RANKING OF SOCIAL MEDIA POSTS WITH RADICAL CONTENT

A Project Report Submitted

for the Course

MA498 Project I

by

Anurag Barfa

(Roll No. 160123005)

and

Ashish Ranjan

(Roll No. 160123006)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

November 2019

CERTIFICATE

This is to certify that the work contained in this project report entitled "Ranking of Social Media Post with Radical Content" submitted by Anurag Barfa (Roll No.: 160123005) and Ashish Ranjan (Roll No.: 160123006) to the Department of Mathematics, Indian Institute of Technology Guwahati towards partial requirement of Bachelor of Technology in Mathematics and Computing has been carried out by them under my supervision.

It is also certified that this report is a survey work based on the references in the bibliography.
Turnitin Similarity: 25 %

Guwahati - 781039 (Dr. Ashok Singh Sairam)
November 2019 Project Supervisor

ABSTRACT

Social media content is a significant part of life for all. People get the most information and news from social media. It is crucial for social media platforms to provide users with the best information they need. So there is a need to get the most relevant and important posts for a particular user. This translates to the problem of ranking social media posts. The main aim of this project is to rank the posts with some radical content as they maybe a threat to society. In this report, we have studied existing papers and thought of new ways in which we can get the features that capture all the details of the post required for ranking. We have also studied how learning to rank algorithms can be used for ranking social media posts.

Contents

List of Figures

Chapter 1

Dataset Preparation

1.1 Getting Twitter data

1.1.1 Scraping

1.1.2 API

1.2 Feature engineering

creating feature from raw data

1.3 Cleaning data

1.3.1 handling missing data

1.3.2 Outliers

friends count likes comments retweets

1.3.3 Converting data types

1.4 Labeling data

1.5 Normalization

1.6 Reducing columns

1.6.1 Dimension Reduction

1.6.2 Feature Selection

1.7 Spliting into training and testing

1.7.1 K-Fold Cross Validation

Chapter 2

Experimental Results

2.1 Performance Metrics

2.1.1 NDCG

why we used NDCG@100 for our training metric and ERR for testing purposes

2.1.2 ERR

2.2 Use of Ranklib library

How it was used and to rank the tweets.

2.3 Tie breaker Algorithm

Algorithm to break the tie between tweets getting same score by model.

2.4 Model Comparison

Results of comparing different models using ranklib.

2.5 Parameter tuning - LambdaMart

2.5.1 label size

2.5.2 No. of trees

2.5.3 No. of leaves

2.5.4 learning rate

2.5.5 Min leaf support

2.6 Model feature statistics

Chapter 3

Building tool for user interface