DEVELOPMENT OF MODEL FOR MULTIMODAL SARCASM DETECTION

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

MASTER OF TECHNOLOGY (M.TECH)
IN
COMPUTER SCIENCE AND ENGINEERING (CSE)

Submitted by
ADITYA CHAUHAN
2K20/CSE/01

Under the guidance of
Dr. Aruna Bhat
(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
NEW DELHI
June, 2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
NEW DELHI

DECLARATION

"I, **Aditya Chauhan**, **2k20/CSE/01** a student of M. Tech. in Computer Science and Engineering declare that the Project Dissertation titled "**Development of Model For MultiModal Sarcasm Detection**" that is submitted to the Department of Computer Engineering, Delhi Technological University by me, in partial fulfillment of the requirement for the award of the degree of Master of Technology is an original work and has not been copied from any source without proper citation. This work has not previously formed the basis for the award of any Fellowship, Degree, Diploma Associateship, or other similar title or recognition"

Place : New Delhi
Date: 27/5/22

Aditya Chauhan
2k20/CSE/01

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

NEW DELHI

CERTIFICATE

This is to certify that the Project Dissertation titled “**Development of Model for MultiModal Sarcasm Detection**” which is submitted by **Aditya Chauhan** of M.Tech , Computer Science & Engineering, Delhi Technological University Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place : New Delhi
Date : 27/5/22

Dr. Aruna Bhat
Associate Professor
Department of CSE
DTU

ACKNOWLEDGEMENT

It gives me great pleasure in expressing my deep gratitude to my mentor **Dr. Aruna Bhat**, Associate Professor, Delhi Technological University for believing in me and providing me with this opportunity to pursue and complete the dissertation work to be submitted in partial fulfillment for the award of degree of M.tech in Computer science and Engineering. I am truly indebted to her for her guidance, suggestions and the support during the entire course of the project. I want to thank her and express my gratitude for guiding and motivating me at every step of this great journey. Her genuineness, diligence, and dedication have always been an inspiration to me. My ambitions have only been made possible by her conscientious efforts.

I am also grateful to Mr. Raju Kumar, who has guided and helped me during the entire process by providing me with valuable inputs and feedback on my work. I also don't want to pass up the chance to acknowledge the contributions of all faculty members in the department for their kind assistance, cooperation along with valuable guidance and feedback given during the initial stages of my project. Also, finally I would like to thank my friends Abhishek, Rajiv and Ankit for their contribution and all the help they lended me that eventually led to the dissertation's completion.

Aditya Chauhan
2k20/CSE/01

ABSTRACT

Sarcasm detection is used to single out natural language statements where intended meaning differs from what the surface meaning implies. A number of tasks in natural language processing areas like analysis of sentiments as well as mining of opinions use sarcasm detection underneath. Many of the key research in the area of sarcasm detection primarily focus on only text-based input. In the present day scenario , there has been a sudden explosion in the amount of multimodal data mainly due to social media. As a result of that, users these days are not just limited to text while expressing themselves , but also make heavy use of visuals like in images and videos. The objective of this research work is to incorporate multimodal data so as to enhance the performance of present sarcasm detection algorithms. Multiple methods that leverage data in the form of image and text, and also as a combination of both, have been presented thus far.We present a unique architecture which works on the Robustly Optimized BERT pre-training approach or RoBERTa which is nothing but facebook modified version of well known model BERT having co-attention layer on top for including the incongruity in the context between attributes of the image and input text. Using Gated Recurrent Unit(GRU), the text-based features are extracted and the features of the image are retrieved by conditioning the image through Feature-wise Linearly Modulated ResNet Blocks. These combined with CLS token from the model RoBERTa are used to obtain the final predictions. In the results, we show performance enhancements from the proposed model when it is stacked against the other latest models with competitive results which make use of the same publicly available Twitter dataset.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

