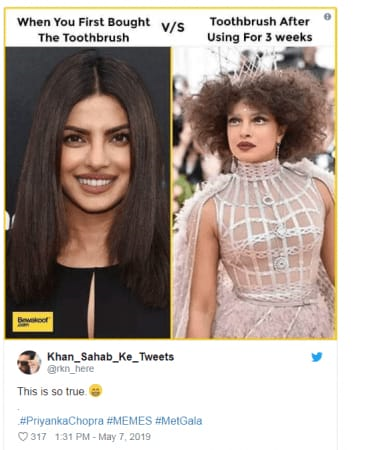
**SRS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **1.** | **Introduction** | | | | |
|  | 1.1 | Overview of the project | | | |
|  | 1.2 | Motivation | | | |
|  | 1.3 | Objectives of the project | | | |
|  | 1.4 | Literature survey (for identifying existing solution) | | | |
|  | 1.5 | Problem definition | | | |
| **2.** | **Proposed System** | | | | |
|  | 2.1 | Description of proposed system with simple block diagram | | | |
|  | 2.2 | Description of Target users | | | |
|  | 2.3 | Advantages/applications of proposed system | | | |
|  | 2.4 | Scope (Boundary of proposed system) | | | |
| **3.** | **Software Requirement Specification** | | | | |
|  | 3.1 | **Overview of SRS** | | | |
|  | 3.2 | Requirement Specifications | | | |
|  |  | 3.2.1 | | Functional Requirements | |
|  |  | 3.2.2 | | Use case diagrams | |
|  |  | 3.2.3 | | Use Case descriptions using scenarios, strictly as per Pressman Template | |
|  |  | 3.2.4 | Non Functional Requirements | | |
|  |  |  | | 3.2.4.1 | Performance requirements |
|  |  |  | | 3.2.4.2 | Model Requirements |

**1. Introduction**

* 1. **Overview of the Project**  
      A meme is an idea, behavior, or style that becomes a trend and spreads by means of imitation from person to person within a culture and often carries symbolic meaning representing a particular phenomenon or theme.  
        
      Over the past years, memes have been widely common all around the internet and especially in social networks.Memes have been shown to serve many purposes, such as entertainment and marketing. However, the wide usage of memes does have a negative side to it, as there are memes that combine certain texts and photos in ways that are being used to hurt a person's or group's feelings, we will refer to those as "Hateful Memes" (HMs).  
       
      Since those HMs could potentially be harmful as they may be used for cyberbullying and incitement, it would be very beneficial to be able to identify a HM on social networks, thus enabling us to block or report the HM. However, at the massive scale of the internet, the task of detecting multimodal hate is both extremely important and particularly difficult.
  2. **Motivation**



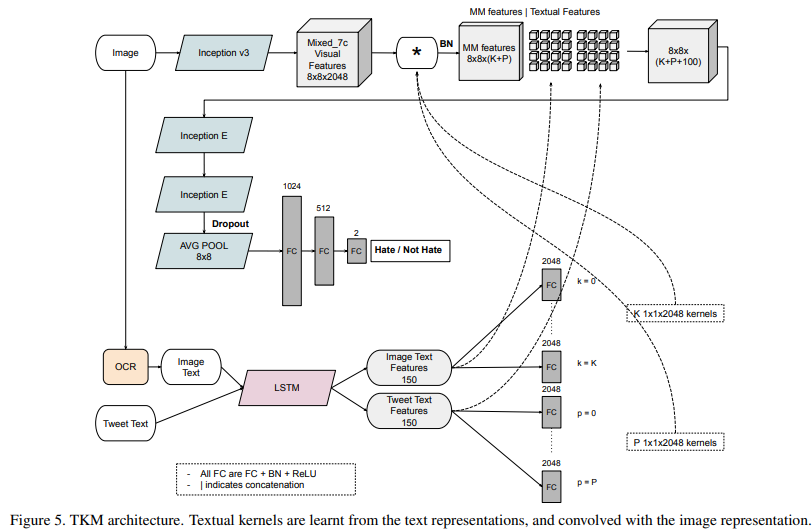


At the scale of the internet, malicious content cannot be tackled by having humans inspect every data point. Consequently, machine learning and artificial intelligence play an ever more important role in mitigating important societal problems, such as the prevalence of hate speech,trolling done over a group or person ,spread of fake news etc, that is been done with the use of memes.

