corpus contained a total of 43,306 text documents after the pre-processing and annotation process. A summary of the collected data is after pre-processing mentioned in the following table3.1.

Table 3.1: Corpus Characteristics

Corpus Attribute	Attribute Value
Total no of Sports News Documents	43,306
Total no of Sentences	9,65,587
Average Sentences per Document	22.30
Total no of Words	1,02,40,032
Total no of Unique Words	2,02,830

3.2.5 Data Distribution

This corpus is comprised of documents relating to sports from five different leading Bengali online news portals namely- Prothom Alo, Daily Samakal, Kaler Kontho, Bhorer Kagoj, and Daily Nayadiganta. A source wise data distribution is presented in the following table 3.2

Table 3.2: Source wise Data Distribution

Data Source	Amount of Data Collected
Prothom Alo	7,595
Daily Samakal	614
Kaler Kontho	27,420
Bhorer Kagoj	5,802
Daily Nayadiganta	1893

The amount of data in the corpus varies depending on the sources. The most documents were submitted by Kaler Kontho (27,420), while the least were contributed by Nayadiganta (1893) and Samakal (614).

The data collected from all these souces are diftributed over four classes : Cricket, Football, Tennis and Athletics. Table 3.1 gives a overview of total number of documents for each category in the corpus which shows that all the categories do not have equal no of documents(Cricket has the highest) as the sports coverage of the news portals depends on the popularity of a particular sports segment. Here, the Cricket class has the most documents (30,032) and the most unique words (1,38,220), while the Athletics class has just 744 documents. On average, all the classes have more than 22 sentences in each text document. The class-wise summation of the unique words is 2,76,622, but for the entire corpus, the number

Table 3.3: Class wise Data Distribution

Category	No. of Documents	No. of Sentences	No. of Words	Average Sentences per Document	No. of unique words
Cricket	30,032	6,80,315	71,69,781	22.65	1,38,220
Football	11,429	2,46,299	26,63,483	21.55	94,050
Tennis	1101	21,041	2,22,152	19.11	22,153
Athletics	744	17,932	1,84,616	24.10	22,199

is 2,02,830, which is less. That means there are some common keywords for every class.

3.3 Proposed Framework of Sports News Classification

Supervised text classification is the process where the classifier model f is trained with a set of documents $D_{train} = \{d_1, d_2, ...d_n\}$ which are already labeled with a predefined set of categories $C_i(i = 1...m)$. The aim of the classifier f is to correctly assign a category C_i to an unseen document D_{test}.

$$f : D \Rightarrow C \tag{3.1}$$

Figure 3.1 shows an abstract representation of the framework that has been used for Sports News sub-classification process. In the pre-processing step raw textual documents are used as input. During this step the entire document-set is tokenized into individual tokens and removes any unwanted words/symbols for further processing.

The noise-free textual documents are used to extract important features during the feature extraction process. The TF-IDF values of various n-grams are used to represent these textual documents numerically in this step. Different feature vectors are generated in order for the classifier models to carry out their respective computations. At this step, the feature space of the entire data set is partitioned into train and test sets. The training data-set was fed into the classifier models to train and generate different Machine Learning models. Following the generation of the classifier models, these trained models provides their prediction results by labelling the unknown sports documents into their respective categories using the test data-set during the prediction step and finally we get the predicted labels from the classifier models.

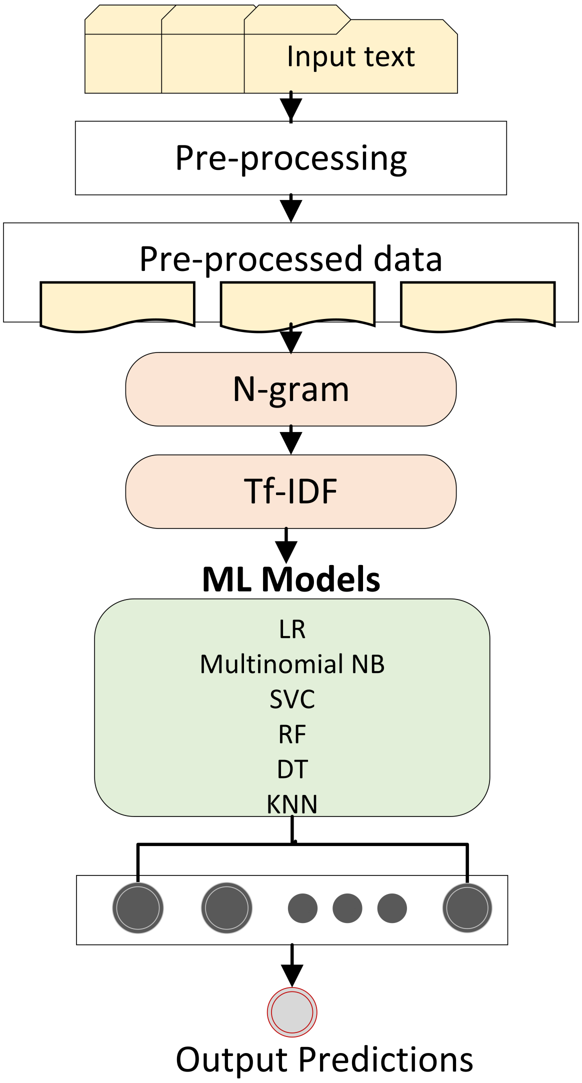


Figure 3.1: Proposed Methodology

3.3.1 Features Extraction

The pre-processed data-set is now ready to be used by the classifier models. To train the models, extracting relevant feature sets from the textual documents is required. In this step, the unorganised text sequence of the given corpus is converted into a structured feature space using the Term Frequency-Inverse Document Frequency(TF-IDF) of different N-grams.

- **N-gram**: By using the n-gram feature, the whole document is divided into n-terms so that each term's TF-IDF value can be calculated to numerically represent the textual document. Different combination of n-grams can be used. Here, we have used three different combination of n-grams:

- Unigram: Represents the vector space of the entire dataset by dividing the entire document-set into each single terms/words(n=1).

