Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

Audio and Video Toxic Comments Detection and Classification

Ms Sangita Holkar
Department of Computer Engineering
Datta Meghe College of Engineering,
Airoli, Navi Mumbai, INDIA

Dr. S. D. Sawarkar
Department of Computer Engineering
Datta Meghe College of Engineering,
Airoli, Navi Mumbai, INDIA

Dr. Shubhangi Vaikole
Department of Computer Engineering
Datta Meghe College of Engineering,
Airoli, Navi Mumbai, INDIA

Abstract: Social networking and online conversation platforms provide us with the power to share our views and ideas. However, nowadays on social media platforms, many people are taking these platforms for granted, they see it as an opportunity to harass and target others leading to cyber-attack and cyber-bullying which lead to traumatic experiences and suicidal attempts in extreme cases. Manually identifying and classifying such comments is a very long, tiresome and unreliable process. To solve this challenge, we have developed a deep learning system which will identify such negative content on online discussion platforms and successfully classify them into proper labels. Our proposed model aims to apply the text-based Convolution Neural Network (CNN) with word embedding, using fastText word embedding technique. fastText has shown efficient and more accurate results compared to Word2Vec and GLOVE model. Our model aims to improve detecting different types of toxicity to improve the social media experience. Our model classifies such comments in six classes which are Toxic, Severe Toxic, Obscene, Threat, Insult and Identity-hate. Multi-Label Classification helps us to provide an automated solution for dealing with the toxic comments problem we are facing.

Keywords: Convolution Neural Network (CNN), Python-tesseract, Fastext

1. INTRODUCTION

In today's world, conversation over online social forums is one of the most common and easy ways of communicating and expressing one's thoughts. These platforms allow to discuss various topics, share information and opinions overa topic. But nowadays maintaining decency and a good level of conduct or behavior over these platforms could be difficult. A lot of abusive content, harassment, jeering, cyber-bullying related activities have become very common on such platforms which have harmful effects on a person's mental and psychological health. This can sometimes lead to detrimental and life-long traumatic effects on an individual. Such type of situations can traumatize users and stop them from expressing their opinions, completely alienating themselves and stop seeking and receiving help from others. The companies which own such online discussion platforms have been working on different solutions such as comment classification techniques, user blocking mechanisms and comment filtering systems. In the comment classification approach, the goal is to classify the comments or sentences based on their toxicity levels into various categories. By categorizing these comments, the action team can take appropriate actions to curb the occurrence and growth of negative influences created with such activities on social platforms. Such a multi-label classification model will make the purpose of social conversation on social media more effective and positive. By automating this comment classification approach, the companies can save their time and manual efforts in moderating these platforms.

1.1 PROPOSED MODEL:

Our proposed model is a system comprising of fastText word embedding technique and CNN which perform multi-label classification of toxic comments. In this system, input as comments will be fed from social sites and images which will be analyzed and sent to word embedding phase. In this phase,sentences are broken into words and embedded into vectors time the user starts stops or restarts the video. This is time based that is, ever time user hits an event the corresponding time is captured against that event. opinions, completely alienating themselves and stop seeking and receiving help from others. The companies which own such online discussion platforms have been working on different solutions such as comment classification techniques, user blocking mechanisms and comment filtering systems. In the comment classification approach, the goal is to classify the comments or sentences based on their toxicity levels into various categories. By categorizing these comments, the action team can take appropriate actions to curb the occurrence and growth of negative influences created with such activities on social platforms. Such a multi-label classification model will make the purpose of social conversation on social media more effective and positive. By automating this comment classification approach, the companies can save their time and manual efforts in moderating these platforms.

The data that we have used for our model is Kaggle's Toxic Comment Classification Dataset on Wikipedia's talk page edits. Using CNN, our aim is to develop a multi-label classification model which classifies the comments based on its toxicity level into 6 different categories toxic, severe-toxic, obscene, threat, insult and identity-hate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

1.2 METHODOLOGY:

Machine or deep learning algorithms cannot process strings (plain text) as inputs. These algorithms are unable to process such strings as raw input. Word embedding technique provides a solution for this issue by transforming the string text into a numerical format or vector format which can be used by model and also this format can be used to find semantic relationships between the associated words by calculating the difference between those two respective vectors, referred as embedding space.are that it generates better word embeddings for rare words, or even words not seen during training because the n-gram character vectors are shared with other words.

