MEMES CLASSIFICATION SYSTEM USING COMPUTER VISION AND NLP TECHNIQUES

Rachana Jadhav
Department of Computer Science and Engineering
Walchand College of Engineering (An Autonomous Institute)
Sangli, India

Prof. Vikas. N. Honmane
Department of Computer Science and Engineering
Walchand College of Engineering (An Autonomous Institute)
Sangli, India

Abstract—A meme is a culturally relevant, and brief form of media that raise a content over the internet. Now a days posting a meme is popular communication medium, due to its multimodal nature. Postings of hateful memes or fooling, cyberbullying are growing gradually. Meme takes a major part in forming people’s trust and perspective. Meme can be quickly post by anybody, and its integrity stands compromised. Hateful and aggressive matter detection have been largely traversed in a form such as text or image. And Memes complicate the task, because some meme can have a good caption and normal pictures, but if combined in some way, they can become offensive. So, it is required to fuse both modality to identify whether a given meme is hateful or not. So here for text classification we found the sequential model like Bi-LSTM and for image we will go with CNN. Late fusion technique is used to combine the image and text mode with EX-OR method to investigate its effectiveness.

Index Terms—Hateful meme, Deep Learning, Natural Language Processing, Late fusion, Image Processing, Bi-LSTM, EX-OR.

I. INTRODUCTION

Now a days social media be it Facebook, Instagram, Twitter or any other social media has gained a lot of enticement since its beginning. Day by day the content generation is growing rapidly. Almost half population of society is present on social media. Most of the present day memes have more impact on people as they are structurally apart from longer form of media. Memes become successful because they are Familiar, short, brief, culturally relevant and quickly understood by the target audience.

Memes can be in many forms like
- Image Dominant: where textual data is less.
- Text Dominant: where textual data is huge
- Text and Image Dominant: where textual and image data are same.

Recently the most central usage of memes to be of images combined with text. Irrespective of the type of meme, the meme may get modified, rehashed or recreated while inter-acting through social media network. In this age memes are decorated fancy pictures that are put forward to be diverting, often as a way to freely disrespect individual control. Additional memes can be spoken recordings, and few memes have a more logical subject. So due to the massive use of social networks, negative impacts need to be automatically limited.

This project serves to propose a technique which uses text mode and image mode processing to classify the memes as hateful or non-hateful and stop spreading hate across internet. This work aims to conduct thorough approaches in computer vision, deep learning and natural language processing.

II. LITERATURE REVIEW

As there was need to detect offensive content author
Because we have to combine data from different modalities.

Here we are going with Late Fusion approach late fusion will classify content for both text and image mode before trying to fuse the results.

A. Memes pre-processing

1. Textual Data Processing

The first step before feeding text data to any machine learning algorithm involves text pre-processing in order to clean the data.

- Remove the unnecessary tags.
- Remove any punctuation or a limited set of special characters like, or, etc.
- Check if the word is not alpha-numeric and should made up of English letters.
- Convert into lowercase form.
- Remove Stop-words which do not add any significance, like is, am, the etc.
- Convert into short into full as these words are not registered in dictionaries. As a result, the main problem is that the same word may have different meanings, depending on the situation.

2. Image Data Processing

- Resizing: The memes collected are of different sizes so need to resized uniformly before being used as input to model therefore, we should establish a base size for all images fed into our algorithms.
- Noise Reduction: Various filters can be used. Most commonly used filters are Mean filter, Gaussian Filter, Median Filter will use one which will improve the system functioning.
- RGB2Gray: Converting color image to gray scale image.
B. Memes Feature Extraction

Textual Feature Extraction:
One hot encoding work straight with categorical variables with representing them as binary vectors. Each word is written or encoded as a one-hot vector, and each one-hot vector is unique, which allows a word to be uniquely identified by its one-hot vector. The reverse is also true, that is two words do not have the same one-hot vector representation. In this respect we can assign ‘non-hateful’ an integer value of 0 and ‘hateful’ the integer value of 1. And its corresponding binary vector [1, 0] for ‘non-hateful’ and [0, 1] for ‘hateful. Therefore, we represent each word and character in the text data as a unique vector, which includes numeric data (1 and 0) as its components. A word is represented as a vector so that a list of words can be represented in a sentence as a vector array or matrix. If we have a list of sentences in which words are hot-encoded, the result is a matrix whose elements are matrices. Then we get a three-dimensional tensor that can be passed to the neural network.