It is very difficult to classify a meme as hateful or non hateful just on the bases of text alone or image alone. Hence there is a need of multimodal architecture which can consider both text and image and give a label to the meme as hateful or non hateful on the basis of the relationship between the text and the image.

* 1. **Objectives of the Project**
* To generate embedding of image and text .
* To fuse the text and image embedding and build a classifier.
* To evaluate the performance of the multimodal architecture.
* To compare proposed architecture with state-of-art technique.  
  1. **Literature Survey**

1. **Exploring Hate Speech Detection in Multimodal Publications ( WACV, 2020)**



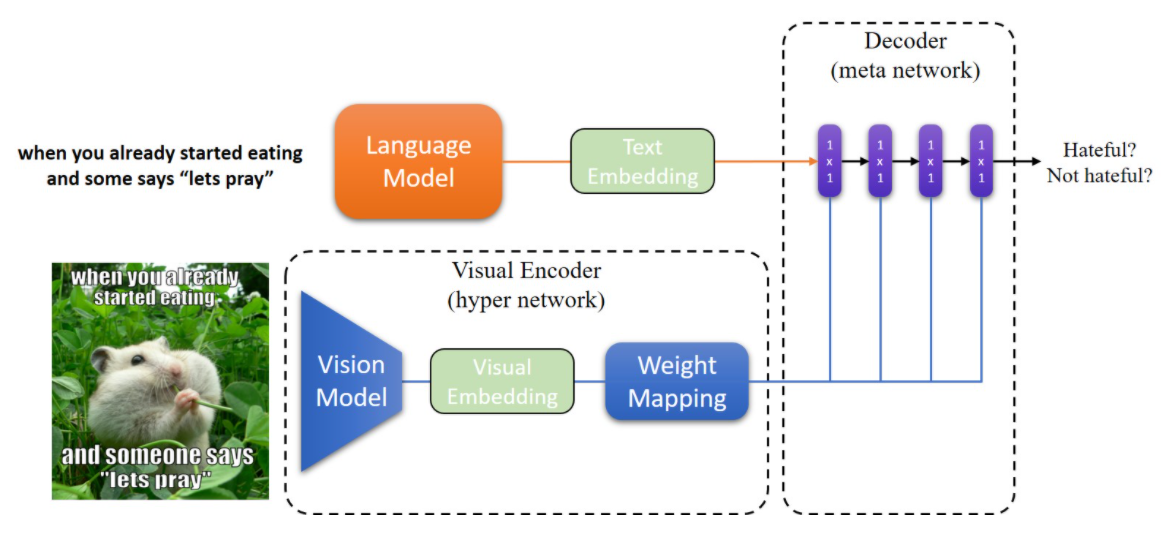
TKM is an architecture which is designed to capture the interactions between the two modalities ( image features and text features ) more expressively .

Language Model :To derive text feature vector or embeddings, LSTM( Long Short Term Memory) model is used, which will help in learning the complete text representation and followed by a Kernels for better representation of the text embeddings.

Visual Model :To derive image feature vector or embeddings , Inception v3 architecture is used which is a pretrained visual model that helps in learning a better feature representation of image

Then combining visual feature embeddings with text feature embeddings, resulting in multimodal feature embeddings. The multimodal feature embedding is then further processed to classify meme as hateful/non-hateful.

1. **Hateful Memes Challenge ,Yuval Nirkin , Assaf Rabinowitz ,Yoni Solel   
    2020, SEP 22**



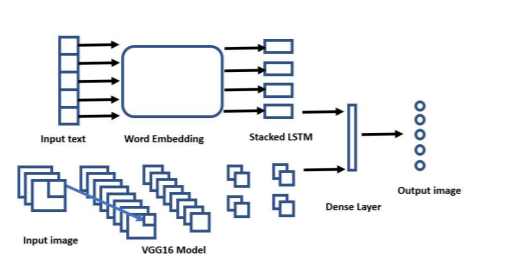
Language Model:  
In this model, text is preprocessed using SBERT, which was published in 2019. SBERT is a modification of the pretrained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embedding.