1.3 Literature Survey

Let's take analysis of different proposed methodologies for efficient class detection system and our proposed method for text clasification. Different CNN approaches are applicable for efficient prediction of class. Various popular methods are:-

- One of the widely used Natural Language Processing & Supervised Machine Learning (ML) task in different business problems is "Text Classification", it's an example of Supervised Machine Learning task since a labelled dataset containing text documents and their labels is used for training a classifier. In today's machine learning applications, support vector machines (SVM) are considered a must try—it offers one of the most robust and accurate methods among all well-known algorithms
- CNN is a class of deep, feed-forward artificial neural networks (where connections between nodes do not form a cycle) & use a variation of multilayer perceptrons designed to require minimal preprocessing. These are inspired by animal visual cortex.
- The result of each convolution will fire when a special pattern is detected. By varying the size of the kernels and concatenating their outputs, you're allowing yourself to detect patterns of multiples sizes (2, 3, or 5 adjacent words).Patterns could be expressions (word ngrams?) like "I hate", "very good" and therefore CNNs can identify them in the sentence regardless of their position..
- Convolution Neural Network (CNN), generally we refer to a 2 dimensional CNN which is used for image classification. But there are two other types of Convolution Neural Networks used in the real world, which are 1 dimensional and 3-dimensional CNNs.
- A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit dynamic temporal behavior for a time sequence.
- Using the knowledge from an external embedding can enhance the precision of your RNN because it integrates new information (lexical and semantic) about the words, an information that has been trained and distilled on a very large corpus of data.The pre-trained embedding we'll be using is GloVe. in any particular application, but it can usually be relied on to be robust and to do quite well.

1.4 Using CNN, our aim is to develop a multi-label classification model which classifies the comments based on its toxicity level into 6 different categories toxic, severe- toxic, obscene, threat, insult and identity-hate.

1.5 Aim and objective

Aim is to use CNN features given by such outstanding FASTEXT method and to increase detection rate for accurate classification. So in proposed system, aim is to increase correct detection of comments and clasify them.

Objective

- Creating a word embedding mechanism that can help to identify the slang or negative terms in the comment.
- Identify the class to which the slang term belongs based on toxicity level and assign corresponding weights.
- Perform an efficient pre-processing unit to make data suitable for analysis
- Create a CNN model to perform classification process by training the model with training data to fit and test to model to evaluate the accuracy rate.
- Improving the model by backpropagation mechanism and managing the trade-off between over-fitting and under-fitting.

2. Stages In Proposed System

1. Training text: It is the input text through which our supervised learning model is able to learn and predict the required class.
2. Feature Vector: A feature vector is a vector that contains information describing the characteristics of the input data.
3. Labels: These are the predefined categories/classes that our model will predict
4. ML Algo: It is the algorithm through which our model is able to deal with text classification (In our case : CNN)
5. Predictive Model: A model which is trained on the historical dataset which can perform label predictions

Above steps can be shown as follows:

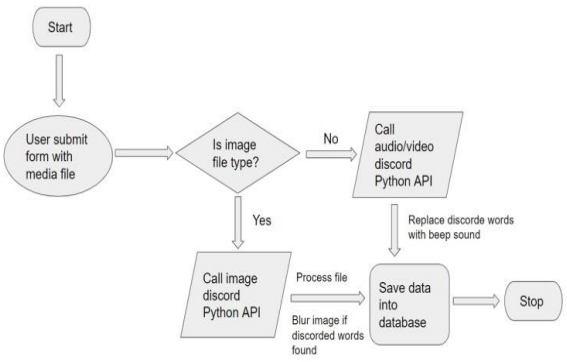
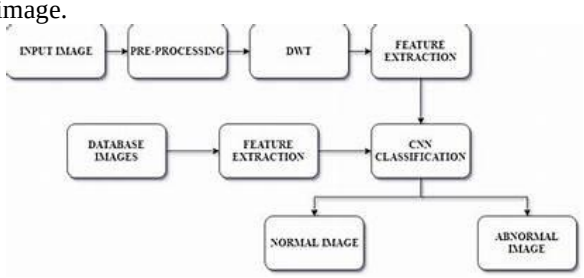


Fig 1: Proposed System

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

2.1 Working

1. Image or the audio file is uploaded from the social sites or other dataset and Once the data is submitted it is separated according to the extension and then the words from the image file and audio file are recognised with the help of python api consisting the image recognistion cnn use feature extraction technique if the image recognition method is not recognised the image is labeled as abnormal image.



2.Data Study and Analysis: Work is to analyze the content, nature of data gathered and try to visualize the data present in terms of number of words and labels provided in the training set. Figure 3 and 4 shows the visualization performed in our project with the Wikipedia dataset. We can derive that the number of comments having less than 200 words is very high in our dataset. Also, we can make a point from that most of the comments belongs to categories of toxic, obscene and insult.