- Unigram+Bigram: Consecutive two terms along with single terms are considered to represent the feature space.

- Unigram+Bigram+Trigram: Consecutive three terms along with two and single terms are considered to represent the feature space.

TF-IDF: The TF-IDF value is used to extract the feature space of different n-grams. We have set the max_df and min_df values of Countvectorizer to reduce the unimportant feature space.
max_df is kept as 0.8 to ignore terms that appear in more than 80% of the documents.
min_df is kept 0.0009 to ignore terms that appear in less than 0.09% of the documents.

The examples of diiferent vector space generated by using the TF-IDF values of different n-grams is shown here.

$$\begin{array}{cc} & \begin{array}{ccccccc} \text{গতিতে} & \text{ছন্দে} & \text{জিতে} & \text{টস} & \text{বাংলাদেশ} & \text{ব্যাটিং} & \text{ম্যাচেই} \end{array} \\ \begin{array}{c} D_1 \\ D_2 \end{array} & \begin{bmatrix} 0 & 0 & 0.471 & 0.471 & 0.471 & 0.335 & 0.471 \\ 0.631 & 0.631 & 0 & 0 & 0 & 0.449 & 0 \end{bmatrix} \end{array}$$

Figure 3.2: Vector space generated by unigram

$$\begin{array}{cc} & \begin{array}{cccccccc} \text{গতিতে} & \text{ছন্দে} & \text{ছন্দে গতিতে} & \text{জিতে} & \text{জিতে ব্যাটিং} & \dots & \text{ম্যাচেই} & \text{ম্যাচেই টস} \end{array} \\ \begin{array}{c} D_1 \\ D_2 \end{array} & \begin{bmatrix} 0 & 0 & 0 & 0.343 & 0.343 & \cdots & 0.344 & 0.344 \\ 0.471 & 0.471 & 0.471 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \end{array}$$

Figure 3.3: Vector space generated by unigram-bigram

$$\begin{array}{cc} & \begin{array}{cccccc} \text{গতিতে} & \text{ছন্দে} & \text{ছন্দে গতিতে} & \text{ব্যাটিং ছন্দে গতিতে} & \dots & \text{ম্যাচেই টস জিতে} \end{array} \\ \begin{array}{c} D_1 \\ D_2 \end{array} & \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0.295 \\ 0.426 & 0.426 & 0.426 & 0.426 & \cdots & 0 \end{bmatrix} \end{array}$$

Figure 3.4: Vector space generated by unigram-bigram-trigram

3.3.2 Classifier Models Preparation

Several ML models are prepared as the baselines including the proposed classifier (SVC). The preparation of the models and its several tuned parameters described in the following:

After extracting important features from the document sets, different Machine Leaning algorithms are used to generate classifier models. The entire corpus is divided into two sets - training and testing data-set. The training data-set is used during the learning phase of the classifier models. After that the learned models are used to predict unknown document categories of the test data-set. Six ML methods (such as LR, NB, SVC, RF, DT, KNN) are implemented to investigate the Bengali Sports News Sub-classification task performance. Before performing the classification task, the various parameters are tuned to prepare the LR, NB, SVC, RF, DT, and KNN models. Table 3.4 shows a list of the parameters used in each model.

Table 3.4: Summary of the Parameters used in ML models

ML Models	**Parameters**
LR	solver='lbfgs', max_iter=600

NB	alpha = 1.0, class_prior=None,
	fit_prior = True

SVC	kernel = 'linear', random_state = 0

RF	criterion='gini', n_estimators=100,
	max_features=None

DT	criterion='gini', splitter='best',
	max_features=None

KNN	n_neighbors= 1

- **LR**: In the LR model, the ‘lbfgs’ optimizer is used with the inverse regularization strength is fixed to 1.0, and a maximum of 600 iterations are taken for solvers to converge.
- **NB**: In this work a Multinomial NB model is used which is most appropriate for features that represent counts or count rates(TF)[26] which is suitable for the given feature sets. The additive smoothing parameter α, of the NB model is set to 1. Prior class probabilities are learned and adjusted according to the data.
- **SVC**: For SVC, the 'linear' kernel is used that takes normal dot product any two given observations to convert the vector space. The 'l2' penalizer is set to avoid the sparse coefficient vectors and the random state is set to 0.
- **RF**: RF is implemented with 100 estimators or trees, and the Gini impurity measure is utilized to measure the quality of split in the tree. If there are at least two samples in an internal node, it is partitioned. During node partitioning, all device features are taken into account. It provides better performance as the prediction value chosen by the majority vote among a number of estimators can end up being better than the performance of any individual estimator.
- **DT**: For DT the 'best' splitter is used to choose the best split at each node. The 'gini' criteria is used as the previous RF model and total no of features are used as maximum feature.
- **KNN**: For the KNN model the similarity score is calculated for each neighbours, so the number of neighbours is taken as 1 as it showed better performance to categorize the imbalanced data-set.

3.3.3 Sports News Class Prediction

After the models have been trained, a test dataset is used to predict the document's class. To predict the class of a document, the trained models calculate a similarity score for the feature space extracted from the test dataset. Different

ML models use the *predict* function in the scikit-learn library to measure the similarity score. For the given feature space of the test document t_i, the ML model calculates the similarity score with the feature value for all the classes(C_i) in the dataset and finally assigns the class label with the highest similarity score.

$$t_i \Rightarrow \max_{Sim_{1,2,\ldots i}} C_i \tag{3.2}$$

3.4 Conclusion

This chapter discusses the methodology for creating a framework for Bangla Sports News sub-categorization. Six distinct feature vectors are created in order to generate a diverse feature space. All of the feature spaces are used to train the classifier models individually. Finally, we train six distinct classifier models on six distinct feature spaces. The following chapter discusses the proposed framework's experimental results analysis.

Chapter 4

Implementation Details

4.1 Introduction

This chapter explains how the system is implemented in practice. The implementation process following the steps of the system architecture discussed in the previous chapter is described here with system requirements and system set up. The impact of the proposed system is also discussed in this chapter.

