Toxic Comment Classification

EE5180 : Introduction to Machine Learning

Anirudh Sagar Gollapalli (EE18B007), Jayadev Joy (EE18B011), Naga Deepthi Chimbili (EE18B043), Varshitha Vadapally (EE18B065)

Introduction

- The social media is a unique place where people can express themselves without compromising their identity. But unfortunately, there are some people who misuse this freedom to spread negativity. Such a situation leads to a toxic environment and should be dealt with in order to enforce safe and healthy conversations. Taking the size and scale of the social media, it is not practical to use manual moderation. Which is why social media giants use Machine Learning to automate this process.

- We aim to build a classifier that detects the level of toxicity for a comment which would be useful to detect toxic comments and hence can be removed.

Problem Statement

What we attempt to do is to build a model which predicts the probability of each type of toxicity for each comment. This is a multi-label classification problem where one comment can belong to more than one label simultaneously. For example, a comment maybe toxic, obscene and insulting at the same time. It may also happen that the comment is non-toxic and hence does not belong to any of the six labels.

Data-Set Description

The Data-set used for this task is sourced from a Kaggle competition and is split into training data and test data. It is composed of comments from Wikipedia talk page edits.

The training data-set consists of a total of 1,59,571 instances with comments and the corresponding multiple binomial labels:

- Severely Toxic
- Toxic
- Obscene
- Threat
- Insult
- Identity Hate

Data-Set Description

Sample instances of the data-set are shown below in the figure:

	A	B	C	D	E	F	G	H
1	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
2	0000997932d777bf	Explanation	0	0	0	0	0	0
3	000103f0d9cfb60f	Shut up	1	0	1	0	1	0
4	000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just t	0	0	0	0	0	0
5	0001b41b1c6bb37e	"	0	0	0	0	0	0
6	0001d958c54c6e35	You, sir, are my hero. Any chance you remember w	0	0	0	0	0	0
7	00025465d4725e87	"	0	0	0	0	0	0

Existing Work

Here are a few papers based on Toxic Comment Classification and their approaches :

Title of the Paper	Key Points from their Abstract
Multilabel Toxic Comment Classification Using Supervised Machine Learning Algorithms	We tested two models: RNN and LSTM. Their performance is significantly better than that of Logistic Regression and ExtraTrees, which are baseline models.
Abusive Comments Classification in Social Media Using Neural Networks	The comments are classified as abusive and non-abusive using convolutional neural network (CNN) model
Toxic Comment Classification	The following work explores the usage of a very basic Logistic Regression classifier and then moves on to explore Deep Learning based approaches based primarily on Sequential Models, especially the Bidirectional LSTM architecture.

Existing Work

Title of the Paper	Key Points from their Abstract
Detecting Abusive Comments Using Ensemble Deep Learning Algorithms	We have applied four classification algorithms: Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine, with Bag of Words features. Deep learning algorithms: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and an ensemble of LSTM and CNN are applied using GloVe and fastText word embedding
Toxic Comment Classification	The current study implemented several algorithms, including Naive Bayes (as a benchmark), Recurrent Neural Network (RNN), and Long Short Term Memory (LSTM).
Convolutional Neural Networks for Toxic Comment Classification	We choose to compare CNN against the traditional bag-of-words approach for text analysis combined with a selection of algorithms proven to be very effective in text classification. The reported results provide enough evidence that CNN enhance toxic comment classification reinforcing research interest towards this direction.

