Negative Comments Multi-Label Classification

Jayant Singh
16010118 ,
IIIT Senapati

Instructor:-
Dr. Nongmeikapam Kishorjit Singh

Introduction

- **This is a classification problem which deals with the comments which contain various type of hate issues.**
- **Negative Comments are one of the cases of cyber bullying.**
- **Aggression and related activities such as trolling peoples, harassing online involves negative comments in various forms.**
- **The task is to extract those comments and label them according to their type.**

Introduction

- **After the introduction of Machine Learning and having data in massive amount now it’s quite logical to make a tool which can tackle this problem faced by people generally.**
- **Text classification became an important tool in various sectors which is helping in quite different ways.**
- **Deep Learning is one of the famous and mostly used approach for doing it and shown greater result also.**

Multi-Label vs Multi-Class

- **Multiclass classification means a classification task with more than two classes; e.g., classify a set of images of fruits which may be oranges, apples, or pears.**
- **Multiclass classification makes the assumption that each sample is assigned to one and only one label: a fruit can be either an apple or a pear but not both at the same time.**
- **Multilabel classification assigns to each sample a set of target labels.**
- **This can be thought as predicting properties of a data-point that are not mutually exclusive, such as topics that are relevant for a document.**
- **A text might be about any of religion, politics, finance or education at the same time or none of these.[1]**

Feasibility Study

Economical and Technical Feasibility

- **The Hardware requirement for this project is a system with basic specifications.**
- **The tools used are open source, libraries used are open source which makes it Economically feasible.**
- **The Design is correct and lead to given requirement with resources easily available to build it, which makes it technically feasible too.**

Requirement Analysis

- **Atleast 4gb Ram system to implement.**
- **Python programming language.**
- **Keras with backend both(Tensorflow and Theano).**
- **Numpy,Padas,scikit-learn(Some major libraries required).**

Related Work

- **Convolutional Neural Networks for Toxic Comment Classification:-**

  CNN used for classification of the toxic-comment dataset provided by kaggle.Three different layer was implemented with different filters including the max pooling within all layers and concatenated at last.[2]

- **Identifying Aggression and Toxicity in Comments using Capsule Network:-**

  The Capsule layer is primarily composed of two sub-layers Primary Capsule Laye and Convolutional Capsule Layer.

Related Work

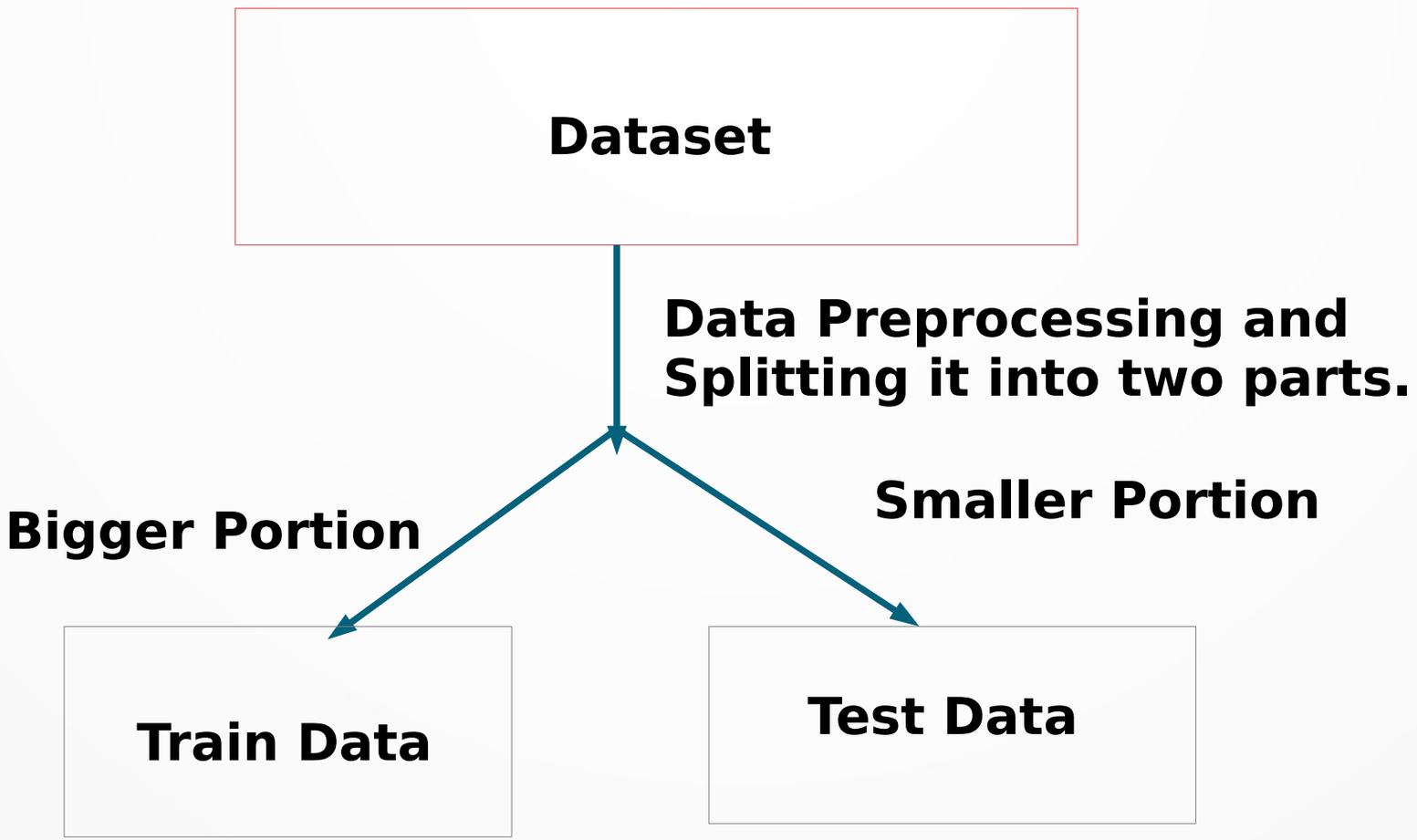
The primary capsules transform a scalar-output feature detector to vector-valued capsules to capture the instantiated features.[3]

