Major Project Proposal Report on

“TEXT BASED FEEDBACK ANALYSIS”

****

in partial fulfillment for the award of the degree of

Bachelor of Information Technology Engineering

**Submitted By:**

Adarsha Wagle [20120001]

Amir Poudel [20120003]

Dikshya Bhujel [20120017]

Kamal Lamichhane [20120027]

**Submittedto**

**Department of Computer and IT Engineering**

**Everest Engineering College**

**Sanepa-2, Lalitpur**

June, 2024

TABLE OF CONTENTS

# 

# LIST OF FIGURES

# Chapter 1: INTRODUCTION………………………………………….……..5

# 1.1 Background …………………………………………………….…..…… 6

# Chapter 2: LITERATURE REVIEW ……………………………….…..…..8

# Chapter 3: METHODOLOGY ………………………………………..…….. 9

# 3.1 Introduction …………………………………………………….…….…. 9

# 3.2 Block Diagram…………………………………………………………....10

# 3.3 Methodology……………………………………………………….....…..10

# 3.3.1. Data Collection and Preprocessing ……………………………......11

# 3.3.2. Text Normalization and Tokenization…………………………….….11

# 3.3.3 Building a Sentiment Analysis………………………………….…..11

# 3.3.4 Testing and Optimization……………………………………………...12

# 3. Machine Learning Methods……………………………………….….13

# 3.3.5 Naive Bayes……………………………………………………….….. 13

# 3.3.6 K-Nearest Neighbor…………………………………………….…..…14

# 3.3.7 Random Forest…………………………………………………..…… 14

# Chapter 4: RESULT AND ANALYSIS………………………………….…...15

# CONCLUSION………………………………………………………….…….20

# REFERENCES ……………………………………………………………….21

# APPENDIX 21

# 

LIST OF FIGURES

Figure 3.2: Block Diagram …………………………………….………10.

Figure 4.1:Classification of precision diagram………………………………………………………………….17.

Figure 4.2: Graph of

Root means square ……………………………………………………18.

Figure 4.3: Graph of

accuracy…………………………………………………………………19.

# 

# 

# 

# 

# 

# 

# Chapter 1: INTRODUCTION

The Internet today contains a huge quantity of textual data, which is growing every day. The text is a prevalent data format on the web, since it is easy to generate and publish. What is hard nowadays is not availability of useful information but rather extracting it in the proper context from the vast ocean of content. It is now beyond human power and time to seed through it manually; therefore, the research problem of automatic categorization and organizing data is apparent.

Textual information can be divided into two main domains: facts and opinions. While facts focus on objective data transmission, the opinions express the sentiment of their authors. Initially, the research has mostly focused on the categorization of the factual data.Today, we have web search engines which enable search based on the keywords that describe the topic of the text. The search for one keyword can return a large number of pages.For example, Google search for the word “startrek” finds more than 2.3 million pages. These articles include both objective facts about the movie franchise (e.g. Wikipedia article) and subjective opinions from the users (e.g. review from critics).

Sentiment analysis aims to uncover the attitude of the author on a particular topic from the written text. Other terms used to denote this research area include “opinion mining” and “subjectivity detection”. It uses natural language processing and machine learning techniques to find statistical and/or linguistic patterns in the text that reveal attitudes. It has gained popularity in recent years due to its immediate applicability in business environments, such as summarizing feedback from the product reviews, discovering collaborative recommendations, or assisting in election campaigns.

## 

## 

## 1.1 Background

The origins of text feedback analysis can be traced back to the early days of content analysis, a research method used to systematically analyze written, spoken, or visual communication. Content analysis emerged in the early 20th century, initially applied in fields like journalism and communication studies to understand the patterns and themes in large volumes of textual data. Researchers manually coded texts, identifying recurring themes and patterns to derive insights.

The development of computational techniques in the mid-20th century marked a significant shift in text analysis. The advent of computers enabled more efficient processing of text data, leading to the creation of early text analysis software in the 1950s and 1960s. These early systems were rudimentary, focusing on word frequency counts and simple keyword searches.

The 1980s and 1990s saw substantial advancements with the rise of natural language processing (NLP) and machine learning. Researchers began developing algorithms that could understand and process human language more effectively. This period saw the introduction of techniques like part-of-speech tagging, syntactic parsing, and the development of basic sentiment analysis tools.