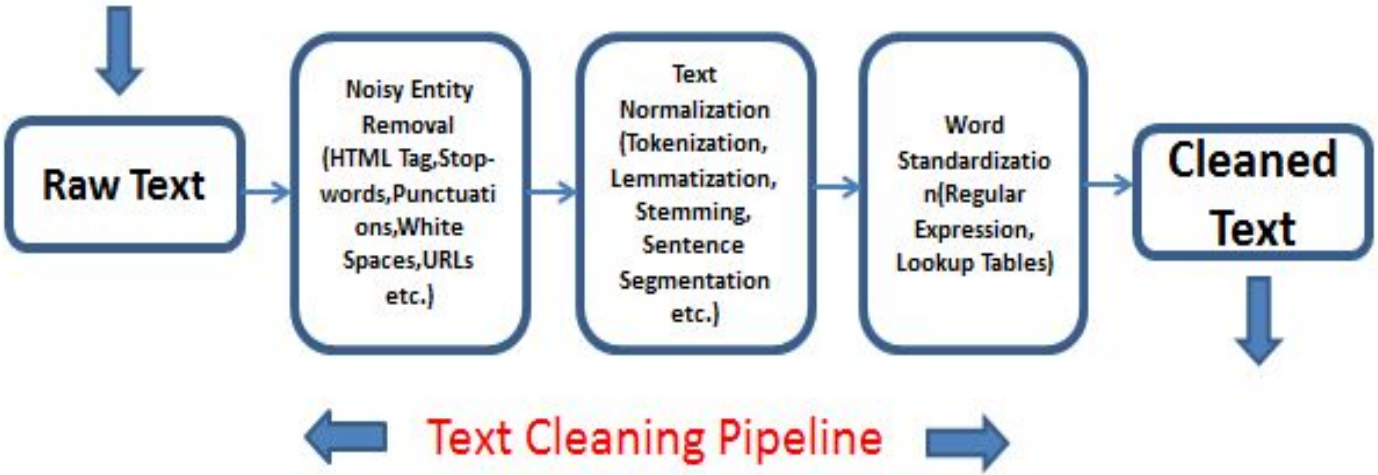
Proposed Work

1. Preprocessing

- Since this is a text based dataset, there are a few specific set of preprocessing tasks that we should do so that we can refine the dataset and help the model perform better.
- First, we would have to remove white spaces, punctuations, etc. Because, they will be of no use to what the text is implying.
- Then, we should also remove stop words. Because, stopwords such as there, over, etc do not contribute to the context or meaning of the text.
- After that, we should lemmatize the text. Because, words like changing, changed , changes etc. more or less mean the same thing, i.e change.

Proposed Work

- After performing all the preprocessing tasks, we would obtain a clean optimal dataset. Which we will then use in the subsequent stages. Such as splitting the dataset into training data and testing data and then training the model.



Proposed Work

2. Models to be used

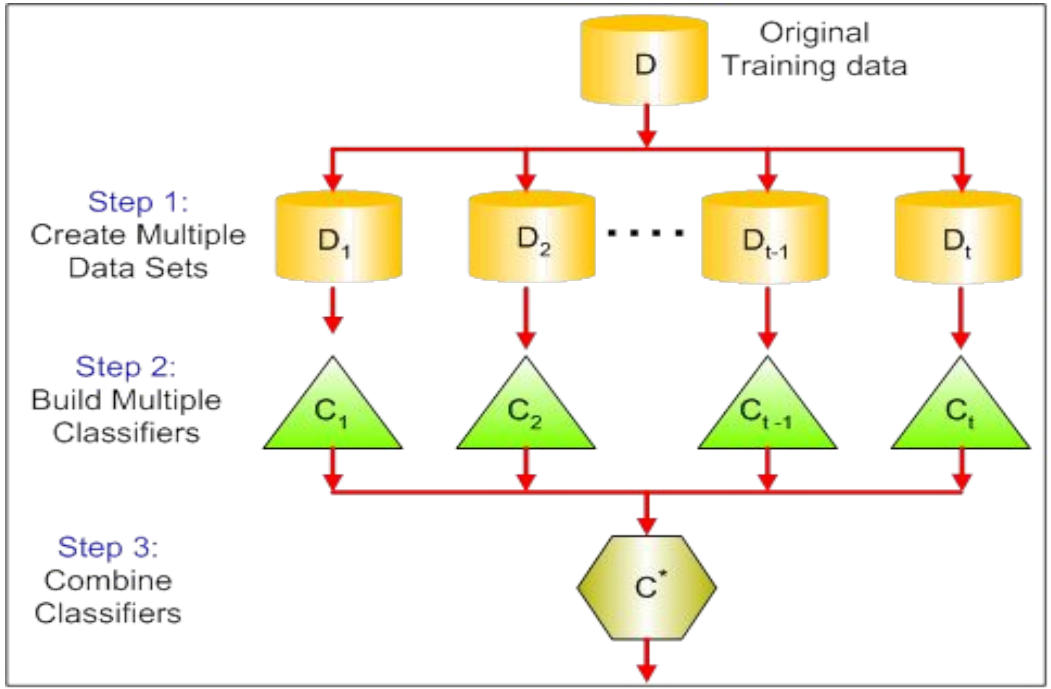
- We first plan on using a few benchmark models such as Naive Bayes Classifier, Logistic Regression to see how baseline models perform. So that, we can clearly understand how other ‘complex’ models perform in comparison to these ‘simple’ models.

- We would then like to use more complex models like Random Forest Classifier and Support Vector Machines (SVM)

- And then we would like to proceed to use even more complex models involving Neural Networks such as Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM).

Proposed Work

- We could use methods like ensembling where we could use multiple models at a time to increase our accuracy.



Proposed Work

- We can also perform Hyperparameter Tuning and also use advanced methods like Word Embedding to get better results.

- We would then, like to compare all the aforementioned models and then find out the best model.

References

- Multilabel Toxic Comment Classification Using Supervised Machine Learning Algorithms
- Detecting Abusive Comments Using Ensemble Deep Learning Algorithms
- Abusive Comments Classification in Social Media Using Neural Networks
- Toxic Comment Classification
- Convolutional Neural Networks for Toxic Comment Classification
- Toxic Comment Classification

Thank You