4.2 System Requirements

The system requirements for the implementation(both hardware and software) of the project is mentioned in the subsequent sections.

4.2.1 Hardware Requirements

The hardware required to implement the system architecture is discussed here. The experiments were implemented on a general-purpose computer with an Intel® Core$^{\text{TM}}$ i7-7700H processor @ 3.60 GHz, 4 cores and 8 logical processors with 8 GB of RAM, and Windows 10.

4.2.2 Software Requirements

The experiments were implemented by installing some required softwares.

- The entire processing of the proposed framework is done using Python programming language on Windows 10 operating system.
- Machine learning models are built with scikit-learn (0.24.1) packages, and processed using numpy(1.12.1) packages.

- A dedicated library for pre-processing Bangla text documents named as BnPreprocessing is used at the pre-processing step.
- The codes were generated and run on Jupyter Notebook IDE.

4.3 System Set up and Runing

The system was implemented by setting up the whole environment in Jupyter notebook using Python programming language. Scikit-learn and numpy packages were installed to build the models using training dataset. During the training setup the input files were converted to CSV files from text documents for further processing. The models were trained on GPU runtime. The trained models were saved in local disk space and tested using the test dataset.

4.3.1 Implementation Snapshot

Our model predicts Sports news categories for the given news documents. The label of the predicted category is given as the output of the system while the news documents are given as input.

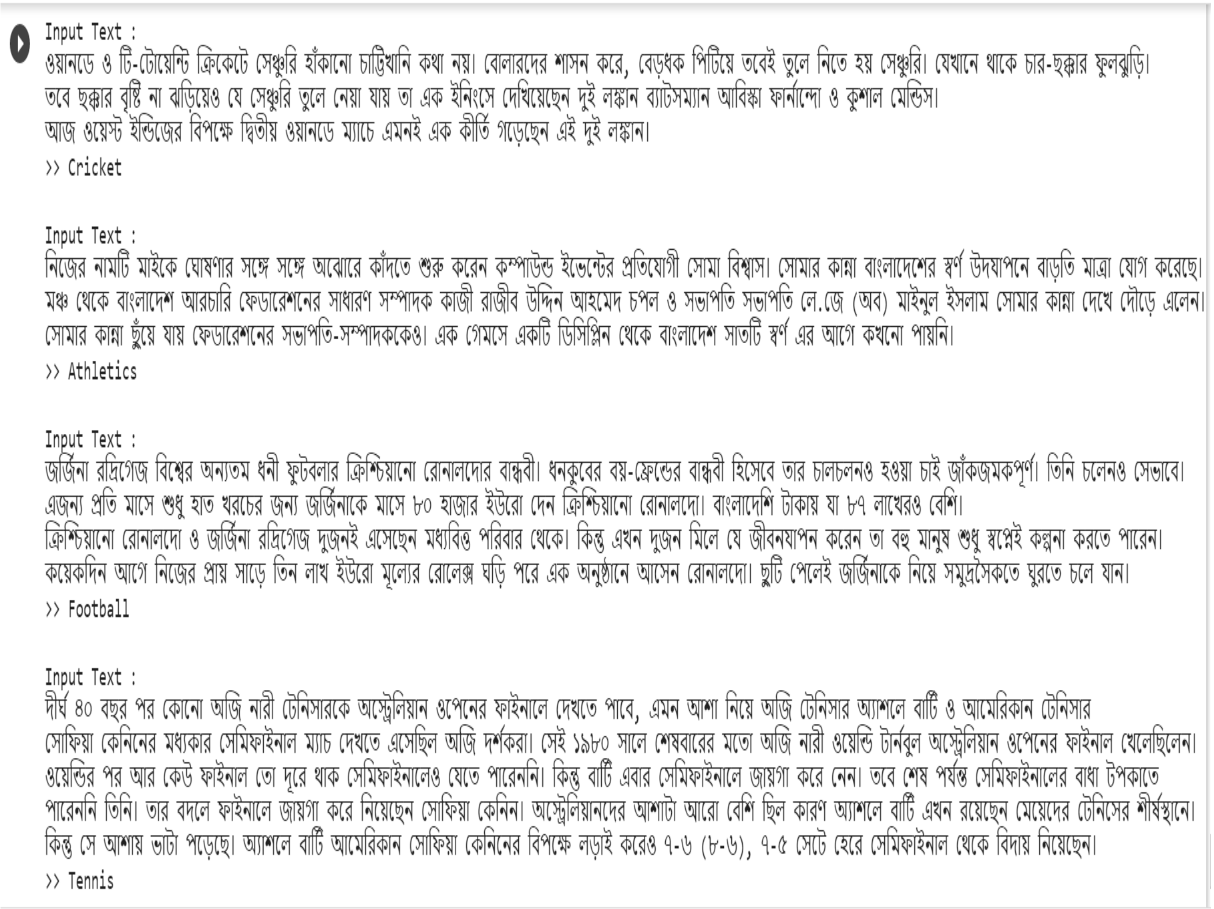
A snapshot of the system input and output is showed in the following figure 4.1

4.4 Impact Analysis

This work has some significant impacts which can be addressed both socially and ethically. As this is an era of electronic technology an automated system of text organization and classification can serve several issues that in occurring in the virtual world.

