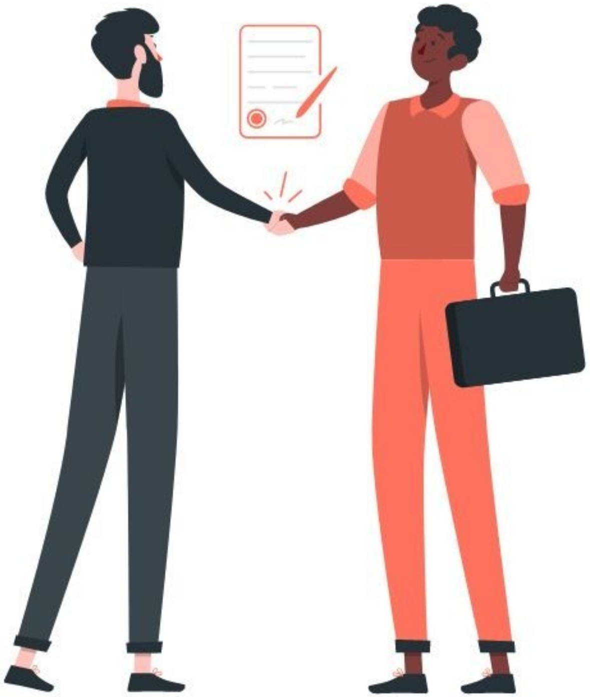
Character-Level Convolutional Neural Networks for Text Classification

SMAI Project



Atharv Sujlegaonkar
Tirth Motka
Jay Ghevariya
Mitul Garg

TABLE OF CONTENTS

01
Abstract

02
Introduction

03
Database used

04
Implementation

05
Deliverables

Abstract

This work looks into the use of convolutional networks on character level instead of traditional methods for text classification. Large datasets are used to show that results are as competitive as other methods. Comparison is done with models such as bag of words,

n-grams, and deep learning models such as word-based ConvNets and RNNs.

PROJECT CHECKPOINTS

GOAL 1

Researching on the topic, exploring various techniques to implement the project

GOAL 2

Studying about the convNet and its implementation

GOAL 3

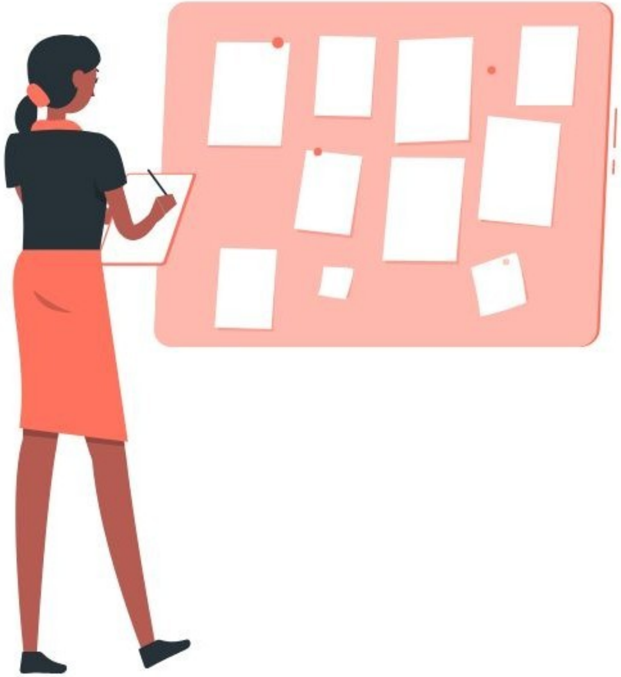
Implementation and drawing out useful inferences

GOAL 4

Error analysis and comparison of our model with different models/ transformers

Introduction

- The ConvNets based text classification model has enabled us to introduce deeper ConvNets with more than 3 layers which provides the model with the ability to extract data in a better way.
- Deeper Neural networks enable to train the same network structures for variety of datasets, and obtain optimal results.



Datasets used

Datasets

AG’s News

The AG's news topic classification dataset is constructed by choosing 4 largest classes from the original corpus. Each class contains 30,000 training samples and 1,900 testing samples. The total number of training samples is 120,000 and testing 7,600.

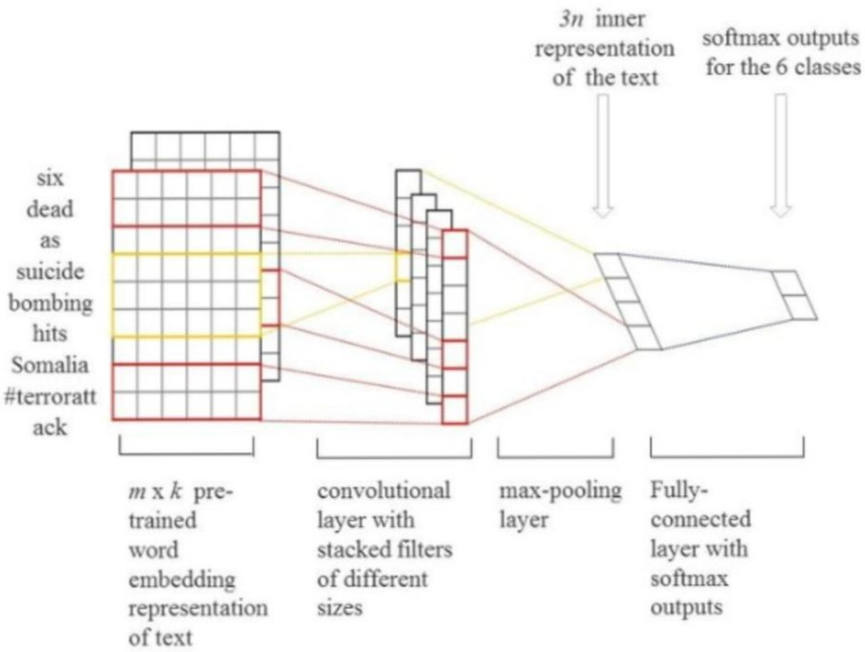
The file classes.txt contains a list of classes corresponding to each label.

DBPedia Ontology

They are listed in classes.txt. From each of these 14 ontology classes, we randomly choose 40,000 training samples and 5,000 testing samples.

Therefore, the total size of the training dataset is 560,000 and testing dataset 70,000. There are 3 columns in the dataset (same for train and test splits), corresponding to class index (1 to 14), title and content.

Implementation


3n inner representation of the text
softmax outputs for the 6 classes
six
dead
as
suicide
bombing
hits
Somalia
#terroratt
ack
m x k pre-trained word embedding representation of text
convolutional layer with stacked filters of different sizes
max-pooling layer
Fully-connected layer with softmax outputs

Deliverables

- training our CNN model on two different datasets for 6-8 epochs.
- Finding results based on variance and F1 score
- Finding accuracy , recall of the model and documenting the results

THANK YOU!!!!!