**Vision Model:** The vision model is based on the Mobilenet V2 architecture, as it is both fast and memory efficient and is well suited for limited hardware. Given an input image  the vision model outputs an embedding vector which is then fed to the weight mapping module.

**Weight Mapping**The weight mapping module receives an embedding vector as input and outputs a set of weights matrices W1,W2,…,WnW1,W2,…,Wn corresponding to the n linear layers in the decoder. The weight mapping module consists of multiple linear layers each receiving a part of the input vector that is directly related to the number of parameters required by the specific decoder layer, the more parameters the larger the relative part of the input vector that will be received.

**Decoder**The decoder is a meta network built as a multilayer perceptron (MLP), it receives its weights from the visual encoder, allowing it to be specialized for each input image. Each linear layer is followed by a batch normalization layer and a ReLU activation layer. Given a text embedding from the language model, the decoder outputs a binary prediction, hateful or not hateful.

1. **Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text.**Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Paul Buitelaar

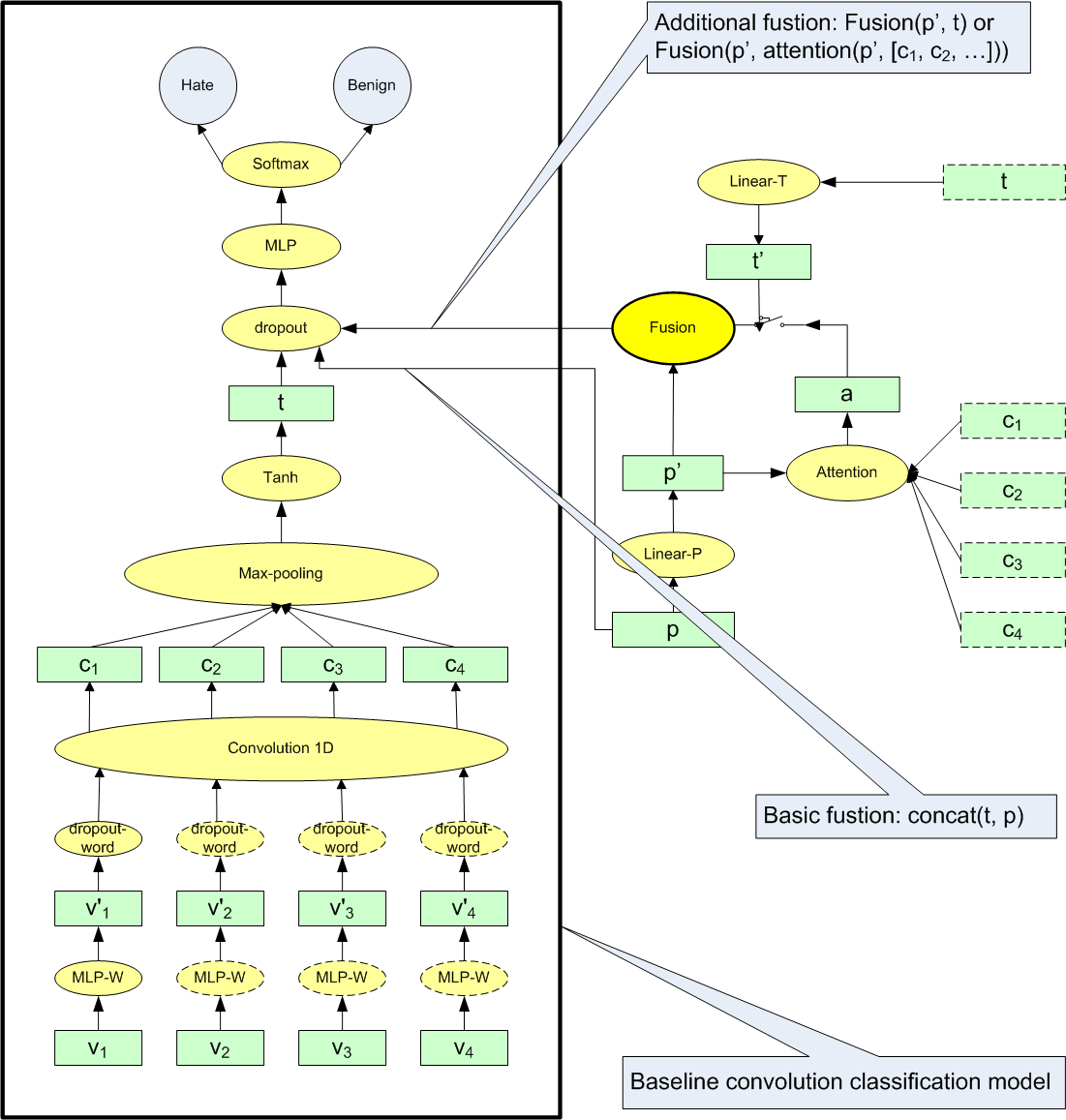


Baseline Models for Textual DataIn this model, first the text is pre-processed using Stacked LSTM .A bag of word approach of treating each word as a separate unit does not preserve the context of the word. LSTM has been used to extract the text feature. It saves the relevant information from the text which could be used later without facing the issue of vanishing gradient descent. In this approach, two LSTMs are stacked together. A stacked LSTM has the capability of building a higher representation of the data. As the output of an LSTM layer has been fed as input into the other. In the architecture for this baseline, stacked LSTMs are used as feature extractors before the data is being sent to the classification layer. Word embeddings are created using a pre-trained GloVe dataset. The use of pre-trained word embedding leverages the contextual meaning of the word globally.

Baseline Model for ImagesA CNN architecture developed by the Visual Geometry Group (VGG) at the University of Oxford has been used to classify the targeted image data (Simonyan and Zisserman, 2014). This specific architecture has 16 layers and is known as VGG16. The model is pre-trained on the ImageNet dataset and has been used as the baseline in our experiments. Images were loaded into an array and changed into a fixed shape as per VGG16 specifications. All the values in the matrix were in the range between 0 and 255.