The explosion of the internet and the advent of social media in the late 1990s and early 2000s generated vast amounts of user-generated content, creating a new impetus for text feedback analysis. Businesses and researchers recognized the potential of analyzing online reviews, social media posts, and other forms of digital feedback to gain insights into consumer behavior and preferences.

The mid-2000s to 2010s witnessed significant progress with the advent of advanced machine learning models and the increasing availability of computational power. Techniques such as latent Dirichlet allocation (LDA) for topic modeling and the use of support vector machines (SVMs) for text classification became mainstream. During this period, sentiment analysis evolved, incorporating more sophisticated approaches that could capture nuances in human language.

In recent years, the development of deep learning and neural networks has revolutionized text feedback analysis. Models like Word2Vec, GloVe, and more recently, transformer-based models like BERT and GPT-3, have dramatically improved the ability to understand and generate human language. These models can analyze complex sentiments, detect sarcasm, and understand context at a level previously unattainable.

Today, text feedback analysis is an integral part of various industries, facilitated by advanced tools and platforms that leverage state-of-the-art NLP techniques. Organizations use these tools to analyze customer reviews, social media interactions, employee feedback, and more, driving strategic decision-making and fostering better engagement with their stakeholders. As technology continues to evolve, the field of text feedback analysis is poised to become even more sophisticated, enabling deeper and more accurate insights from textual data.

## 

## 

## 

## 

## Chapter 2: LITERATURE REVIEW

The study evaluates the effectiveness of different machine learning algorithms and feature extraction methods for sentiment analysis of reviews and comments. The experiments are conducted using supervised (Naive Bayes, Decision Trees, Maximum-Entropy) and unsupervised (K-Means Clustering) learning algorithms with various feature extractors, such as plain bag-of-words and bag-of-words using frequencies from the movie reviews corpus. Results showed that Naive Bayes and Maximum-Entropy algorithms performed best for classifying sentiments in comments, with around 45% accuracy in distinguishing between positive, negative, and neutral sentiments. For binary sentiment classification, Naive Bayes and Decision Trees performed equally well, achieving around 67% accuracy.

Further experiments focused on the use of the existing movie review corpus for training classifiers. The results revealed that the performance of classifiers trained on a subset of 1000 positive and 1000 negative sentences from the movie review corpus was lower than expected, achieving around 58% accuracy. This was attributed to the difference in the number of training and test sentences and the presence of many words in the training corpus that did not appear in the test corpora. Additionally, the study highlighted the challenges associated with using small, manually annotated corpora and the potential for errors in annotations.

The conclusion emphasizes that while Naive Bayes combined with a bag-of-words feature extraction using negated words was the most effective method, the study's findings should be interpreted cautiously due to several threats to validity, such as small corpus size, potential annotation errors, and topic-specific language. The authors suggest that further research with larger and more diverse corpora is needed to confirm their findings and explore the applicability of these techniques to other types of comments and communities.

The methodology involved several key steps: data acquisition from online sources, preprocessing to clean the text, feature extraction to identify sentiment-bearing words, and finally classification using the Naive Bayes algorithm. The preprocessing phase included steps like removing special symbols, expanding abbreviations, and stemming. Feature extraction focused on identifying terms with strong sentiment orientation, using measures like term frequency-inverse document frequency (TF-IDF) and counting polar words. The Naive Bayes classifier then used these features to distinguish between positive and negative reviews.

The results of the experiments showed that the Naive Bayes classifier performed the best, achieving an accuracy of 81.45%. The Random Forest classifier followed with an accuracy of 78.65%, while the K-Nearest Neighbors algorithm lagged behind with 55.30% accuracy. These results highlight the effectiveness of the Naive Bayes algorithm in handling the sentiment analysis task for movie reviews. Additionally, the paper emphasizes the need for further experimentation with different algorithms or hybrid methods to potentially increase the accuracy of sentiment classification.

The paper concludes that accurate sentiment analysis can significantly benefit various domains by enabling the development of intelligent systems that provide comprehensive reviews of movies, products, and services. These systems can help users make informed decisions without the need to read through individual reviews. The research also acknowledges that testing a broader range of algorithms and incorporating more diverse datasets could enhance the robustness and accuracy of sentiment analysis models.