C. Memes Classification

1) Image Classification using CNN:
CNN image classification takes input images, processes them and classifies them into specific categories (such as dogs, cats, tigers, lions). The computer treats the input image as an array of pixels and depends on the image resolution. You will see H x W x D (H= height, W= width, D = size). For example, a 6 x 6 x 3 matrix image of RGB (3 refers to RGB values) and a 4 x 4 x 1 matrix image of grayscale Technically speaking, for the deep learning CNN model used for training and testing, each input image goes through a series of convolutional layers with filters (kernels), pools, and fully connected layers (FC), and the softmax function is applied For objects with probability values to be classified are between 0 and 1. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.

2) Text classification using Bi-LSTM
Bidirectional LSTM or bi-LSTM is a sequence processing model that consists of two LSTMs: one receives input data in the forward direction, and the other receives input data in the opposite direction. Bi-LSTM effectively increases the amount of information available to the network by improving the context available to the algorithm. (For example, know which word immediately precedes a word in a sentence) The Bi-LSTM neural network consists of LSTM modules that work bi-directionally, including past and future contextual information. Bi-LSTM can learn long-term dependence without obtaining repeated contextual information, so it exhibits excellent performance in sequential modeling tasks and is widely used in text classification. In the LSTM network, the Bi-LSTM network has two parallel layers, which propagate in two directions by passing forward and backward to capture the dependencies in the two contexts. The structure of the Bi-LSTM network is shown in Figure.
D. Fusion of Text and Image

Ex-OR Prediction: If any one of them be a text or image is hateful then it will result hateful. The following snippet of code will show you the exact idea that has been implemented:

```python
In [54]: def predict(image, text):
    #Image Prediction
    image = np.array(image) / 255
    image = image.reshape(1, 40000).astype("float32")
    image_class = model.predict(image)

    #Text Prediction
    text_list = []
    text_list.append(text)
    text_padded = pad_sequences(text_list, maxlen=100, padding='post')
    text_class = model.predict_classes(text_padded)

    aexor weight age prediction
    if (image_class == text_class == 1):
        return 1
    else:
        return 0
```

Fig. 5. Fusion

IV. SIGNIFICANCE

To stop spreading offensive content on online social media. Also, few brands are marketing with memes to connect with their followers or consumer on social media so it can be supportive in advertising and promotion or publicity of products and services. Some people said or done something mean or cruel to another person online so this system will be helpful for Cyber bullying monitoring. Trend analysis i.e. you can get information from past and you will come up with an idea for future in areas like intellectual analysis, finding the popularity of a political party.

V. RESULTS AND DISCUSSION

We worked on hateful memes classification through approaches in Natural Language Processing and Deep learning with help of different libraries like NLTK, Keras, Tensorflow, SkLearn, Matplotlib, and Scikitplot to get work done. We have used a sequential model for text classification like Bi-LSTM which gives average accuracy of 87% And for image Convolutional Neural Network is used with average accuracy of 35%. Finally, Late fusion of meme with EX-OR prediction of both the form that is text and image is done which gives 87% overall accuracy.

The most popular performance measures such as accuracy, precision, sensitivity, F1 Score are used to evaluate the performance of the proposed model.

- Accuracy = Total No. of Correct Predicted / Total No. of Observations
- Precision = Total No. of correct predicted Category Observations / Total No. of Predicted Category Observations
- Recall = Total No. of correct predicted Category Observations / Total No. of Actual Correct Category Observations
- F-Measure is the arithmetic mean of precision and recall.

Fig. 6. Confusion Matrix

The following GUI will give you the exact idea that has been implemented:

Fig. 7. Hateful
VI. CONCLUSION AND FUTURE WORK

We proposed to investigate the important problem of memes classification system using computer vision and NLP techniques. Suggests a method that could be beneficial to classify memes with fix visual and textual features. Late Fusion will classify content for both text and image before trying to fuse the results. Eventually, further research and work to identify hateful meme is in progress and will give a more refined classification scheme for meme.

In the future, we plan to extend this work to other multimodal feature extraction methods to improve training on specific data sets. In addition, social media trends and patterns are changing rapidly, so it is necessary to capture memes in real time with respect to a particular domain so as to find the influential entities. This work can be extended to collect this data in real time and train deep learning models to identify hateful memes.

REFERENCES


[4] Shervin Malmasi, MarcosZampieri “Detecting Hate Speech in Social- Media”


