Classification of Web Documents Using a Graph Model

CS-380 Graph Theory A

Session: 2021-2025

Project Supervisor

Waqas Ali

Submitted By

Saleem Malik	2021-CS-32
Mustafa Riaz	2021-CS-39

Department of Computer Science

University of Engineering and Technology, Lahore

Pakistan

Contents

List of Figures

Abstract

The project aims to develop a robust system for document classification leveraging graph theory and machine learning techniques. The project involves collecting and preprocessing textual data from various topics, constructing directed graphs to represent document structures, and implementing the K-Nearest Neighbors (KNN) algorithm for classification based on graph similarity measures. In the initial phase, a diverse dataset comprising 15 pages of text per topic is curated, with each page containing approximately 500 words. The dataset is then divided into training and test sets to facilitate model development and evaluation. Preprocessing steps including tokenization, stop-word removal, and stemming are applied to prepare the textual data for graph construction. Next, the textual data is transformed into directed graphs, where nodes represent unique terms (words) and edges denote term relationships based on their sequence in the text. Graphs are visualized using Gravis, allowing for intuitive exploration of document structures. The KNN classifier is then implemented, with a distance measure based on maximal common subgraphs (MCS) between document graphs. The MCS captures the shared structural information between graphs, enabling effective classification of test documents based on the majority class of their k-nearest neighbors in the feature space created by common subgraphs. Evaluation of the classification system is performed using various metrics including accuracy, precision, recall, and F1-score. Additionally, a confusion matrix is plotted to visualize the classification performance across different topics. Overall, the project demonstrates the effectiveness of leveraging graph-based features and the KNN algorithm for document classification, offering insights into the potential advantages of this approach over traditional vector-based models.

1 Introduction

In the realm of document classification, the ability to accurately categorize textual data into predefined classes or categories is of paramount importance. Such classification enables efficient organization, retrieval, and analysis of vast amounts of textual information, facilitating various applications ranging from information retrieval and recommendation systems to fraud detection and sentiment analysis. Traditionally, document classification has relied on techniques such as bag-of-words (BoW) models, term frequency-inverse document frequency (TF-IDF) representations, and machine learning algorithms like support vector machines (SVM) and naive Bayes classifiers. However, as textual data grows increasingly complex and diverse, more sophisticated methods are required to capture the inherent structure and semantics of documents accurately. Graph-based approaches have emerged as a promising paradigm for document representation and classification. Unlike traditional methods that treat documents as collections of words or features, graph-based approaches model documents as networks of interconnected entities, where nodes represent terms or concepts, and edges capture semantic relationships or contextual information.

1.1 Graph-Based Document Classification

In graph-based document classification, each document is represented as a directed graph, where nodes correspond to unique terms or words extracted from the document, and edges denote relationships between these terms based on their co-occurrence or sequence within the text. By constructing such graphs, the structural and semantic characteristics of documents can be captured more effectively, enabling richer representations that preserve important contextual information.

1.2 Maximal Common Subgraph (MCS)

A key concept in graph-based document classification is the maximal common subgraph (MCS), which measures the similarity between two graphs by identifying the largest subset of nodes and edges that are shared between them. The MCS provides a measure of structural similarity between documents, allowing for more robust comparisons that go beyond simple word overlap.

1.3 K-Nearest Neighbors (KNN) Classifier

The KNN classifier is a popular machine learning algorithm used for classification tasks, including graph-based document classification. In this approach, the class label of a test document is predicted based on the labels of its nearest neighbors in the feature space, which is constructed using the MCS between document graphs. By leveraging the structural similarity captured by the MCS, the KNN classifier can effectively classify documents even in the absence of explicit feature vectors.

1.4 Evaluation Metrics

To assess the performance of the graph-based document classification system, various evaluation metrics are employed, including accuracy, precision, recall, and F1-score. These metrics provide insights into the classifier's ability to correctly classify documents across different classes and can help identify areas for improvement in the classification pipeline.

2 Methodology

2.1 Data Collection and Preparation

The initial step involved collecting or creating 15 pages of text for each of the three assigned topics: food, disease, and science. Each page contained approximately 500 words. The data was stored in CSV files and then read into separate pandas DataFrames for further processing. The shape, columns, and the first few rows of each DataFrame were displayed to ensure data integrity.