VGG architecture has two convolution layers both with Relu as an activation function. The output of the activation function has been fed to the max-pooling layer which later has been followed by a fully connected layer which also uses “Relu” (Wang, 2017) as an activation function. Instead of a fully connected layer, a Global Average Pooling layer has been used which later is connected to a Dense layer with the Sigmoid activation function to predict class probability.

Multimodal ApproachEarly Fusion method is used as fusion technique.the text and image modalities in their vector form have been fed into the classifier. In this architecture, both modalities are required to classify the offensive content. A new vector has been formed by the concatenation of both modalities which represents a meme as a whole and hence can be used for classification as hateful or not.

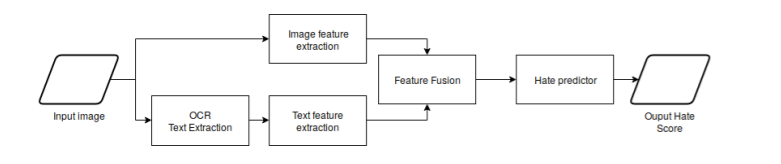
**4. Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification**

This architecture uses Convolutional text model for text classifier with various fusion methods for combining both image and text features.  
  
Convolutional text model   
For each word in a piece of text, the pre-trained embeddings [v1, v2, ..., vn] is retrieved. Then using MLP-W new word embeddings [v’1 , v’2 , ..., v’n ] which helps model to be more robust against word embeddings features. Then pretrained CNN(Convolutional Neural Networks) will generate context embeddings[c1,c2…cn] for text and finally generating text embeddings(t)  
  
Visual Model   
Pre-train deep neural network on millions of photos on a social network platform.Then use those learned features, to extract features for new photo of hate domain.

Attention Model  
 The idea of attention is to use the information of a vector (called query) to weighted-sum a list of vectors (called context) where the context vector [c1, c2, ..., cn] is from text, while the query vector is the photo vector p’( photo features ). As a result, this model will focus on specific aspects of both textual and photo features to generate a attention vector ( a ).

Fusion  
 Gated Summation approach is applied to text and photo features fusion, which helps in dynamically controlling the combination of visual and textual features.