The paper "Sentiment Analysis on Nepali Movie Reviews using Machine Learning" by Ashok Kumar Pant and Abhimanu Yadav presents a study on using machine learning techniques to classify the sentiment of Nepali movie reviews as positive or negative. The researchers created a dataset of 500 movie reviews, with an equal split between positive and negative sentiments. They employed various natural language processing (NLP) techniques for preprocessing the text, such as removing noise, handling negations, and performing part-of-speech tagging. The core of their approach was using a Naive Bayes classifier to perform the sentiment classification, achieving precision, recall, and F-score metrics of approximately 79%.

The methodology involved several key steps: data acquisition from online sources, preprocessing to clean the text, feature extraction to identify sentiment-bearing words, and finally classification using the Naive Bayes algorithm. The preprocessing phase included steps like removing special symbols, expanding abbreviations, and stemming. Feature extraction focused on identifying terms with strong sentiment orientation, using measures like term frequency-inverse document frequency (TF-IDF) and counting polar words. The Naive Bayes classifier then used these features to distinguish between positive and negative reviews.

The results of the study showed that the Naive Bayes classifier could effectively classify the sentiment of Nepali movie reviews with relatively high accuracy. The authors concluded that their approach could be beneficial for various applications, such as marketing and product selection, where understanding public sentiment is crucial. They also suggested that future work could explore using larger datasets and other machine learning techniques like support vector machines (SVM) and neural networks to potentially improve classification performance.

# 

# Chapter 3:METHODOLOGY

## INTRODUCTION

Sentiment analysis is a natural language processing task that deals with the identification and extraction of subjective information or the opinion from the given text documents.It determines the attitude or the contextual polarity of the

document. Sentiment analysis carries the basic task of classification of the expressed opinion in a document into ”positive”,”negative”, or ”neutral” class. Beyond polarity, sentiment classification can be used with the emotional states such as”happy”, ”sad”, and ”angry.”

Recently, sentiment analysis has taken great interest as the rise of social media such as blogs and social networks.With the proliferation of reviews, ratings, recommendations and other forms of online expression, online opinion has turned into a kind of virtual currency for businesses looking to market

their products, identify new opportunities and manage their reputations. It can also be used to make decisions to purchase or to use services by individuals. Ad markets can use sentiment analysis to place ads on praised sites. And,sentiment analysis can also be used for opinion retrievals .

# 

# 3.2 Block Diagram

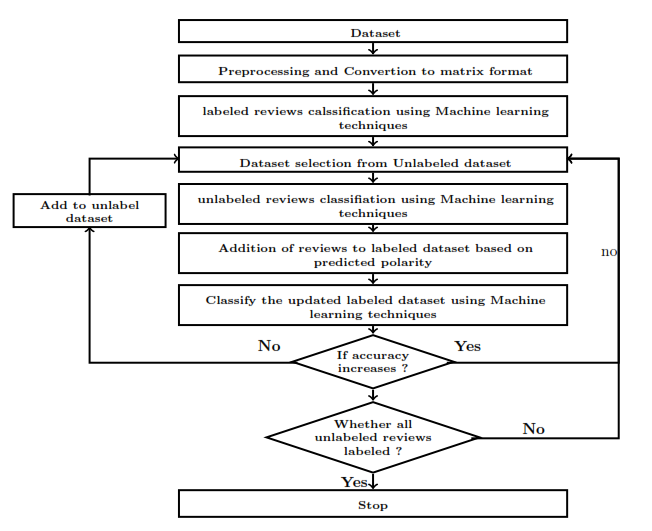


Fig 3.2: Block Diagram

# 3.3 Methodology

## Step 1: Data Collection and Preprocessing

Pre-processing the data is the process of cleaning and preparing the text for feature extraction and classification. In this stage, noise and uninformative text are removed from the input text document . Keeping those words makes the

dimensionality of the problem is high and hence the classification is more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis.

## Step 2: Normalization

In order to make full use of the unique sentiment resource information in text sentiment analysis tasks, this paper extracts four kinds of features: word features, part of speech features, position features, and dependency syntax features.The word features are combined with the other three kinds of features separately to form three new combined features which are input into a multi-channel convolutional neural network. Then, the features extracted from different channels are concatenated and input to the multi-head attention layer. Finally the sentiment classification results are obtained through the sentiment classification layer.

