**WEBPAGE CLASSIFICATION**

**USING NATURAL LANGUAGE PROCESSING**

**AND DEEP LEARNING**

A Synopsis Submitted in Partial Fulfillment of the Requirements

for the Degree of

**Bachelor in Technology**

In

**Computer Science & Engineering**

by

Aman Mundey - 01615602715

Anurag Sharma – 02215602715

Nikhil Gaur – 05915602715

Nishant Chhetri - 06015602715

**Department of Computer Science & Engineering**

**Dr. Akhilesh Das Gupta Institute of Technology and Management, New Delhi**

**Guru Gobind Singh Indraprastha University, New Delhi**

**August,2018**

**Synopsis**

1. **Introduction**

Today everything is on the internet and billions of websites are there which have thousands of webpages on different categories of news, sports, education, etc. To process such huge number of pages and categorize them manually by a human cannot be imagined. But AI comes to rescue. Using Natural Language Processing in Deep Learning Algorithms is very new field in comparison to traditional NLP methods. There are still many challenging problems to solve in Natural Language and this project is an attempt to solve such a problem.

1. **Motivation**

Webpage classification is a process of assigning webpage to one or more category labes eg. news, sports, education,etc. Web page classification is a vital task for imporving the quality of web search, constructing and expanding web directories, helping question and answer systems, web content filtering, assisted web browsing and knowledge base construction and many more. Therefore creating a well defined pipeline for this process will prove to be very important for many existing problems.

1. **Related Work**

For decades machine learning approaches targetting NLP problems have been based on shallow models eg. SVM, NB and linear models. In the last few years, neural networks based on dense vector representations have been producing superior results on various NLP tasks. This trend is sparked by word embeddings and deep learning methods. Also, simple deep learning frameworks outperform most state of the art approaches in several NLP tasks.

1. **Proposed Method**

The problem that will be tackled in this project will be multiclass single label classification. It will be a hard and flat based classification problem.

This is a very different problem than normal text classification as the data of website will be processed from its hyperlink and because of uncontrolled nature of webcontent advanced technique of data preprocessing will be required .

Text Preprocessing concepts like Lemmatization, Stemming, Stopwords, term frequency, Inverse document frequency, etc will be used.

Natural Language Processing concepts like bagofwords, ngram, word embeddings (word2vec) , semantic and syntax analysis, etc will be used.

We will use deep learning algorithms to train our model such as convolutional neural network, recurrent neural networks and recursive neural networks.

1. **Methodology**

Coding will completely be done in python programming language. Various Machine learning and NLP libraries will be used like sklearn, nltk , pytorch, keras etc.Text preprocessing will be perform with pandas, regular expression(re) , sklearn.feature\_extraction ,etc. Feature Engineering will be performed of onpage features and features of Neighbors in order to extract maximum out of hyperlinks. Results will be interpreted in the form of accuracy score and different machine learning models will be compared in terms of accuracy and training time. Goal will be to classify a new and unseen webpage accurately using the classifier.

1. **Plan of Work**

Aug 20- Aug 30 : Preparing and Preprocessing data.

Sep 1 - Sep 15 : Feature Engineering and Natural Language Processing.

Sep 15 - Sep 30: Model Training and validation

Oct 1 - Oct 25: Improving accuracy and comparing various models

Oct 26 : Final Report Submission

**References**

[1] Web Page Classification: Features and Algorithms by XIAOGUANG QI and BRIAN D. DAVISON from Lehigh University

[2] MACHINE LEARNING ALGORITHMS IN WEB PAGE CLASSIFICATION W. A. AWAD Math.& Computer Science Dept., Faculty of Science, Port Said University, Egypt. Scientific Research Group in Egypt (SRGE)

# [3] Recent Trends in Deep Learning Based Natural Language Processing

[ReviewArticle]Tom Young Information and Electronics, Beijing Institute of Technology, Beijing, China

[4] Sentiment Analysis of a document using deep learning approach and decision trees by Arman S. Zharmagambetov LLC AlemResearch Almaty, Kazakhstan and Alexandr A. Pak Institute of ICT Almaty, Kazakhstan.

[5] DEEP NEURAL NETWORK LANGUAGE MODEL RESEARCH AND APPLICATION OVERVIEW

FU-LIAN YIN, XING-YI PAN, XIAO-WEI LW, HUI-XIN LIU , Communication University of China, Beijing, 100024, China