4.4.1 Social and Environmental Impact

Sports news attracts a large group of audiences of different ages. It inspires the youngsters of our generation to get motivated by the success stories of the legendary players. With the help of an automated system, they can closely follow their inspirational role models and go through different articles for guidance and



```
Input Text :
ওয়ানডে ও টি-টোয়েন্টি ক্রিকেটে সেঞ্চুরি হাঁকানো চাট্টিখানি কথা নয়। বোলারদের শাসন করে, বেড়ধক পিটিয়ে তবেই তুলে নিতে হয় সেঞ্চুরি। যেখানে থাকে চার-ছক্কার ফুলঝুড়ি।
তবে ছক্কার বৃষ্টি না ঝড়িয়েও যে সেঞ্চুরি তুলে নেয়া যায় তা এক ইনিংসে দেখিয়েছেন দুই লঙ্কান ব্যাটসম্যান আবিষ্কা ফার্নান্দো ও কুশাল মেন্ডিস।
আজ ওয়েস্ট ইন্ডিজের বিপক্ষে দ্বিতীয় ওয়ানডে ম্যাচে এমনই এক কীর্তি গড়েছেন এই দুই লঙ্কান।
>> Cricket

Input Text :
নিজের নামটি মাইকে ঘোষণার সঙ্গে সঙ্গে অঝোরে কাঁদতে শুরু করেন কম্পাউন্ড ইভেন্টের প্রতিযোগী সোমা বিশ্বাস। সোমার কান্না বাংলাদেশের স্বর্ণ উদযাপনে বাড়তি মাত্রা যোগ করেছে।
মঞ্চ থেকে বাংলাদেশ আরচারি ফেডারেশনের সাধারণ সম্পাদক কাজী রাজীব উদ্দিন আহমেদ চপল ও সভাপতি সভাপতি লে.জে (অব) মাইনুল ইসলাম সোমার কান্না দেখে দৌড়ে এলেন।
সোমার কান্না ছুঁয়ে যায় ফেডারেশনের সভাপতি-সম্পাদককেও। এক গেমসে একটি ডিসিপ্লিন থেকে বাংলাদেশ সাতটি স্বর্ণ এর আগে কখনো পায়নি।
>> Athletics

Input Text :
জর্জিনা রদ্রিগেজ বিশ্বের অন্যতম ধনী ফুটবলার ক্রিশ্চিয়ানো রোনালদোর বান্ধবী। ধনকুবের বয়-ফ্রেন্ডের বান্ধবী হিসেবে তার চালচলনও হওয়া চাই জাঁকজমকপূর্ণ। তিনি চলেনও সেভাবে।
এজন্য প্রতি মাসে শুধু হাত খরচের জন্য জর্জিনাকে মাসে ৮০ হাজার ইউরো দেন ক্রিশ্চিয়ানো রোনালদো। বাংলাদেশি টাকায় যা ৮৭ লাখেরও বেশি।
ক্রিশ্চিয়ানো রোনালদো ও জর্জিনা রদ্রিগেজ দুজনই এসেছেন মধ্যবিত্ত পরিবার থেকে। কিন্তু এখন দুজন মিলে যে জীবনযাপন করেন তা বহু মানুষ শুধু স্বপ্নেই কল্পনা করতে পারেন।
কয়েকদিন আগে নিজের প্রায় সাড়ে তিন লাখ ইউরো মূল্যের রোলেক্স ঘড়ি পরে এক অনুষ্ঠানে আসেন রোনালদো। ছুটি পেলেই জর্জিনাকে নিয়ে সমুদ্রসৈকতে ঘুরতে চলে যান।
>> Football

Input Text :
দীর্ঘ ৪০ বছর পর কোনো অজি নারী টেনিসারকে অস্ট্রেলিয়ান ওপেনের ফাইনালে দেখতে পাবে, এমন আশা নিয়ে অজি টেনিসার অ্যাশলে বার্টি ও আমেরিকান টেনিসার
সোফিয়া কেনিনের মধ্যকার সেমিফাইনাল ম্যাচ দেখতে এসেছিল অজি দর্শকরা। সেই ১৯৮০ সালে শেষবারের মতো অজি নারী ওয়েন্ডি টার্নবুল অস্ট্রেলিয়ান ওপেনের ফাইনাল খেলেছিলেন।
ওয়েন্ডির পর আর কেউ ফাইনাল তো দূরে থাক সেমিফাইনালেও যেতে পারেননি। কিন্তু বার্টি এবার সেমিফাইনালে জায়গা করে নেন। তবে শেষ পর্যন্ত সেমিফাইনালের বাধা টপকাতে
পারেননি তিনি। তার বদলে ফাইনালে জায়গা করে নিয়েছেন সোফিয়া কেনিন। অস্ট্রেলিয়ানদের আশাটা আরো বেশি ছিল কারণ অ্যাশলে বার্টি এখন রয়েছেন মেয়েদের টেনিসের শীর্ষস্থানে।
কিন্তু সে আশায় ভাটা পড়েছে। অ্যাশলে বার্টি আমেরিকান সোফিয়া কেনিনের বিপক্ষে লড়াই করেও ৭-৬ (৮-৬), ৭-৫ সেটে হেরে সেমিফাইনাল থেকে বিদায় নিয়েছেন।
>> Tennis
```

Figure 4.1: Implementation Snapshot

inspiration. It also helps several local and global advertising companies detect the popular areas of sports to endorse their products using text mining facilitated by an automated classification system. Thus our work of sports news classification can have several social and economic impact on the current environment.

4.4.2 Ethical Impact

With the growing popularity of different sports and sports-persons, it is essential for the organizers and the players to follow the ethical code and conduct. Several issues come in the news about match-fixing, ball tempering, and other unauthorized activities published in the report to spread awareness among mass audiences. These ethical issues can be addressed and quickly investigated by implementing our automated system.

4.5 Conclusion

This section covers the implementation of the proposed models' system architecture as well as other system specifications. Before moving on to the result analysis, which is outlined in the next section, this section illustrates the system implementation as well as the input and output cases of the system. This section ends with a discussion of the work's social and ethical implications.

Chapter 5

Results and Discussions

5.1 Introduction

The detailed experimental study of various ML methods in the established corpus is explained in Chapter 5. The performance of implemented models is also investigated in this chapter using various evaluation measures (such as precision, recall, accuracy, and the $f1$-score). This Chapter also includes a thorough comparison of different approaches with current techniques.A detailed error analysis of the proposed model is included at the end of this Chapter for better understanding.

5.2 Experiments

The performance of our model is evaluated for four different sports categories: Cricket, Football, Tennis, and Athletics. Each category contains documents of varying sizes, and the performance of the classifier model also varies according to the total number of documents of that class. The performance of the proposed framework is evaluated by comparing it with three different feature spaces and and six distinct supervised classifier models.