## Step 3: Tokenization

Features in the context of opinion mining are the words, terms or phrases that strongly express the opinion as positive or negative. This means that they have a higher impact on the orientation of the text than other words in the same text.

There are several methods that are used in feature selection, where some are syntactic, based on the syntactic position of the word such as objectives, and some are univariate, based on each feature’s relation to a specific category, and some are multivariate using genetic algorithms and decision trees based on features subsets.

## Step 4: Building a Sentiment Analysis Model

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or

that were not recognized as class examples (false negatives).

A. Precision

Precision (also called positive predictive value) is the number of correctly classified positive examples divided by the number of examples labeled by the system as positive.

B. Recall

Recall(also called sensitivity)is the number of correctly classified positive examples divided by the number of positive examples in the test dataset.

## 

## STEP 4 :TRAINING AND TESTING DATASET

The testing dataset is a smaller subset of the data that is kept separate from the training data. This set is used to evaluate the model's performance and generalization ability. By testing on this unseen data, we can gauge how well the model will perform in real-world scenarios. The testing dataset should be representative of the overall data distribution to ensure a fair evaluation.

The dataset is typically split using a certain ratio, commonly 70-80% for training and 20-30% for testing. This can be done using random sampling to ensure that both sets are representative of the entire dataset.

# 3. MACHINE LEARNING METHODS

## 3.1 Naïve Bayes

It is a technique based on Bayes’ Theorem. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is easy to build and particularly useful for very large datasets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

P C|X = P(X|C) ∗ P(C)/P(X) (1)

P(C|X) is posterior probability of class C

P(C) is prior probability of class C

P(X|C) is the probability of a predictor given the class.

P(X) is the prior probability of the predictor.

## 3.2 K- Nearest Neighbour

K-NN is the simplest of all machine learning algorithms. The principle behind this method is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a

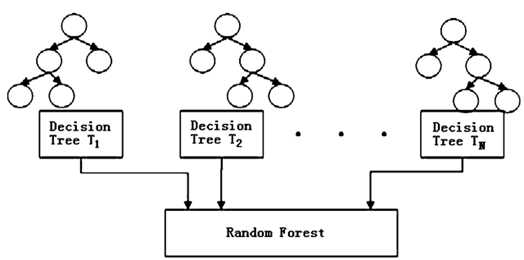
user-defined constant or vary based on the local density of points. The distance can be any metric measure. Standard Euclidean distance is the most common choice for calculating the distance between two points. The Nearest Neighbours

have been successful in a large number of classification and regression problems, including handwritten digits or satellite image processing and so on.

## 3.3 Random Forest

Random Forests are the learning method for classification and regression. It constructs a number of decision trees at training time. To classify a new case it sends the new case to each of the trees. Each tree performs classification and outputs a class. The output class is chosen based on majority voting that is

the maximum number of similar classes generated by various trees is considered as the output of the Random Forest.



## Fig 3.3 Random Forest Algorithm

# 

# 

# Chapter 4:RESULT AND ANALYSIS

The continuous collection and analysis of text feedback can significantly enhance project outcomes. Each day, feedback is gathered from various sources such as user interactions, through <https://rms.aarohasoft.com/feedback> or any other sources.This feedback is promptly categorized into key areas like usability, functionality, performance, and customer satisfaction. For instance, user feedback might reveal recurring issues with the application’s login process, indicating a critical area needing immediate attention.

Quantitative analysis is conducted to determine the sentiment of the feedback, with tools identifying whether the tone is positive, negative, or neutral.Quantitative analysis is conducted to determine the sentiment of the feedback, with tools identifying whether the tone is positive, negative, or neutral. For example, a daily sentiment analysis might show that 70% of feedback is positive, while 20% is negative and 10% neutral, highlighting general user satisfaction but also pinpointing areas of discontent. Qualitative analysis further deepens understanding by examining the specifics of the negative feedback. If users frequently mention app crashes during a specific function, this identifies a clear and actionable area for the development team.