VGG	Visual Geometry Group Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
OCR	Optical Character Recognition
GRU	Gated Recurrent Unit
SVM	Support Vector Machine
NLP	Natural Language Processing
LSTM	Long Short Term Memory
ESD	Emoji Sarcasm Detection
RoBERTa	Robustly Optimized BERT pre-training approach
FiLM	Feature Wise linear modulation

CHAPTER1: INTRODUCTION

1.1 OVERVIEW

A phenomenon in which the actual meaning comes out totally different than the literal meaning can be referred to as Sarcasm. Sarcasm can be used as a way of poking fun, taking jibes and in some cases to insult someone too. Today social media is growing at a commendable pace and so is the sarcasm on these platforms. Not all posts and messages have straight-forward meanings on social media as these days people have become quite fluent at speaking sarcasm. When sarcasm is difficult to decipher even for us human beings owing to the contextual understanding it requires in order to comprehend it, sarcasm detection making use of machines or algorithms is certainly going to be a very tricky task. Detecting sarcasm or understanding the underlined meaning is one of the key requirements that needs to be met as the speaker's actual message needs to be known in a number of different applications like opinion mining, product reviews, sentiment analysis etc. Traditionally or in the earlier researches, this has been limited to only the text based methods and was mainly treated like a problem of text categorization. Some signs based on lexicals, language markers and some frequent utterances were being emphasized on inorder to detect sarcasm out of texts. Besides that on social media platforms, hashtags and emoticons have a significant part in sarcasm detection as they are likely to provide necessary hints and clues for that purpose. The text-only methods fall short when it comes to understanding the contextual part in order to completely get the intended meaning. Moreover with the explosion in the amount of data and the ever popular social media, it just isn't enough to solely focus our attention on texts as messages are being conveyed using multi-modalities these days. This can be seen on twitter, instagram, facebook and also on other social media sites where users post more with images along with the texts. At times the image is completely opposite of the caption specified along with it and that is where sarcasm can be detected considering both the modalities and not just the one based on texts. So it is not enough to focus only on text based modality for sarcasm detection as real meaning might be conveyed by the stark contrast between the image and the text. Moreover in the motivation section, some examples are being shown where the image and text have completely different meanings and that shows the reliance on other modality as well as help in the detection of sarcasm.

1.2 MOTIVATION

Only considering text based modality will not exactly help when the other modality conveys the exact opposite of what is in the text. That is where we realize that there is a dire need to incorporate other modalities too for the purpose of sarcasm detection. This can also be explained with the help of below examples .

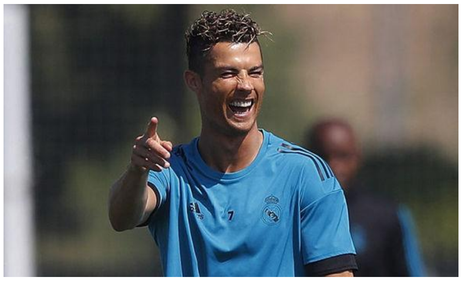


Fig.1.1 “The 7th Ballon d'Or was meant to be Messi’s”

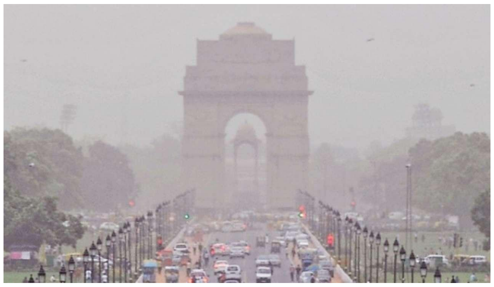


Fig.1.2 "Delhi winters are an absolute heaven"

Take the case depicted in Fig.1.1 which says “The 7th Ballon d’Or was meant to be Messi’s” which on reading seems to be conveying the positive sentiment as if the speaker truly believes that it is Messi who rightly deserved it. But on looking at the other modality, which is the image, it is clearly visible that someone is laughing while speaking the same which leads us to the conclusion that it is being made fun of or in other words the sentence has been said sarcastically. The complete meaning can only be found out when we have considered both the modalities thus signifying the importance of other modality in case of sarcasm detection. Similarly in Fig.1.2 , the caption reads “Delhi winters are an absolute heaven” which conveys a positive sentiment if we only look at the text but when we consider image modality also, it shows the foggy weather of Delhi winter which signifies that it is very polluted and hence can be considered nowhere close to a heaven. Thus this example also conveys that there is a need to consider the other modality.