5.2.1 Experimental Data

As this is a supervised machine learning problem, the entire dataset was divided into two portions, Train and Test. For training purpose, 80% of the data samples are taken randomly, containing 34,644 documents. The remaining 20% samples having 8662 documents are used for testing to evaluate the classifier performance. The entire division of the train and test set is showed in the following table 5.1.

Table 5.1: Data Distribution over train and test set

Categories	**Train-set(80%)**	**Test-set(20%)**
Cricket	23,988	6044
Football	9141	2288
Tennis	912	189
Athletics	603	141
Total Corpus	34,644	8662

5.2.2 Evaluation Measures

The proposed emotion detection system was evaluated two different phases: training phase and test phase. Several measures such as precision (P), recall (R), and F_1-score are considered for model's evaluation.

- **Precision:** Calculates the number of documents (d_i) that actually belong to class (C) among the samples (d_i) labeled as class (C). It represents the percentage of correctly predicted labels for a given class from all the predicted labels of that class.

$$\rho = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5.1}$$

- **Recall :** Calculates how many documents (d_i) are correctly labeled as class (C) among the total number of documents (d_i) of class (C). It represents the percentage of correctly predicted labels from all the actual number of labels of that class in the test set of documents.

$$R = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5.2}$$

- **F1-score :** F1-score is the harmonic mean of precision and recall. It is calculated using the following equation,

$$f_1 = \frac{2 * \rho * R}{\rho + R} \tag{5.3}$$

  Where, ρ represents Precision and R represents Recall.

- **Weighted F1-score:** For imbalanced dataset Weighted F1-score is used to evaluate the performance of the classifier models. It is defined by the following equation.

$$WF = \frac{1}{N} \sum_{i=1}^{c} F_i n_i \tag{5.4}$$

where,

$$N = \sum_{i=1}^{c} n_i \tag{5.5}$$

here, N, F_i and n_i denotes total samples in test-set, f1-score and number of samples in class(c_i).

5.2.3 Results Analysis

The evaluation results for the developed models and a detailed errors analysis are presented in the subsequent subsections.

5.2.4 Model Evaluation

The outcomes of the proposed classifier models are showed with respect to different evaluation metrics - Precision, Recall and F1-score in Table 5.2. From the evaluation results of individual classes it is evident that the classifier models showed better performance in identifying Cricket news documents as it has the highest number of test and train set and worked badly for Athletics news documents. For all the sports categories it is evident that the performance of Unigram+Bigram+Trigram feature is better for most of the classifier models, so it is chosen as the optimum feature space for our proposed framework. But for KNN classifier model using Unigram feature space gives better results when the number of neighbours is $k = 1$. From all the classifier models NB and SVC model performed the best for identifying Cricket documents with an class-based F1-score of 98.53%. However the LR and RF models also performed well but the KNN models showed poor Precision, Recall and F1-scores as the KNN model does not perform well for imbalanced corpus size. To determine the overall performance of the classifier models Weighted F1-score is presented in the next table 5.3.

Table 5.2: Evaluation results of different models on the test set

Unigram Feature Space												
	Cricket			**Football**			**Tennis**			**Athletics**		
	P	R	F	P	R	F	P	R	F	P	R	F
LR	97.91	99.02	98.46	96.78	95.80	96.29	95.91	86.77	91.11	94.69	75.88	84.25
NB	98.26	98.78	98.51	95.99	96.15	96.07	96.51	87.83	91.96	90.98	78.72	84.41
SVC	98.14	98.87	98.51	96.77	95.67	96.22	94.41	89.41	91.84	91.67	85.82	86.64
RF	97.94	98.89	98.41	96.51	95.50	96.00	94.38	88.88	91.55	91.45	75.89	82.94
DT	97.50	98.10	97.80	95.87	93.31	94.57	83.17	88.89	85.93	73.03	78.72	75.77
KNN	74.59	99.26	85.17	92.83	20.37	33.41	87.14	32.28	47.10	89.36	29.78	44.68
Unigram+Bigram Feature Space												
LR	97.89	98.96	98.42	96.52	95.80	96.16	95.91	86.77	91.11	94.55	73.76	82.87
NB	98.46	98.58	98.52	94.37	96.72	95.53	96.34	83.60	89.52	92.16	66.67	77.37
SVC	98.11	98.86	98.48	96.68	95.60	96.13	94.97	89.95	92.39	92.37	85.82	88.97
RF	97.95	98.82	98.39	96.46	95.41	95.93	93.33	88.89	91.06	92.26	79.43	85.49
DT	97.68	98.06	97.87	95.65	94.10	94.87	88.24	87.30	87.77	72.44	80.14	76.09
KNN	73.11	99.35	84.23	91.53	14.16	24.53	83.64	24.34	37.70	92.31	25.53	40.00
Unigram+Bigram+Trigram Feature Space												
LR	97.87	98.99	98.43	96.61	95.80	96.20	95.91	86.77	91.11	95.41	73.76	83.20
NB	98.48	98.58	98.53	94.09	96.77	95.41	96.30	82.54	88.89	93.81	64.54	76.47
SVC	98.13	98.88	98.52	96.68	95.63	96.15	94.97	89.95	92.39	92.37	85.82	88.97
RF	97.95	98.87	98.41	96.46	95.37	95.91	94.38	88.89	91.55	92.56	79.43	85.50
DT	96.76	96.77	96.77	92.85	92.53	92.69	85.57	87.83	86.68	74.83	75.89	75.35
KNN	73.01	99.35	84.17	91.57	13.77	23.94	83.34	23.81	37.04	92.50	26.24	40.89

In table 5.3 the Weighted F1-scores of individual classifier models are defined. From the table it is observed that the SVC model has the highest score of 97.60%(WF) for Unigram+Bigram+Trigram feature space. Other than this the

Table 5.3: Performance evaluation using Weighted F1-score

Feature Space	**Weighted F1-score**					
	LR	**NB**	**SVC**	**RF**	**DT**	**KNN**
Unigram	97.44	97.38	97.53	97.32	96.33	70.00
Unigram+ Bigram	97.41	97.19	97.57	97.37	96.50	66.73
Unigram+ Bigram+ Trigram	97.43	97.14	**97.60**	97.39	96.28	66.53

LR model also showed a good performance with a higher WF score of 97.43%. Again the overall performance of KNN model was poor.