3.Data Pre-processing: To prepare data for model, we here remove all punctuation characters, conversion into lowercase letters, removing stop-words and appending the numeric digits to it.

4.Embedding word into vectors: The processed comments are converted into vector format, in numerical representation using fastText.

5.Partitioning into training and test set: Splitting the data into training and test data with a 70:30 ratio respectively.

6.Training on CNN: First layer performs convolutions over the embedded word vectors using multiple filter sizes, then pooling layers and the dense layers, model will be saved.

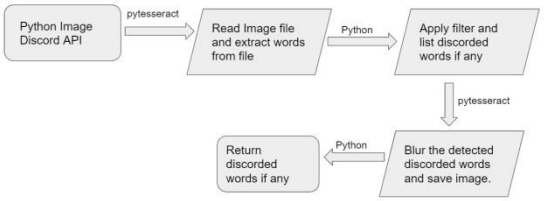
7.Save model: Saving model checkpoints for future predictions. This also helps while tuning or adjusting the parameters of the model for accuracy improvement

8.Load model: Load model for predictions on the new unseen data, that is, unseen comments.

9.Visualizing with labels: Visualizing or showing tagged or label comments on the platform end. Efficient visualizing the results with a well storytelling of the data or insights can help organizations to take proper measures for their platforms and reduce the chances of such activities to happen, thereby creating a positive impact on the digital world.

2.2 Preprocessing dataset

Analysing Our Data :



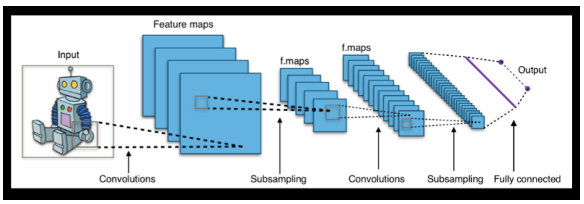
1. Python-tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and “read” the text embedded in images.

2. Python-tesseract is a wrapper for Google’s Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Additionally, if used as a script, Python-
tesseract will print the recognized text instead of writing it to a file.

3. Convolutional Neural Network (CNN) is one of the popular neural networks widely used for image classification. When an image is fed to CNN, the convolutional layers of CNN are able to identify different features of the image. The ability to accurately extract feature information from images makes CNN popular.

4. A neural network is a network of neural layers. Each layer contains what is called neurons. How we connect these neurons make up the configuration of a Neural Network.

5. CNN is a special type of neural network. A convolution neural network consists of an input layer, convolutional layers, Pooling(subsampling) layers followed by fully connected feed forward network.



A fully connected neural network is a network in which each neuron from one layer is connected to all neurons on its adjacent layers.

Consider a simple 100×100 pixel image. And let’s say, we have 20 hidden layers with 50 neurons in each layer. So for training the network, the total number of parameters in this fully connected neural network to process 100×100 pixel image would be 100x100x50x20 + bias which is more than 10000000 parameters. Omg! So yeah, this is rightly known as ‘Parameter Explosion’. Thus we need a mechanism that can work well on images but using much lesser parameters

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

than that is used in a fully connected feed-forward neural network. Yes…CNN works with a lesser number of parameters and hence also less prone to overfitting!

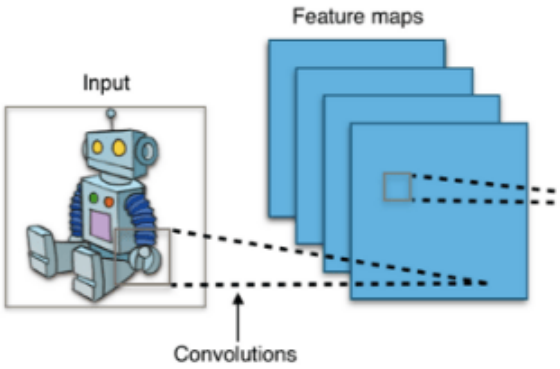
I. CONVOLUTIONAL LAYERS:

CNN network uses the mathematical concept called Convolution and hence the name. Convolutional Layers are formed by applying the sliding window function to the matrix representing the image. This sliding window function is known as kernel or filter which is nothing but another matrix. In the below example, a 3 x 3 kernel is slid across the 10 x 10 input matrix. The 3 x 3 kernel matrix (K) is multiplied with every 3 x 3 matrix available from the input matrix (I).

$$S_{ij} = (I * K)_{ij}$$

where S is the resultant matrix. (8 x 8 matrix in the above example). The resultant matrix is also known as the 'Feature map'. This operation of matrix multiplying the kernel with image sections is known as convolution.

In CNN, on each convolutional layer, the process does not only use just 1 filter but uses multiple such filters(/kernels) the result of which is we get multiple convoluted matrices. Each convoluted matrix (feature map) gives information on different aspects of the image. In the below picture there are 4 kernels used.