- **Hate Speech Detection Using Natural Language Processing Techniques:-**
- **The dimension of the word vector is set to 100 at first and thus the embedding layer passes an input feature space that has a 3-dimensional tensor of shape (None, 100, 300) .**
- **The output of this layer is then fed into a 2D convolutional layer with filter layers of 3, 4, and 5, each having a 100 feature map.[4]**

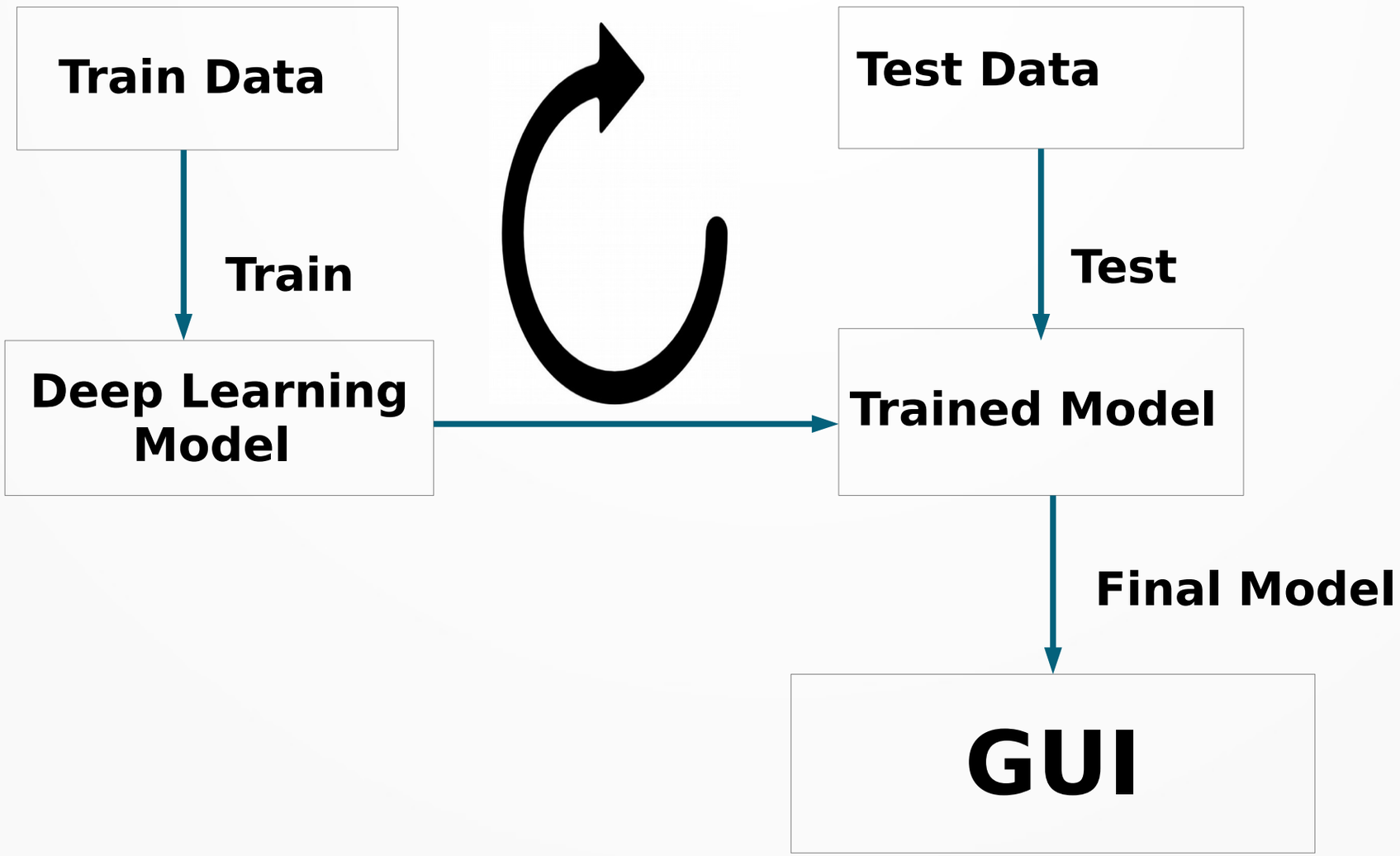
Dataset

- **Provided by kaggle.**
- **The Dataset Contains wikipedia talk page comments which are labelled in 6-different classes.**
- **One comment can be of multiple class signify by having value 1.**
- **0 if doesn’t belong to that class if for all label a comment have 0 in it then we say normal comment without toxicity.**
- **Toxic,severe_toxic,obscene,threat,insult,identity_hate are the classes.**

System Design



System Design



System Design

System Design

Data Preprocessing

- **As the dataset contains wiki talk page comments it consist of a lot of html tags removing those is primary.**
- **Punctuation marks are removed.**
- **Stop words are removed.**
- **Stemming has been done.**

System Design

- **Sentence are the converted into a matrix of vectors.**
- **Pre-Existing Embedding used to map the word to their vectors of 300 dimension(Glove vectors).**
- **Having variable length lead to do padding to the vectors by inserting extra ‘0’.**
- **Data are splitted in two sections test and train.**

System Design

Deep Learning Models

- **Previous work have been done on CNN so here we tend to use different model for our project.**