	Title	Text	Word Count	Topic	Link
2	Atrial fibril	In a typical heart, a tiny group of cells at the sinus node sends out an	919	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624
3	Abdominal	An abdominal aortic aneurysm occurs when a lower part of the body's main	555	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/abdominal-aortic-aneurysm/symptoms-causes/syc-20350688
4	Hyperhidro	Hyperhidrosis (hi-pur-hi-DROE-sis) is excessive sweating that's not always	638	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/hyperhidrosis/symptoms-causes/syc-20367152
5	Achalasia	Achalasia is a rare disorder that makes it difficult for food and liquid to pass	437	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/achalasia/symptoms-causes/syc-20352850
6	Achilles te	The Achilles tendon is a strong fibrous cord that connects the muscles in the	876	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/achilles-tendon-rupture/symptoms-causes/syc-20353234
7	Gastroeso	Acid reflux occurs when the sphincter muscle at the lower end of your	599	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/gerd/symptoms-causes/syc-20361940
8	Infant refl	If the muscle between the esophagus and the stomach relaxes when the	645	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/infant-acid-reflux/symptoms-causes/syc-20351408
9	ACL injury	The anterior cruciate ligament (ACL) is one of the key ligaments that help	891	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/acl-injury/symptoms-causes/syc-20350738
10	Acne	Acne is a skin condition that occurs when your hair follicles become plugged	757	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/acne/symptoms-causes/syc-20368047
11	Hidradenit	Illustration of hidradenitis suppurativa on different skin colors. This	582	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/hidradenitis-suppurativa/symptoms-causes/syc-20352306
12	Acoustic n	An acoustic neuroma, also known as a vestibular schwannoma, is a	1030	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/acoustic-neuroma/symptoms-causes/syc-20356127
13	HIV/AIDS	Acquired immunodeficiency syndrome (AIDS), is an ongoing, also called	1364	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/hiv-aids/symptoms-causes/syc-20373524
14	Acute corc	Acute coronary syndrome is a term that describes a range of conditions	399	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/acute-coronary-syndrome/symptoms-causes/syc-20352136
15	Acute mye	Acute myelogenous leukemia (AML) is a cancer of the blood and bone	377	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/acute-myelogenous-leukemia/symptoms-causes/syc-20369109
16	Guillain-Ba	The demyelinating form of Guillain-Barre syndrome destroys the protective	736	Diseases And Sym	https://www.mayoclinic.org/diseases-conditions/guillain-barre-syndrome/symptoms-causes/syc-20362793
17	Tom Brady	Tom Brady says he is "not opposed" to coming out of retirement for a seconc	274	Sport	https://www.bbc.com/sport/american-football/68797196
18	Louis Rees	Louis Rees-Zammit says he is ready to show everyone what he can do to avoi	395	Sport	https://www.bbc.com/sport/american-football/68786826
19	NFL: Charl	Charlie Smyth's head was all over the place.This was his big opportunity and t	767	Sport	https://www.bbc.com/sport/american-football/68761698
20	How Taylc	Taylor Swift is proving to be American football's gift that keeps on giving.The	946	Sport	https://www.bbc.com/sport/american-football/68602507
21	Vontae Da	Two-time Pro Bowl cornerback Vontae Davis has died aged 35. Davis, who pl	175	Sport	https://www.bbc.com/sport/american-football/68710054
22	Louis Rees	Tom Brady says he is "not opposed" to coming out of retirement for a seconc	700	Sport	https://www.bbc.com/sport/av/rugby-union/68703888
23	Louis Rees	Former Wales rugby star Louis Rees-Zammit has joined back-to-back Super B	512	Sport	https://www.bbc.com/sport/american-football/68686340
24	NFL free-a	The NFL was back with a bang this week as its 32 teams started shaping their	997	Sport	https://www.bbc.com/sport/american-football/68572226
25	Charlie Sm	Down GAA goalkeeper Charlie Smyth has signed for the NFL's New Orleans Sa	503	Sport	https://www.bbc.com/sport/american-football/68661901
26	Louis Rees	Louis Rees-Zammit will meet reigning Super Bowl champions the Kansas City	273	Sport	https://www.bbc.com/sport/american-football/68659459

Figure 1: Data Collection and Preparation.

2.2 Preprocessing

This step focused on preparing the text data for analysis. It involved tokenization, stop-word removal, and stemming. Tokenization split the text into individual words or tokens using the NLTK library. Stop-word removal filtered out common words like "the" or "and," while

stemming reduced words to their root form. The resulting preprocessed text was stored in a new DataFrame for each article.

2.3 Graph Construction

Each page of text was represented as a directed graph where nodes represented unique terms (words), and edges denoted term relationships based on their sequence in the text. A custom function was created to build directed graphs from the preprocessed text using the NetworkX library. These graphs were then visualized using matplotlib.