5.2.5 Comparisons with Baselines

Finally the Unigram+Bigram+Trigram feature space is selected to compare the overall performance of the proposed framework. The figure 5.1 plots the F1-scores of different classifier models to classify four categories of sports news documents.

The figure shows the comparison of the F1-scores for different methods for the chosen feature space. It also shows the individual F1-scores for each class.

5.2.6 Comparison with Existing Techniques

As per the recent works , no significant work has been done for the sub-categorization of the Sports news documents in Bengali. Therefore, this research adopted several recent techniques that have been explored on similar tasks for general news classification. For proper evaluation previous methods have implemented on our developed dataset and their performance are compared with the proposed technique. Table 5.4 shows the comparison in terms of Weighted F1-score. In [9] they proposed a system using the LR with TF-IDF values of the entire feature space for classifying news categories. In [18] they proposed SVC based classifier model

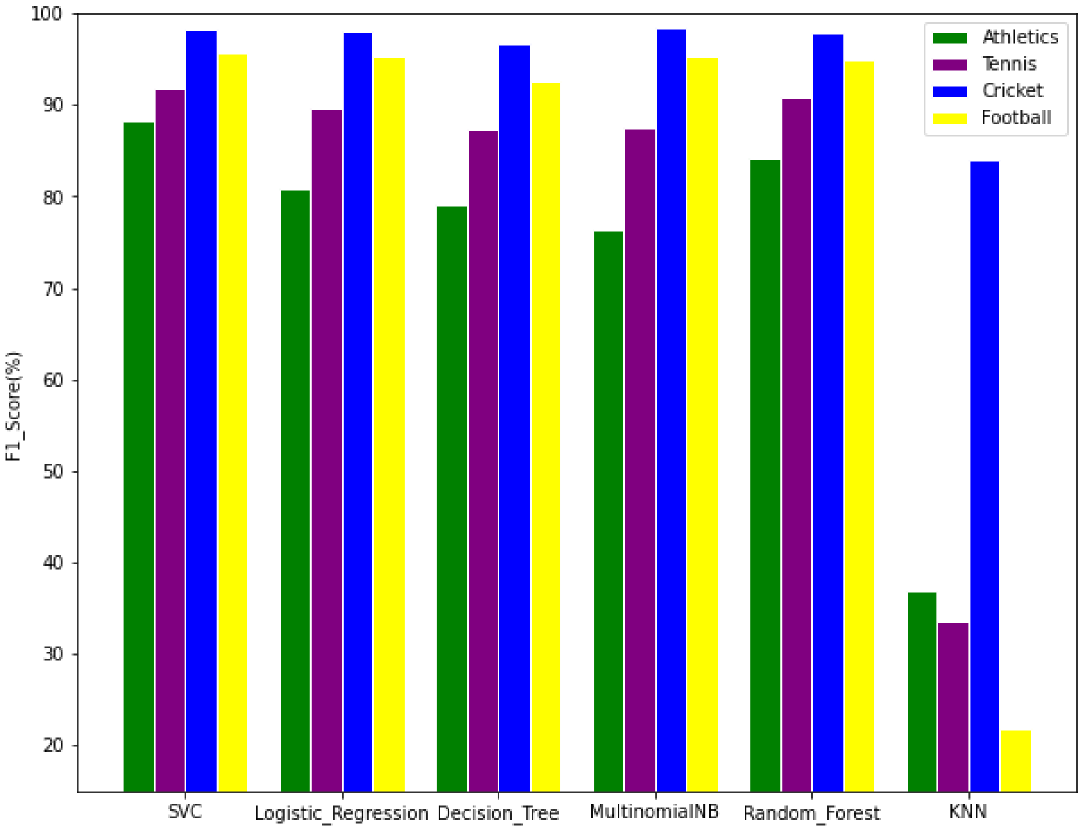


Figure 5.1: Comparison of different ML models using unigram-bigram-trigram feature

Table 5.4: Comparison between the proposed and other existing techniques

Techniques	Weighted F1 Score
Shah et al.[9]	97.01
Mandal et al.[18]	96.11
Islam et al.[16]	97.24
Proposed	**97.60**

with sigmoid kernel using the TF-IDF feature space. Another work is done for news classification in [16] where uni-gram feature space is addressed with C-SVM for multiclass classification. All these methods are compared with our proposed method training them on our dataset.

5.3 Discussion

From the above result analysis, the following points can be discussed to give better insights :

- For each classifier model, the F1-score varies in each category depending on the size of the documents of that category.

- As the corpus is imbalanced, the class Cricket has the highest number of documents, hence got the highest percentage in F1-score and Athletics has the lowest score.
- The proposed model using a diverse Unigram+Bigram+Trigram feature space gives better results than the other feature spaces because the combination of three n-grams takes the syntactic similarity into consideration while using only single n-gram doesn't perform well.
- Among all the classifier models SVC and LR has the the highest performance measure and KNN model has the lowest evaluation score as it is not a standard model for text classification.
- The SVC model using 'linear' kernel gives the highest performance score as it is the best fit for multi-class classification problem.