In a similar manner, when we read caption from Fig.1.3 , it says “Working on weekends too is fun" which conveys the message that the concerned person is really hard working and likes to work even on the weekends, but on looking at the picture we get a clear idea that the person simply doesn't want to work on weekends, in fact he hates it which is a general idea as nobody usually wants to work on the weekends.

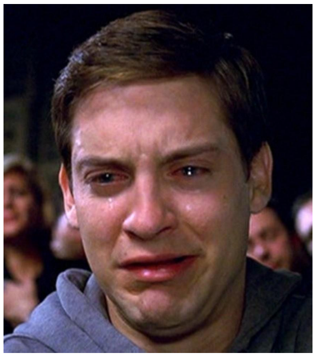


Fig.1.3 "Working on weekends too is fun"

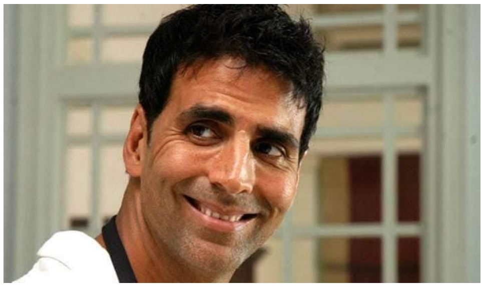


Fig.1.4 "Nobody can cancel our Leh-Ladakh trip"

Also from Fig.4's caption which says "Nobody can cancel our Leh-Ladakh trip", we seem to get a sense of belief that this time the trip will surely happen but on looking at the picture, it can be seen that there is an element of mischief over here and the trip can still be canceled due to any reason or any particular person. In other words it simply conveys the message that the trip has been canceled so many times that it is now very difficult to say with certainty that it will take place no matter what. This is clearly an example of sarcasm where it has been said in order to take a jibe at listeners who likely are the responsible ones for canceling the trip earlier.

In all the above examples we have observed that merely relying on the text doesn't actually serve the purpose of sarcasm detection as some important message is likely to be conveyed by the other modality. Thus in the age of multimedia and social media, it doesn't make sense at all to ignore the other aspect of data and solely focus on the textual part. Thus there is a dire need to consider image modality too in our methods for detecting sarcasm. Moreover sarcasm is widely common these days on online social media sites and can be used for numerous applications. These mainly include advertising, opinion mining, hate detection to name a few. The use of sarcasm was limited to humor or some light hearted harmless fun but these days it is taking drastic avatars like cyberbullying, hate speeches, racism etc. Thus to counter these issues, there is a need for sarcasm detection on online social media. Thus inspired by the degree to which sarcasm detection can make an impact in making social media better and the role of multimodal data in today's social media world lead us to go in depth and research more in this area.

1.3 PROBLEM STATEMENT

Traditionally sarcasm detection finds analogy with the problem of text categorization and moreover it has been treated like one. A lot of emphasis has been laid on the language constructs and other identifiers in the sentence for this purpose. Some of the language markers and lexicals have been used in the past for the same purpose as the sole focus was identifying sarcasm through text based patterns. Thus only one modality, that is the one based on texts, has been considered overlooking the other modalities. But with today's data explosion due to social media sites, relying solely on the texts is not enough as some of the key information might be conveyed through other forms of data like images.

