**MAKERERE UNIVERSITY**

**SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY**

**CONCEPT PAPER FOR RECESS FINAL PROJECT**

A Concept Paper submitted to the School of Computing and Informatics Technology

For the Study Leading to a Project Proposal in Partial Fulfillment of the

Requirements for the Award of the Degree of Bachelor of

Science in Software Engineering of Makerere University.

|  |  |  |
| --- | --- | --- |
| NAME | REG NO | STD. NO |
| KYAKA HERMAN CEASER | **15/U/6851/PS** | **215011936** |
| MALITI EMMANUEL | **17/U/5926/EVE** | **217006711** |
| KABWAMA RONALD | **15/U/5398/EVE** | **215018454** |
| KAIJUKA DONALD | **13/U/5658/PS** | **213004661** |

**Table Of Contents**

Contents

[**1** **Introduction** 3](#_Toc14360517)

[**1.1** **Title** 3](#_Toc14360518)

[**1.2** **Background.** 3](#_Toc14360519)

[**1.3** **Problem Statement.** 3](#_Toc14360520)

[**1.4** **Aim and Objectives** 3](#_Toc14360521)

[1.4.1 Aim or general objective 3](#_Toc14360522)

[1.4.2 Specific Objectives 3](#_Toc14360523)

[**2** **Methodology (Data Pipeline)** 3](#_Toc14360524)

[**2.1** **Data collection** 3](#_Toc14360525)

[**2.2** **Data Analysis** 4](#_Toc14360526)

[**2.3** **Data visualizations and Graphical Analysis.** 4](#_Toc14360527)

[**2.4** **Predictive Modelling** 4](#_Toc14360528)

[**3** **References.** 4](#_Toc14360529)

# **Introduction**

## **Title**

Review prediction from text comments

**Definitions.**

A review is a formal assessment of something with the intention of instituting change if necessary. Knowledge of how the commenting community reacts to a specified topic.

## **Background.**

Data set: Movie Reviews, <https://www.kaggle.com/nltkdata/movie-review>.

This data was first used in Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of the ACL, 2004.

The Data set contains 1000 positive and 1000 negative reviews classified under a column called Tag. The reviews are for comments got from Movie websites which are identified by the html\_id column and are given by different users on those websites who are identified by the sent\_id. The comments themselves are grouped under the text column.

## **Problem Statement.**

There is a lot of time wasted in reading individual comments to award the review to the subject topic. This however can be automated by involving machine learning to generate the review.

## **Aim and Objectives**

### Aim or general objective

To Develop a Prediction Model that uses the existing datasets to predict whether a review is positive or Negative.

### Specific Objectives

* To collect and analyze datasets to be used in training our prediction model.
* To Identify the variables that are to be used in rating subjects and be able to predict the likelihood of good or bad review basing on those variables.
* To use the collected data to visualize the relationships between the chosen variables.
* To design a prediction algorithm that will be used to predict review.
* To implement the algorithm using Machine learning tools like Natural Language Processing Tool Kit, Naïve Bayes classifier and the sample datasets chosen.
* To test the accuracy of the implemented Predictive model.

# **Methodology (Data Pipeline)**

## **Data collection**

This includes all the techniques used to collect data.

We shall use Kaggle, a dataset repository on the web to get sample datasets that satisfy the variables listed and download them to our local server for data mangling.

## **Data Analysis**

We shall use Pandas, a python Library to read the dataset into a pandas dataframe.

We shall them Do Exploratory analysis to determine the features of our Data sets by using methods on Pandas dataframe like describe(), info() and the head() method to get an insight into our dataset.

After that, we shall do feature Engineering on our data by creating new features in our dataframe to help us understand the dataset. Feature columns to be created include the Text\_length column that describes the length of each comment in the text column.

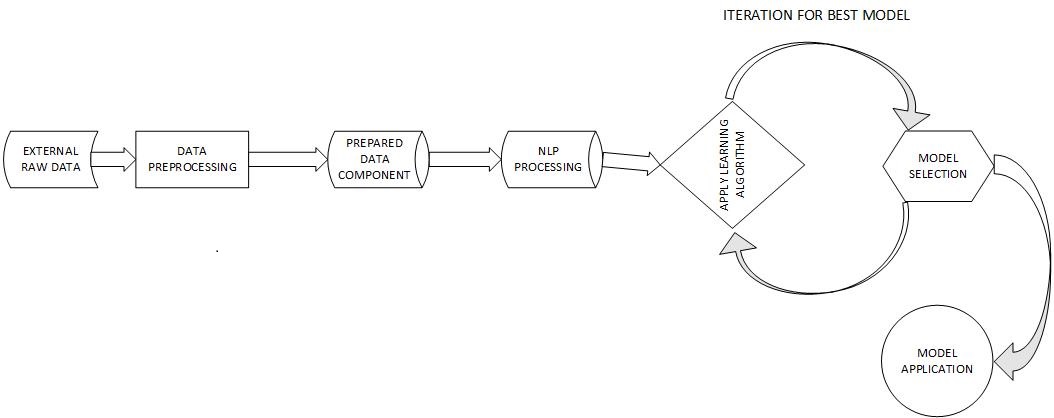
## **Data visualizations and Graphical Analysis.**

We shall then use a histogram, pair plots to visualize the relationship between textlength and the review, which is either positive or negative.

## **Predictive Modelling**

Before we Train Our Model, we shall have to preprocess the Text data by Tokenizing it and removing any punctuations and stopwords such that data is ready to be fit into our Machine Learning Model.

We shall then use Naïve Bayes classification together with the CountVectorizer to train our model and come up with a pipeline. The pipeline will get the vector count of the comments, pass them through the Term frequency Inverse document Transformer and then use MultinomialNB classification to predict the tag of the review.



# **References.**

1. <https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa>
2. <https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook>
3. <https://www.eurekalert.org/pub_releases/2019-06/uot-sra062119.php>