Coventry University | FACULTY OF ENGINEERING, ENVIRONMENT & COMPUTING

(7151CEM)

Computing Individual Research Project

Student Name: Vedan Yadav Gokul

SID: 11479310

Project Title: Classification of Wikipedia Articles

I can confirm that all work submitted is my own: Yes

Section A – Ethics Application

✓ **I submitted my ethics application and my application has been approved. I include my ethics certificate in the appendix as evidence.**

☐ **I submitted my ethics application and my application is currently under review.**

☐ **I have not submitted my ethics application.**

Section B – Project Proposal

1. Research Question, Problem Statement or Topic for Investigation

The aim of the project is to compare the performance of different Machine Learning and Deep Learning techniques, reason their performance and identify the best performing techniques on Wikipedia article classification. Once the best performing techniques are identified, a weighted combination of these techniques is used to try to further increase the performance on the data.

The idea of this project is about classification of Wikipedia articles into their respective classes. Depending on the information present in the form of text in the particular article, it is classified and placed into one of the predetermined classes.

The data going to be used in the project is publicly available Wikipedia data. As this project is aimed to categorize Wikipedia articles by using text classification, there are not many suitable datasets that are ready and publicly available that fit the exact scale of data and diversity of classes needed for this project. Hence, I will need to gather data manually with the help of MediaWiki API. Wikipedia contains 41 main categories that correspond to one of Wikipedia's major topic classifications. Out of these 41 categories, ten categories will be chosen, keeping in mind maximum diversity between the categories. 500 articles from each category are extracted and a balanced dataset with around 5000 total samples containing 10 classes will be made. This customized dataset will be used in our project for further purposes.

The dataset is then cleaned to match the needs of each individual model separately before applying the model on the dataset. The data set will be divided into training and testing sets with stratified split so that the model will generalize properly on the entire data set. Each individual model will be tuned based on the metrics evaluated on the validation set which is divided as part of the training set. As per the optimal parameters obtained, the model will be tested on the test set and the performance of the model will be evaluated using test set. After this process, limitations and improvements of each model will be identified and reported. At the end, the performance of all the models are compiled together and then compared to give us an idea of their relative performance.

2. Intended user or group of users and their requirements

According to Wikipedia, it currently contains over 6 million articles in the English language alone and more than 55 million when we consider all the languages. The categorization of Wikipedia articles is done manually one by one by editors, which is a time consuming process and prone to human errors. And the fact that Wikipedia is a community driven website and the information is edited by its users, it poses different challenges in terms of analyzing the data, one of which is classification. The data can lack structure, grammar and integrity.

There is a requirement for publicly available datasets in this domain which can used by future users and researchers for their needs. An understanding of the performance of different methods is necessary on not only Wikipedia data, but other community driven website data like github, reddit, etc. By creating a balanced and diverse dataset, this project can provide users with a much needed public dataset and also act as a guide for users who want to create a customized dataset of their own from data of a publicly available source. It will provide future readers with an estimation of the performance of different models and the reasons for the said performance on text classification of Wikipedia articles which can give an idea of how different algorithms behave on similar data extracted from community driven websites. The project also gives the users a brief look at how different segments of a Wikipedia article taken as an input can affect the final performance of an algorithm.

3. Systems requirements, project deliverables and final project outcome

The first deliverable of the process is the acquisition of the data and creation of the dataset. It is followed by a thorough literature review of the available methods of text classification and Wikipedia articles classification. Then based on the literature review, choosing the suitable methods which may give better results on our data. Application of these methods and the results of their performance are then recorded. Followed by trying to identify the reasons for the given performance of the algorithms. Identifying ways to improve the performance of the said models with possible modifications. At the end, a weighted combination of the best performing algorithms is applied on the data to further improve the accuracy and the weights are modified until they reach their optimal state. This outcome of this project will be the creation of a dataset, clear comparison of the performances of the methods on the dataset, results of weighted combination of algorithms, final best accuracy obtained and further methods to try in the future.