5.3.1 Error Analysis

To analysis how the models works in predicting the document categories we've presented a confusion matrix of the proposed model in Figure 5.2

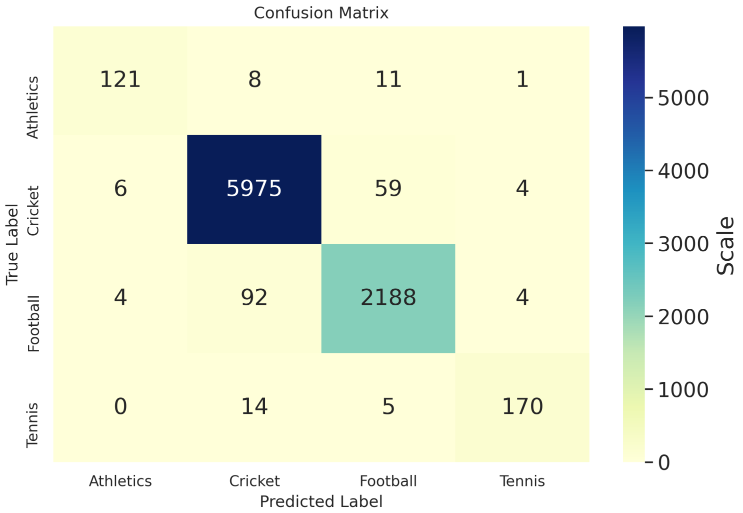


Figure 5.2: Confusion Matrix of the proposed work

From the figure, we can see that the proposed model fails to detect the classes with a smaller number of documents(Athletics, Tennis). Still, the classes with the highest number of documents(Cricket, Football) can be easily detected for most cases. For Cricket, a total of 5975 documents were correctly predicted.

The Cricket documents mainly get confused with Football documents. Football and Tennis have the highest false-positive rate for Cricket, but Athletics mostly gets confused with Football documents. The presence of similar sports in the Olympics that resembles the keywords of football may be the reason.

From the confusion matrix, we get a quantitative representation of the correctly predicted documents for each class. But the contextual analysis with a detailed view of which type of documents can be correctly predicted by our proposed and other ML models is illustrated using some examples in Table 5.5. When we look into the examples, it is evident that most of the Tennis and athletics documents get confused with Cricket and football. When there is the word 'World-cup' which is commonly used for Football or Cricket, but in the actual news, it is used for athletics, the model makes wrong predictions. Also, when other Olympic games are classified as athletics, there is a large portion of news about Football at the Olympics; in these cases, the model gets confused. Increasing the number of athletics and tennis documents in the dataset can improve the models' performance to learn sports-specific features.

5.4 Conclusion

This chapter provides the results of our proposed Sports news sub-classification framework and also investigates performance evaluation from various aspects. The comparative performances across every sports news categories, for different feature spaces as well as for different machine learning methods are represented here. From all these perspective, our proposed framework using Unigram+Bigram+Trigram feature space proved to be superior for SVC and LR models. The next chapter concludes this thesis and provides future recommendations.

Table 5.5: Contextual representation of different classifier models in predicting Sports news doceumnets. The wrong predictions are marked in red. C, F, T and A denotes the labels Cricket, Football, Tennis and Athletics.

Example	SVC	NB	LR	DT	RF	KNN
জন ইসনার সেমিফাইনালে ডেল পোত্রোকে হারিয়ে ফাইনালে উত্তীর্ণ হন। ম্যাচ শেষে উচ্ছ্বসিত ইসনার বলেছেন যদিও দীর্ঘদিন ধরেই নিজের সেরা টেনিস খেলে আসছি,কিন্তু কোনটাতেই শুরুটা ভাল হয়নি।	T	F	T	T	T	C
টেনিসের খুব পরিচিত দৃশ্য -ম্যাচ শেষে বিজয়ী গ্যালারিতে ছুড়ে দিচ্ছেন আর্ম বা হেয়ার ব্যান্ড। খুশির ছোঁয়া বেশি লাগলে তোয়ালেটাও উড়ে যায় দর্শকদের দিকে। তোয়ালে নিয়ে খেলোয়াড়-দর্শকের এ টানাটানিতে তোয়ালে-সংকটে আছে কর্তৃপক্ষ।	T	F	C	T	T	C
তাঁর এখন থাকার কথা ছিল আজারবাইজানে,কিন্তু বাড়িতে অলস সময় কাটাচ্ছেন শ্যুটার আবদুল্লাহ বাকি। ভিসা না পেয়ে আগস্টের বিশ্বকাপ শ্যুটিংয়ে যেতে পারলেনেনা কমনওয়েলথের রুপাজয়ী-শ্যুটার	A	F	F	C	F	C
দক্ষিণ কোরিয়ায় আর মাত্র একদিন পর পর্দা উঠবে শীতকালীন অলিম্পিকের। এর মধ্যেই দুঃসংবাদ পেল রাশিয়া, রাষ্ট্রীয় পৃষ্ঠপোষকতায় অ্যাথলেটদের ডোপিংয়ের অভিযোগে এবারের অলিম্পিক থেকে রাশিয়াকে নিষিদ্ধ করেছে অলিম্পিক কমিটি।	F	F	F	A	F	F

Chapter 6

Conclusion

This Chapter summarizes the thesis with highlighting the major contributions of this work including a few weaknesses in the current implementation. This Chapter also provides the few recommendations for further improvements of the proposed system.

6.1 Conclusion

Categorization and organization of text documents have a wide area of implementation, and there is only a small amount of work that has been conducted for Bangla text classification. Moreover, most of the works are executed using an insufficient dataset without any proper validation of the dataset's quality. This problem is needed to be addressed if we want to build an efficient classifier model.

In this paper, a sub-categorization method of the Bangla Sports news documents has been studied. It investigates the performance of the classifiers by proposing a technique that contributes to the area of text classification by creating a large corpus to analyze different feature spaces on multiple ML models. A corpus of 43,306 labelled documents has been used to train six supervised Machine learning models. A detailed description of the corpus creation process is given with proper quality assessment. Finally, we trained the models using the feature space generated from the TF-IDF values of different n-grams. These models were evaluated using the Weighted F1-score as a performance measure. From the results, it has been observed that the Unigram+Bigram+Trigram feature space gives the highest performance for most of the classifier models. For this imbalanced corpus, SVC and Logistic Regression classifiers provide satisfactory performance with a highest weighted f1-score of 97.60%. The analysis results of our work can be used as a benchmark for further research works, and our developed corpus can highly contribute to this arena.