The insights derived from daily feedback analysis are then used to prioritize tasks. High-impact issues like app crashes are addressed immediately, while lower-priority suggestions, such as aesthetic improvements, are scheduled for future updates. This ongoing process ensures that the project adapts rapidly to user needs and maintains a high level of quality and user satisfaction.

This continuous loop of gathering, analyzing, and acting on feedback not only improves the project incrementally but also fosters a user-centric approach, ultimately leading to a product that closely aligns with user expectations and experiences.

## 

## 

## Fig 4.1: Classification precision of k-NN, Naïve Bayes,andDecision Tree

## 

## 

## 

## 

## 

## Fig 4.2: Graph of Root Mean Squared Error of Algorithm

## 

## 

## 

## 

## 

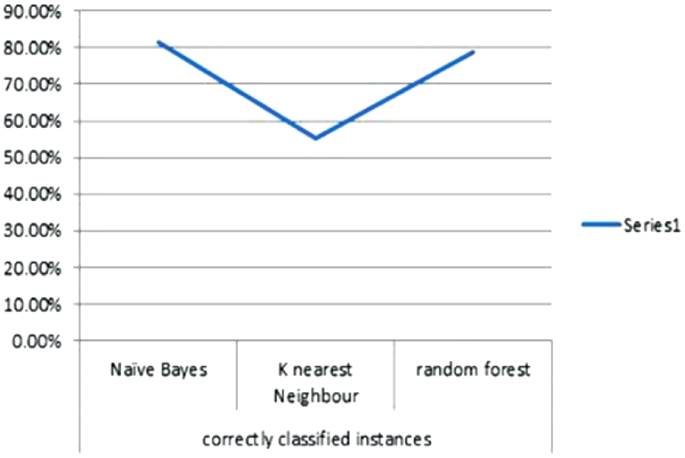
## 

## 

## 

## 

## 



## Fig 4.3: Graph of Accuracy of various Algorithms

## 

## The algorithms performed were Naïve Bayes, K-Nearest Neighbour, Random Forest. The best results were given by the Naïve Bayes classifier. The Naïve Bayeclassifier achieved 81.45% accuracy, Random Forest classifier we achieved 78.65% accuracy, K-Nearest Neighbour classifier achieved 55.30% accuracy.As only few algorithms were tested , it is required to test other algorithms or create hybrid methods so that accuracy of the results can be increased. Finding the polarity of the reviews can help in various domains. Intelligent systems can be developed which can provide the users with comprehensive reviews of movies,products, services etc. without requiring the user to go through individual reviews, he can directly take decisions based on the results provided by the intelligent systems.

## 

## 

## 

## 

## 

## 

## CONCLUSION

Conducting a text feedback analysis offers valuable insights into customer perceptions, preferences, and pain points. By systematically evaluating feedback, patterns and themes can be identified, allowing businesses to address specific issues and enhance overall customer satisfaction. For example, if recurring complaints about customer service responsiveness are noted, targeted training and process improvements can be implemented to resolve these concerns. Additionally, positive feedback can highlight areas of success that can be leveraged for marketing and further improvement.Ultimately, this analysis not only helps in refining products and services but also fosters stronger customer relationships by showing a commitment to listening and responding to their needs.

## REFERENCES

[1] Sentiment Analysis of Movie Review Comments.Available at:

<https://www.scss.tcd.ie/Khurshid.Ahmad/Research/Sentiments/K_Teams_Buchraest/mvie%20review%20review.pdf>.

[2]**Sentiment Analysis of Movie Reviews using Machine Learning Techniques.**

**Available at:**

[**https://www.researchgate.net/publication/321843804\_Sentiment\_Analysis\_of\_Movie\_Reviews\_using\_Machine\_Learning\_Techniques**](https://www.researchgate.net/publication/321843804_Sentiment_Analysis_of_Movie_Reviews_using_Machine_Learning_Techniques)**.**

**[3]** Sentiment Analysis on Nepali Movie Reviews using Machine Learning.Available at:

<https://www.academia.edu/12075554/Sentiment_Analysis_on_Nepali_Movie_Reviews_using_Machine_Learning>.

[4]**Short Text Sentiment Analysis Based on Multi-Channel CNN With Multi-Head Attention Mechanism. Available at:**

<https://ieeexplore.ieee.org/document/9335574>.