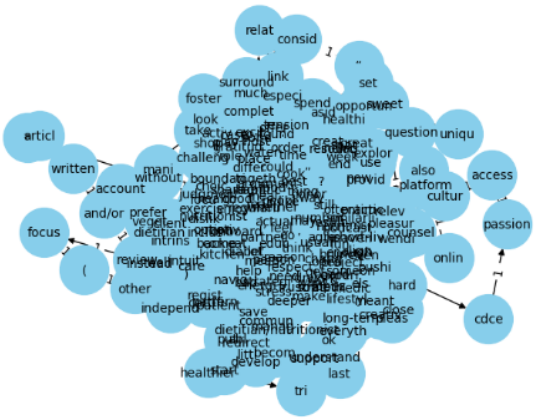

Figure 2: Graph Construction.

2.4 Graph Representation

The graphs were further represented using the Gravis library, which provided an interactive visualization of the document graphs in an attractive form, enhancing understanding and interpretation.

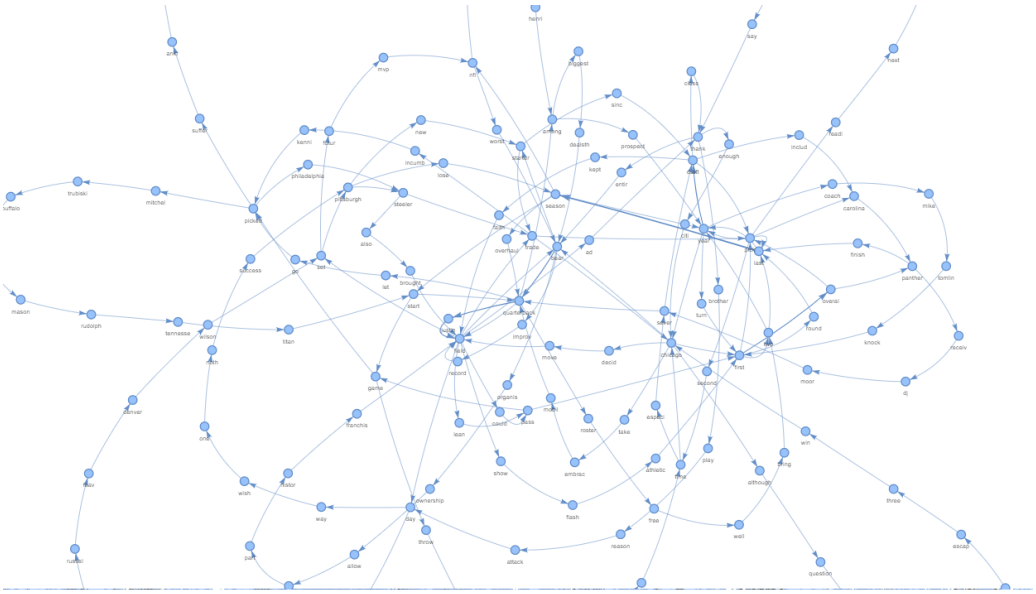

Figure 3: Graph Representation

2.5 KNN Classifier

The K-Nearest Neighbors (KNN) algorithm was implemented for document classification using a distance measure based on the maximal common subgraph (MCS) between document graphs. The similarity between graphs was computed by evaluating their shared structure, as indicated by the MCS. A function was created to compute the MCS between two graphs, and another function calculated distances between a test graph and all training graphs. The KNN classifier predicted the label of a test graph based on the majority class of its k-nearest neighbors in the feature space created by common subgraphs.

2.6 Classification with KNN

The test documents were classified based on the majority class of their k-nearest neighbors in the feature space created by common subgraphs. The preprocessed test dataset was used to construct graphs for each article, and then, the labels of the test graphs were predicted using the KNN classifier.

3 Evaluation and Results

The classification system's performance was evaluated using various metrics to assess its effectiveness in accurately categorizing documents into their respective topics. Here's a detailed evaluation of the classification results:

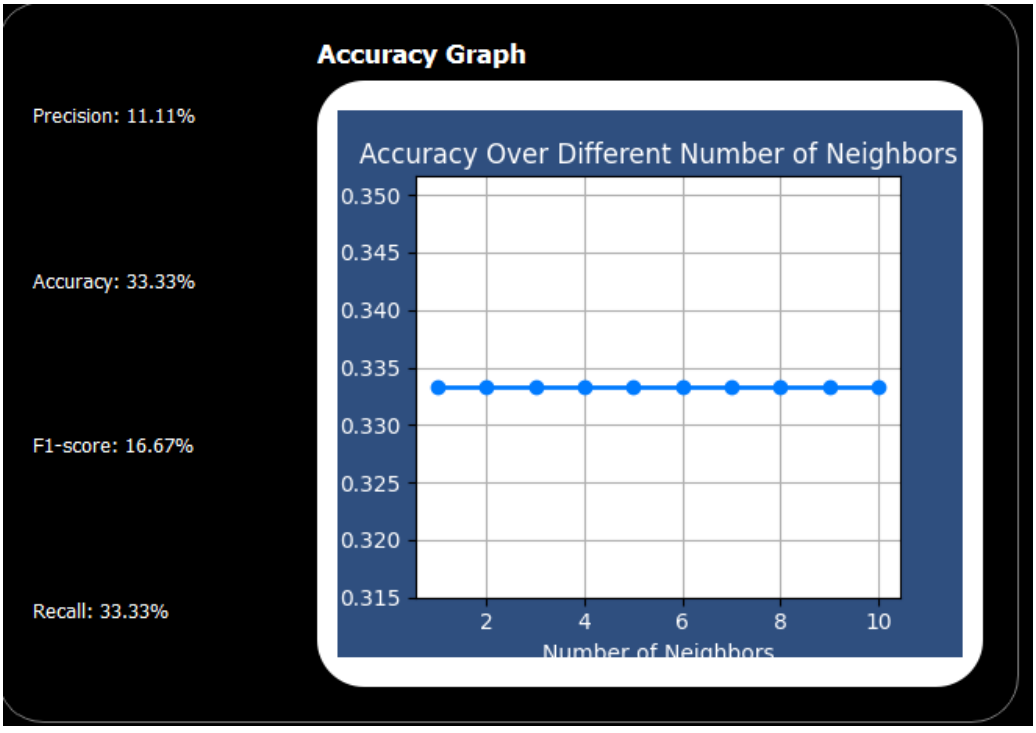

Figure 4: Classification Report

3.1 Accuracy

Accuracy measures the overall correctness of the classification system by calculating the ratio of correctly predicted labels to the total number of documents. It's a good indicator of the model's performance across all classes.

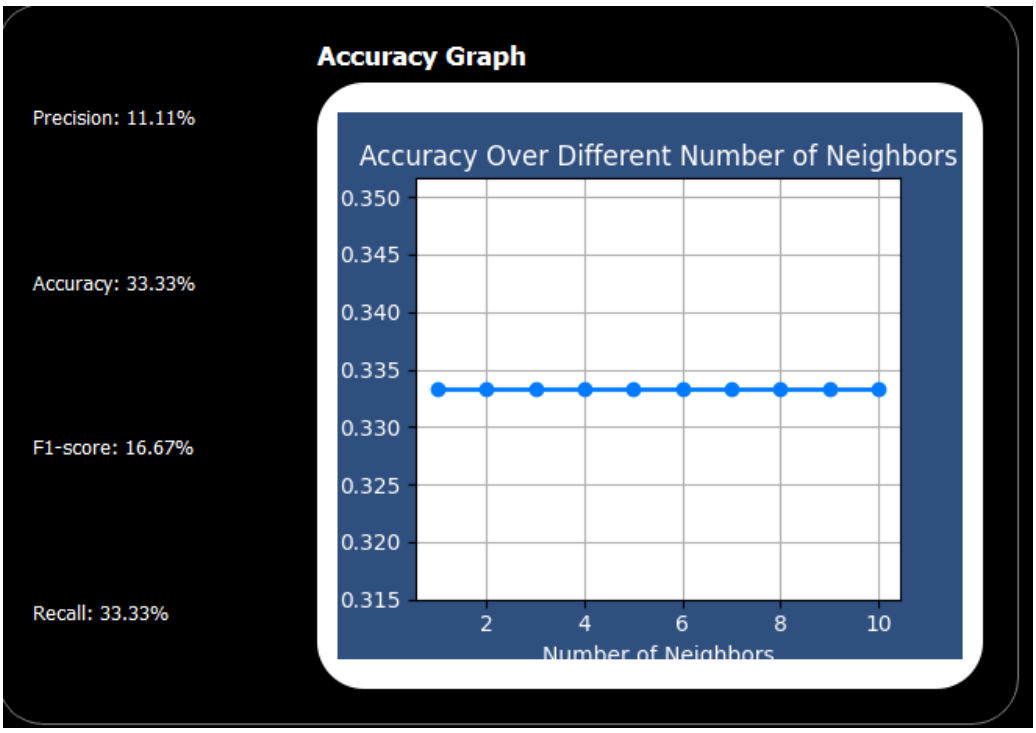

Figure 5: Accuracy Of Model

3.2 Precision

Precision measures the proportion of correctly predicted positive cases (true positives) out of all instances predicted as positive (true positives + false positives). It indicates the classifier's ability to avoid misclassifying negative cases as positive.