6.1.1 Limitations

- Did not work with all the Sports news categories.
- Did not address the other advanced DL methods for building the classifier models.
- Could not ensure balanced class size while developing the corpus.

6.2 Future Recommendations

As there are various categories of texts producing every day, the text classification methods has an open area of improvement. Some of the aspects of this problem is addressed in this work which can be further improved using various methods in future. Some future recommendations are presented here,

- In this work, we mainly used the machine learning models for the subcategorization process. The performance of the classifier models can be further improved using the Deep learning methods of text classification which will be investigated in future.
- Various ensemble methods can also be implemented to investigate the class dependent features to improve the model's performance.
- We intend to extend the corpus with a balanced class size with proper annotation to develop the DL models. Thus this work can reach its state-of-the-art model to investigate Bangla text classification on Sports news documents.

References

[1] M. T. Alam and M. M. Islam, “Bard: Bangla article classification using a new comprehensive dataset,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, IEEE, 2018, pp. 1–5 (cit. on p. 3).

[2] M. Thangaraj and M. Sivakami, “Text classification techniques: A literature review.,” *Interdisciplinary Journal of Information, Knowledge & Management*, vol. 13, 2018 (cit. on p. 8).

[3] G. Kaur and K. Bajaj, “News classification and its techniques: A review,” *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 18, no. 1, pp. 22–26, 2016 (cit. on p. 9).

[4] M. Bednarek, *Evaluation in media discourse: Analysis of a newspaper corpus.* A&C Black, 2006 (cit. on p. 9).

[5] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Information*, vol. 10, no. 4, p. 150, 2019 (cit. on p. 15).

[6] V. K. Vijayan, K. Bindu, and L. Parameswaran, “A comprehensive study of text classification algorithms,” in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017, pp. 1109–1113 (cit. on p. 15).

[7] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, “Text classification using machine learning techniques.,” *WSEAS transactions on computers*, vol. 4, no. 8, pp. 966–974, 2005 (cit. on p. 15).

[8] T. S. Zakzouk and H. I. Mathkour, “Comparing text classifiers for sports news,” *Procedia Technology*, vol. 1, pp. 474–480, 2012 (cit. on p. 15).

[9] K. Shah, H. Patel, D. Sanghvi, and M. Shah, “A comparative analysis of logistic regression, random forest and knn models for the text classification,” *Augmented Human Research*, vol. 5, no. 1, pp. 1–16, 2020 (cit. on pp. 15, 37, 38).

[10] N. Bidi and Z. Elberrichi, “Feature selection for text classification using genetic algorithms,” in *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*, IEEE, 2016, pp. 806–810 (cit. on p. 15).

[11] A. M. El-Halees, “Arabic text classification using maximum entropy,” *IUG Journal of Natural Studies*, vol. 15, no. 1, 2015 (cit. on p. 15).

[12] T. Zia, Q. Abbas, and M. P. Akhtar, “Evaluation of feature selection approaches for urdu text categorization.,” *International Journal of Intelligent Systems & Applications*, vol. 7, no. 6, 2015 (cit. on p. 15).

[13] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev, and A. Alizade, “Empirical study of online news classification using machine learning approaches,” in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, 2018, pp. 1–6 (cit. on p. 15).

[14] W. Zhang, X. Tang, and T. Yoshida, “Tesc: An approach to text classification using semi-supervised clustering,” *Knowledge-Based Systems*, vol. 75, pp. 152–160, 2015 (cit. on p. 16).

[15] C. Li, G. Zhan, and Z. Li, “News text classification based on improved bi-lstm-cnn,” in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, IEEE, 2018, pp. 890–893 (cit. on p. 16).

[16] M. S. Islam, F. E. M. Jubayer, and S. I. Ahmed, “A support vector machine mixed with tf-idf algorithm to categorize bengali document,” in *2017 international conference on electrical, computer and communication engineering (ECCE)*, IEEE, 2017, pp. 191–196 (cit. on pp. 16, 17, 38).

[17] S. Al Mostakim, F. Ehsan, S. M. Hasan, S. Islam, and S. Shatabda, “Bangla content categorization using text based supervised learning methods,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, IEEE, 2018, pp. 1–6 (cit. on pp. 16, 17).

[18] A. K. Mandal and R. Sen, “Supervised learning methods for bangla web document categorization,” *arXiv preprint arXiv:1410.2045*, 2014 (cit. on pp. 16, 17, 37, 38).

[19] A. N. Chy, M. H. Seddiqui, and S. Das, “Bangla news classification using naive bayes classifier,” in *16th Int’l Conf. Computer and Information Technology*, IEEE, 2014, pp. 366–371 (cit. on p. 16).

[20] F. Quadery, A. Al Maruf, T. Ahmed, and M. S. Islam, “Semi supervised keyword based bengali document categorization,” in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, IEEE, 2016, pp. 1–5 (cit. on p. 16).

[21] P. Dhar, M. Abedin, *et al.*, “Bengali news headline categorization using optimized machine learning pipeline.,” *International Journal of Information Engineering & Electronic Business*, vol. 13, no. 1, 2021 (cit. on pp. 16, 17).

[22] M. M. H. Shahin, T. Ahmmed, S. H. Piyal, and M. Shopon, “Classification of bangla news articles using bidirectional long short term memory,” in *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2020, pp. 1547–1551 (cit. on p. 16).

[23] D. M. Eler, D. Grosa, I. Pola, R. Garcia, R. Correia, and J. Teixeira, "Analysis of document pre-processing effects in text and opinion mining," *Information*, vol. 9, no. 4, p. 100, 2018 (cit. on p. 20).

[24] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, IEEE, vol. 3, 2003, pp. 1661–1666 (cit. on p. 20).

[25] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, "Automatic labeling of topics," in *2009 Ninth International Conference on Intelligent Systems Design and Applications*, IEEE, 2009, pp. 1227–1232 (cit. on p. 21).

[26] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *Australasian Joint Conference on Artificial Intelligence*, Springer, 2004, pp. 488–499 (cit. on p. 27).