- **Adding a memory unit to the model could help in giving better acuracy so we try with RNNs.**
- **GRU Model:**
- **Uses update gate and reset gate to solve vanishing gradient problem.**
- **Basically, these are two vectors which decide what information should be passed to the output.**

System Design

- **The update gate helps the model to determine how much of the past information (from previous time steps) needs to be passed along to the future.**
- **Reset gate is used from the model to decide how much of the past information to forget.**
- **We are using GRU over other model as it is easy to modify and doesn't need memory units, therefore, faster to train than LSTM and give as per performance.**

Work In-Progress

- **Data are preprocessed by removing the html tags present as well as punctuations are removed, stop-words are vanished as well.**
- **Words are converted in vectors using embeddings(Glove).**
- **Working on models which will be suitable for the given data and give better accuracy.**

Refrences

- [1]https://scikit-learn.org/stable/modules/multiclass.html
- [2] **Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, andVassilis P Plagianakos. “Convolutional neural networks for toxic comment classification.” arXiv preprint arXiv:1802.09957, 2018**
- [3] **Saurabh Srivastava, Prerna Khurana, and Vartika Tewari. “Identifying aggression and toxicity in comments using capsule network.” In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying(TRAC-2018), pp 98–105, 2018**

Refrences

- [4] ***Shanita Biere, Sandjai Bhulai, and Master Business Analytics. "Hate speech detection using natural language processing techniques." 2018***

Thank You