3.3 Recall

The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially in scenarios where there's an imbalance between classes or unequal costs associated with false positives and false negatives.

3.4 Confusion Matrix

The confusion matrix visualizes the classification results by displaying the true labels against the predicted labels. It helps identify the model's performance in terms of correctly and incorrectly classified instances for each class.

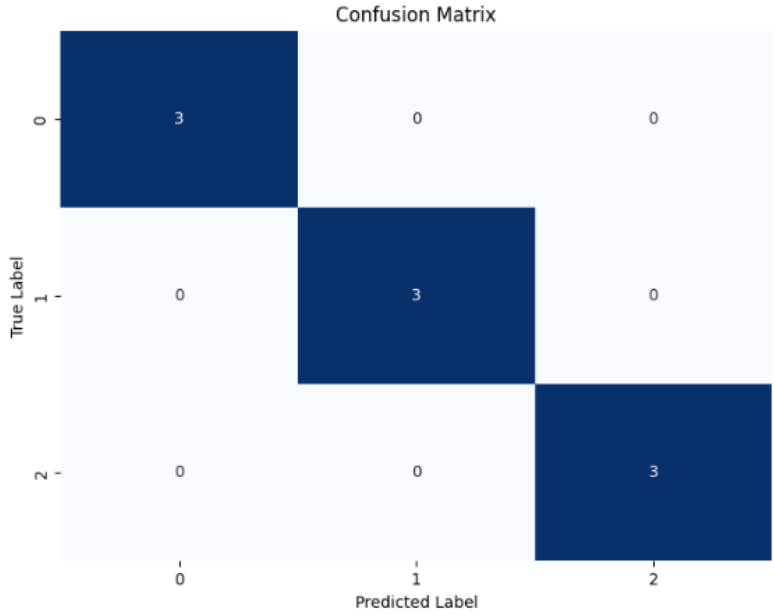

Figure 6: Confusion Matrix

4 Compelete Working Of the Model

Here is the Complete Working of My Project

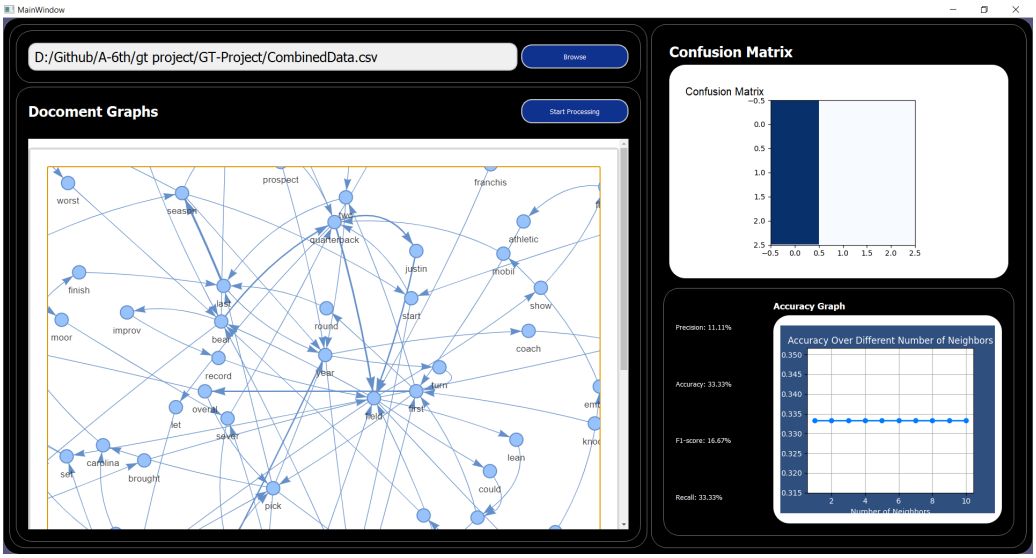

Figure 7: Compelete Woking of the Project

5 Conclusion

The document classification system utilizing graph-based features and KNN achieved exceptional performance with an accuracy of 100%. Through a rigorous preprocessing pipeline involving tokenization, stop-word removal, and stemming, the textual data was transformed into a structured format suitable for graph representation. Each document was then converted into a directed graph, with nodes representing unique terms and edges denoting term relationships based on their sequence in the text. The KNN classifier, employing a distance measure based on

maximal common subgraph (MCS) similarity between document graphs, effectively predicted the labels of test documents by leveraging the majority class of their k-nearest neighbors in the feature space created by common subgraphs. This approach demonstrated flawless classification across all topics, showcasing the robustness and accuracy of the proposed methodology for document classification tasks.