Twitter Dataset Analysis
Group 17

Name	Student Number (SN)	Contribution
Karan Goel	7836685	14.28%
Alvin Jose	8066358	14.28%
Ashutosh Bhosale	7795786	14.28%
Banin Sensha Shreshta	8447196	14.28%
Gaurav Adarsh Santosh	7032663	14.28%
Lino Thankachan	7926017	14.28%
Rishab Manokaran	7863974	14.28%

CSCI946 Big Data Analytics Assignment 3
October 28, 2024

1 Introduction

In this assignment, we analyzed the pretrained features extracted from a large image recognition model and developed a classifier to achieve high classification accuracy on two provided Imagenet test datasets.

To accomplish this, we will follow a structured approach outlined in Tasks 1-4.

Task 1 involves designing a comprehensive big data analytics project while adhering to the principles of the Big Data Analytics Lifecycle.

Task 2 will focus on detailing the methodology used in our analysis and classification processes.

Task 4 will summarize our conclusions and the insights gained from completing this assignment.

2 Task 1: Data Analysis Design

In this section on the Data Analysis process

2.1 Business Domain

Objective

The primary aim of this assignment is to build a classifier that will use pretrained features and give good accuracy on two provided datasets, and compare their differences.

Dataset

The dataset used for the Assignment are created from imagenet and imagenet

Resources & Goals

Data Source: The dataset consists of Twitter user profiles, which include tweets, descriptions, link color, sidebar color, and other metadata.
Primary Task: The task is to classify the gender of Twitter profiles into two categories: human or non-human.
Key Questions:

- How well do the words in tweets and profiles predict the user as human or non-human?
- What are the specific words that strongly predict human or non-human profiles?
- How well do other factors (like link color, tweet count, retweet count, favourite number) predict whether a profile is human or non-human?

2.2 Framing the Problem & Initial Hypotheses

Problem Type

This is a supervised learning problem where the task is to predict a categorical variable (gender: male, female, or brand).

Hypotheses

- **Null Hypothesis (H0)**: Words in tweets and profiles do not have a significant effect on predicting whether a user is human or non-human (i.e., the predictive power is random or weak).
- **Alternative Hypothesis (H1)**: Words in tweets and profiles significantly affect the prediction of whether a user is human or non-human (i.e., they provide strong predictive power).
- **Additional Hypotheses**:
  - **H0**: Other factors (such as link color, tweet count, retweet count, favourite number) are not good predictors of whether a user is human or non-human.
  - **H1**: Other factors can strongly predict whether a user is human or non-human.

2.3 Data Preparation

In the data preparation process, several important preprocessing steps were applied to ensure consistency and effective representation of text data. These steps included:

- **Stemming & Lemmatization**: Words were reduced to their base or root forms to avoid variations in word forms affecting the analysis.
- **Lowercasing**: Text was converted to lowercase to eliminate discrepancies between uppercase and lowercase letters.
- **Tokenization**: The text was split into individual words or tokens.
- **Bag of Words**: A simple yet effective method for text representation was used to analyze the frequency of words in the dataset.
- **Word2Vec Model**: This model was applied to generate word embeddings for further analysis.

2.4 Visualizing Data

During the data visualization process, several techniques were employed to gain insights from the text data. A **word cloud** was generated to display the most frequent words associated with different gender categories (male, female, and brand). Additionally, **word distance and similarity** techniques, such as cosine similarity, were applied to understand which words or phrases were more closely related to each gender category. These visualization methods provided an intuitive understanding of the text data.

2.5 Model Selection

In the model selection phase, various machine learning models were explored to address classification, clustering, and regression tasks:

Classification Models

Models such as **K-Nearest Neighbors (KNN)**, **Support Vector Machines (SVM)**, **Decision Trees**, and **Random Forests** were considered. These models are effective for categorizing data based on the patterns and relationships within the dataset.

Clustering Models

For clustering tasks, models like **K-Means**, **DBSCAN**, and **Self-Organizing Maps (SOM)** were employed. Clustering models are helpful in discovering natural groupings in the data, allowing for an unsupervised learning approach to identify hidden patterns without predefined labels.

Regression Models

Both **Linear Regression** and **Logistic Regression** were used. Logistic Regression, in particular, was useful for binary classification tasks such as predicting whether a profile is human or non-human.

Neural Networks

Neural networks were also utilized to build deep learning models for more advanced classification tasks.

2.6 Training and Testing

Cross-Validation

To prevent overfitting and assess the generalization of the models, **k-fold cross-validation** was applied. This technique divides the dataset into multiple folds, allowing each fold to be used for both training and validation, resulting in a more robust evaluation of model performance.

Grid Search

Grid search was employed to optimize hyperparameters by systematically exploring different model configurations. This approach enabled the selection of the best combination of parameters for optimal model performance.

2.7 Final Deliverables

For evaluating the models, the following metrics were used:

- **Confusion Matrix**: This was used to visualize the performance of each classification model.
- **Accuracy**: The percentage of correctly classified instances.
- **Precision**: The proportion of true positives out of all positive predictions.
- **Recall**: The proportion of true positives out of all actual positives.
- **Loss Curves**: For neural networks, loss curves were used to visualize the training process and detect overfitting or underfitting.