**5. Multimodal Hate Speech Detection in Memes,Benet Oriol S`abat**



OCR - Pytesseract

The first step is to extract the text from the image. This task is called Optical Character Recognition (OCR) and there are several implented systems that detect text in an image and extract it. One of the most popular and usable is Tesseract5 . We will treat this as a text detector black box, but we want to mention that this library actually uses state-of-the art Deep Learning algorithms to extract the text. We used the pytesseract Python wrapper6

Text embedding - BERT

The text detected by the OCR is encoded in a BERT (Bevlin 2018 ) embedding, a state of the art model that has been proved to perform very well in different tasks. This model turns a (sub)word sequence into a 768-dimentional feature vector. It is basically a transformer architecture and is pre-trained in a self-supervised way, which is a kind of encoder-decoder architecture.

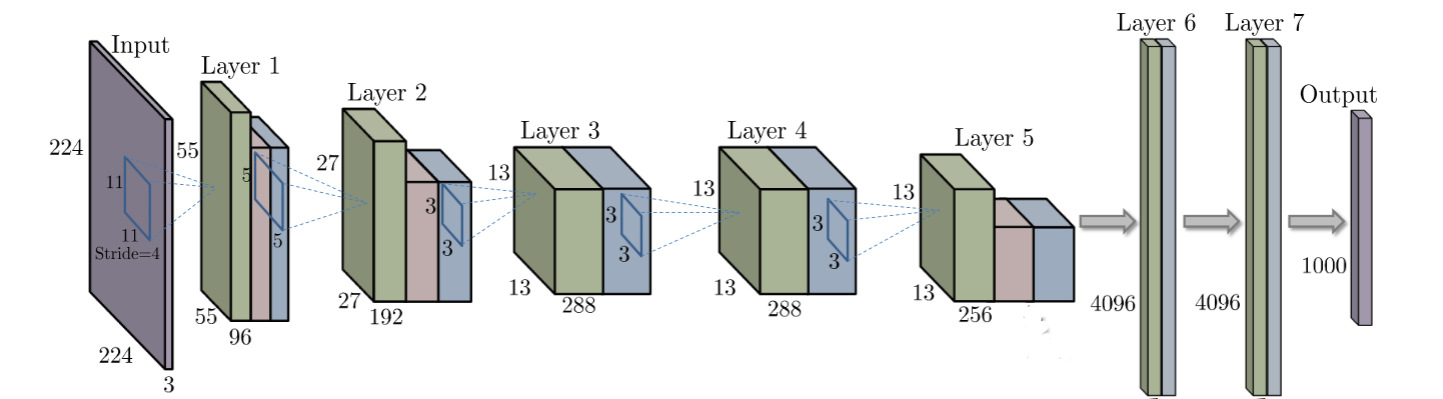
Encoder-decoder architectures have been used as the top performing models for sequence to sequence tasks such as machine translation. The encoder maps an input sequence of symbol representations (*x*1*, ..., xn*) to a sequence of continuous representations *z* = (*z*1*, ..., zn*). Given *z*, the decoder then generates an outputssequence (*y*1*, ..., ym*) of symbols one element at a time. At each step the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next.

LSTMs (Schmidhuber 2005 [16]) have been used for a long time to model both encoder and decoder. However, the state-of-the art is currently the Transformer model.

Image embedding - VGG16

VGG are (Simonyan 2014 [32]) are a set of image classifification architectures proposed for the first time by the Visual Geometry Group in the university of Oxford. The architectures in the VGG-N models differ in the number on layers - which is N - but have a very similar architecture.All of them have convolutional layers at the beginning and end up with a fully connected network

for classification.



Convolutional Neural Network architecture. Purple represents the input 3 channel matrix and the output vector. Each block is a layer, consisting of the vectors after convolution, max pooling and non-liner activation in green, red and blue, respectively.

* 1. **Problem Definition**

To develop multimodal architecture which classifies the memes as hateful or non hateful.