4. Primary Research Plan

The first task of this project is acquiring the data and creating the dataset required for our project. Then comes the application of simple models first like Logistic Regression to understand the complexity of the data set. After understanding the complexity of the dataset, performing a literature review of the existing models and choosing which may work the best in our case is the

next step. The data is cleaned to match the needs of each individual model separately before applying the model on the dataset. The data set will be divided into training and testing sets with stratified split so that the model will generalize properly on the entire data set. Each individual model will be tuned based on the metrics evaluated on the validation set which is divided as part of the training set. As per the optimal parameters obtained, the model will be tested on the test set and the performance of the model will be evaluated using test set. After this process, limitations and improvements of each model will be identified and reported. Any modifications which can lead to further improvement in accuracy are given a try and applied. The performance of all the models are compiled together and then compared to give us an idea of their relative performance. Application of weighted combination of the best performing methods and balancing the weights to get the best possible results, finishes the project. Considering there is approximately six weeks from the day of project proposal to project presentation, below is the suggested timeline for carrying out this project

Timeline:
Week 1: Acquiring the data and creating suitable dataset. Figuring out the complexity of the data.
Week 2: Application of basic text classification models and literature review to find out best suitable methods for the dataset.
Week 3: Data Cleaning and application of the chosen models
Week 4: Analyzing the performance of the models and identifying the reasons for the resulted performance. Discover areas of improvement and their application.
Week 5: Using weighted combination of the best performing algorithms to increase accuracy. Finding future ideas associated with the project.
Week 6: Compiling everything that is done in the weeks before and putting it on paper in the form of the project report.

5. Initial/Mini Literature Review

A major challenge for many analyses of Wikipedia dynamics e.g., imbalances in content quality, geographic differences in what content is popular, what types of articles attract more editor discussion is grouping the very diverse range of Wikipedia articles into coherent, consistent topics [1]. The methods used to tackle these challenges were Wikipedia's category network, WikiProjects, and external taxonomies. But these have limitations when it comes to tackling the language barrier and only apply to a small subset of articles, since Wikipedia has more than 300 languages these methods cant be used for a large dataset containing articles with different datasets .In this paper [1] the authors provide a solution to the language barrier of the Wikipedia articles by using a language agnostic approach. They use page to page link based network approach instead of the text based approach. A similar project idea to mine can be developed here, which is to identify all the network based approaches available and perform similar actions like in my project.

Different Machine Learning Techniques for text classification are explored in this paper [2]. SVM, Multiclass SVM, Naive Bayes and Random Forest algorithms are tested on the 20newsdataset from sklearn datasets. The performance of the different algorithms are given in the results. This is similar to what we are doing in my project. The Machine Learning Techniques and the approach

is similar except that the dataset is different and not community edited. The data is perfectly pre classified unlike in our case where the classification process is complicated and is done manually one by one by editors.

This paper [3] has a survey of all the text classification models which include RNN based models, CNN based models, Capsule Neural Networks based models, Attention based models, GNN based models and Hybrid models. The paper describes tasks of Sentiment Analysis, Topic Analysis, News Categorization, Question Answering and Natural Language Inference as text classification tasks which can be solved using the above mentioned models. All the models mentioned in this paper can be studied and possibly implemented on our dataset to check the relative performances of the different models.

6. Bibliography

1. Isaac Johnson, Martin Gerlach, and Diego Sáez-Trumper. 2021. Language-agnostic Topic Classification for Wikipedia. Companion Proceedings of the Web Conference 2021. Association for Computing Machinery, New York, NY, USA, 594–601. DOI:https://doi.org/10.1145/3442442.3452347

2. Athanasios Tzimourtas, Spyros Bakalakos, Panagiota Tselenti, and Athanasios Voulodimos. 2021. An exploration on text classification using machine learning techniques. In 25th Pan-Hellenic Conference on Informatics (PCI 2021). Association for Computing Machinery, New York, NY, USA, 247–249. DOI:https://doi.org/10.1145/3503823.3503869

3. JinXiong Yang, Liang Bai, and Yanming Guo. 2020. A survey of text classification models. In Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI 2020). Association for Computing Machinery, New York, NY, USA, 327–334. DOI:https://doi.org/10.1145/3438872.3439101

APPENDIX

Certificate of Ethical Approval

Applicant:	Vedan Gokul
Project Title:	Classification of Wikipedia articles

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	24 Feb 2022
Project Reference Number:	P134433