This concept of using different kernels gives CNN the ability to preserve the spatial structure of the image as the value is calculated based only on its surrounding pixels. Also, note the same kernel matrix is used to produce a convolutional matrix. This is why there are so many fewer parameters on CNN when compared to a feed forward network taking an image to process. In other words, on CNN, the image is transformed into much smaller dimensions before it is fed into the fully connected layers of the CNN

The kernel values are learned automatically by the network using backpropagation. The parameters that we have to decide are the number of kernels and the size of the kernel matrix to be used on each layer in the network. In general 3×3 kernel is used on small images and 5×5 or 7×7 and more on larger images. Generally, fewer filters (16, 32) are used at the input layer, and more filters used at subsequent deeper layers.

The output of the convolution is a feature matrix that is smaller than the input matrix. In the below example, convolving 3 x 3 kernel with 10 x 10 image produces an 8 x 8 feature map. Also, in the convolution process, the corner pixels of the image matrix are covered less number of times when compared to the non-corner pixels. This obviously impacts the quality of information extracted from the corners of the images.The corners of the images are processed the same as the inner parts of the image, one solution is to have the original image matrix surrounded by a new border like below. This is called padding.The rectified output of the convolution layer is then passed to the pooling layer. Pooling is used to reduce the size of the input matrix to the subsequent layer. Max pooling is the most commonly used pooling method.

CNN can contain multiple convolution and pooling layers. Finally, the output of the last pooling layer of the network is flattened and is given to the fully connected layer. The number of hidden layers and the number of neurons in each hidden layer are the parameters that needed to be defined. Sigmoid and Softmax activation functions are used at these layers to output the class probability.

2.2.1 Data Trimming

Using character n-grams can create a more robust network as partial components of words are often shared. However, for our case, we use a combination of words and punctuation, leaving out the parts-of-speech and not using characters.

After, the n-grams have been created, the features are then averaged (pooling) and send to the hidden variables. Then we apply a softmax activation function to the output, that's it!

Training set:

Training set comprises of all columns with label

Example: Sentence: This is a test phrase.

1-Gram (Unigram): [This, is, a, test, phrase, .]

2-Gram (Bigram): [This is, is a, a test, test phrase, phrase.]

3-Gram (Trigram): [This is a, is a test, a test phrase, test phrase.]

To clarify, we are using word embeddings as opposed to character embeddings.

Character embeddings for the above would look something like this:

Sentence: This is a test phrase.

3-Gram (Trigram): [This, his, is i, s is, s a, a t, a te, tes, test, est, st p, t ph, phr, phra, hras, rase, ase, se ., e . , .]

CONCLUSIONS

Our model based on multilabel classification using fast Text and CNN, is useful in detecting toxic and abusive comments on social media platforms and categorizing them according to their toxicity. We have presented multiple approaches for toxic comment classification using fastText and Word2Vec. Here using the classification obtained, social media platforms can

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

implement this system and curb negative influences on social media. As we have tested between fastText and Word2Vec and concluded from our result, that fastText is more accurate and is more accurate when dealing with slangs, jargons, typing mistakes and short forms used.

ACKNOWLEDGMENTS

.This System Is Implemented Under Guidance Of Prof S.D .Sawarkar And Prof Shubhangi Vaikole ,Department Of Computer Engineering ,Airoli, India

REFERENCES

[1] Thedora Chu, Max Wang, Kylie Jue. “Comment Abuse Classification with Deep Leraning.” Stanford University.
[2] Karthik Diankar, Roi Riechart, Henry Lieberman. “Modeling the Detection of Textual Cyberbullying.” Massachusetts Institute of Technology, Cambridge MA 02139 USA.
[3] Xin Wang, Yuanchao Li, Chengjie Sun, Baoxum Wang and Xialong Wang. “Polarities of Tweets by Composing Word Embeddings with Long Short Term Memory.” 7th International Joint Conference of Natural Language Processing. July-2005.
[4] S. V. Georgakopoulus, A. G. Vrahatis, S. K. Tasoulis, V. P. Plagianakos. “Convolutional Neural Networks for Toxic Comment Classification.” arXiv:1802.099574v1 [cs.CL], 27 Feb 2018.
[5] Manav Kohli, Emily Kuehler and John Palowitch. “Paying attention to toxic comments.” Stanford University.
[6] E. Wulcyzn, N.Thain and L.Dixon. “Ex-Machina-Personal attacks seen at scale.” 2017 International World Wide Web Conference Committee (IW3C2), ACM 978-1-44501-4913- - 10/17/04