Multimodal Architecture

Non Hateful

Output

Input

1. **Proposed System  
    2.1 Description of proposed system with block diagram**



Input

Output

Hateful/Non-Hateful

Visual Module  
(Inception v3)

Text Module  
(Sbert)

Decoder

Performance Evaluation

She hates Jews but she didn’t mean to say she hates Jews

* Our system will have two inputs : Image Text, Image
* Text module [ Sbert ] will process the text to extract or generate text embeddings. Similarly Visual module [ Inception v3 ] will process the image to generate image embeddings.
* In Decoder module, text embeddings and image embeddings obtained from text and visual module are fused and fed to the classifier, which will then classify meme as hateful or non-hateful.
* Once the classification is done, performance evaluation is carried out to check our system accuracy.

**2.2 Description of Target Users**

Target Users for our system are majorly **Social Media** platforms. These platforms require detecting hateful speech as there are multimodal data( For ex : image, text ) involved.

**2.3 Advantages/Applications of proposed system**

1. This system will help in detecting multimodal hateful memes .
2. This system can be used in controlling the hate content that is being posted on internet every minute.
3. This system can be incorporated as an gateway in many Social Media platforms , to monitor every meme before it is been posted.

**2.4 Scope(Boundary of proposed system)**

1. This system will deal with only images( meme ) with English text in it.
2. This system does not uses OCR, hence the embedded text will have to be provided in the input along with the meme.
3. This system will be efficiently able to generate an accuracy of 70% .

**3. Software Requirement Specification**

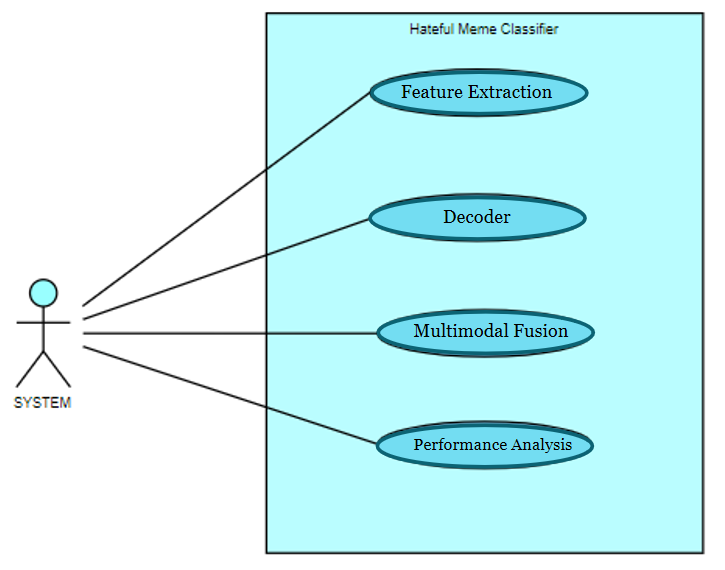
* 1. **Overview of SRS**

1. Our system models are required to be pre trained first.
2. The image and text model are pre training , generates the embedding
3. The fusion module clubs these embedding and generates a feature vector which is generated considering the relationship between both the modal(image and text)
4. The classifier then labels the the given input meme as hateful or non hateful .
5. The framework used for the model is Keras upon Linux platform.
   1. **Requirement Specifications**

**3.2.1 Functional Requirements**

1. The system shall be able to generate image and text representations .
2. The system shall be able to classify the meme as hateful or non hateful using provided embedding .
3. The system shall be able to generate the probability of the meme being hateful .
4. The system shall be comparable to the state of the art techniques.

**3.2.2 Use Case Diagram**



**3.2.3 Use Case Descriptions**

Feature Extraction : In this process we generate the embedding for the text and the image that is provided in the form of input .

Multimodal Fusion : In this process we club the obtained image and text embedding and generate a feature vector which represents the features of both the image and text.

Decode : This process takes the feature vector and then tries to generate a label for the provided label in terms of hateful or non hateful

Performance Analysis :This process is used to measure how accurate our model is, in classifying the meme as either hateful or non hateful.

**3.2.4 Non Functional Requirements**

* + - 1. The system should be at least 70% accurate .
      2. The system should not under-perform on limited computing resources.
      3. In building the system , the framework used should be Keras ,on Linux platform.