Thus it is the combination of multiple modalities that should be put into use for the purpose of sarcasm and not just the text based one should be emphasized anymore. The main focus should be to make the best possible use of modalities available for sarcasm detection as all the modalities combined convey the exact message of the speaker or author. A key aspect that we found to be missing after our research is the interaction between different modalities. Most of the research focuses on extracting embeddings from different modalities using various deep learning based methods and combining them usually by concatenation to feed it to the fully connected neural network for the purpose of prediction. In our work we limit our work only to text and image based modalities. When text and image convey the meaning that is in stark contrast to each other we call it incongruity. Incongruity is the basis of finding the relationship between different modalities in our work and this relationship is the key aspect found to be missing in our research so this is the main area we will be focusing on. In this work , a method based on deep learning considering the text and image modality is proposed to capture incongruity and ultimately serve the purpose of sarcasm detection.

CHAPTER2: RELATED WORK

A certain level of understanding is required between the author/speaker and the receiver as it is considered to be a very contextual event[1]. Bamman in [1] leveraged background information about the post's authors , audience's relationship with him/her and the kind of situation they currently find themselves in so as to achieve better accuracy for prediction. In [2], a method is proposed by Joshi which is based on text incongruity, mainly between different sentences for the purpose of sarcasm detection. Khattri provides empirical proof in [3] that the past tweets of an author can give more information for identifying sarcasm. They utilize the author's previous emotion on entities constituting a tweet to figure out if it's sarcastic. To identify the sarcasm in messages or posts on social media like Twitter, Wang in [30] deploys a context-based technique that employs dialogues, such as a series of tweets. Schifanella basically uses convolution on hand-selected features as well as on deep learning-based features from the images as well as texts to create predictions utilizing two different modalities in [5]. Numerous research based on multimodal sentiment analysis (Wang in [31]; [32]) focus on video inputs, where multiple data formats such as text, image, and voice mix nicely.

Poria and others [21] use several kernel learning to mix various modalities. Zadeh in [32] built a layer for fusion employing outer product instead of just basic concatenation to add some additional characteristics. Gu employs a number of attention strategies and blends text and audio for each word in [33]. Gu presented an approach centered on fusion of modalities in [10] to illustrate the underlying value of having several modalities.

Schifanella was the first to propose two strategies for detecting sarcasm using multimodal aspects of data[5]. One that combines or, to be more exact, concatenates textual as well as semantic attributes retrieved from external dataset training. The other just leveraged a few deep learning features through ImageNet which is trained on image parameters and also the text features passed through a visual neural net incorporating both image and text modality for prediction. Gajarla [11] evaluated many classification algos, like SVM, on multiple high level characteristics retrieved with VGG-ImageNet using their own dataset. In [34], Y. Tay introduced a new multi-dimensional model for modeling incongruity that focused on intra-attention. In [22], Wu combined sentiment based features, embeddings, and syntactic information to develop a multi-tasking model by making use of an LSTM which is densely connected. According to the authors of [14], CNN may be used to encode and represent characteristics collected from data generated by the movement of the eye or gaze. The authors of [19] used an RNN or Recurrent Neural Network architecture for texts. A model was suggested by them that can process data from a variety of sources, including image transcripts, images, and texts. 'VGG-16' is used to retrieve features from the image, whereas contextual data is obtained with the help of Bi-GRU.

In [12], authors say that sarcasm detection may be considered like a binary classifier's task. They hypothesized that rather than manually picking features for the purpose of feature extraction, they could use the components that are visually reasoned or the image cues like an input for CNN in order for sarcasm detection. In [8], the authors propose a fusion-based model that takes into account attributes of an image, features of an image and the text based features as three distinct modes. The image features, attributes and text based features are represented in the form of vectors by extracting them using a pre-trained ResNet and bidirectional LSTM respectively. WIth the concept of weighted average these vectors are then fused into one single vector rather than just simple concatenation and then for prediction purposes passed on to a fully connected neural net. The authors in [9] used BERT and ResNet models for encoding the textual and image data respectively and used a gate which they called bridge. They proposed a two-dimensional intra-attention layer for extracting the relationship between the image and text. Our work is mainly inspired by [8],[20] but there is also resemblance with [7] where BERT is used for modeling both intramodal and intermodal incongruity by making use of co-attention and self-attention mechanisms. Intra-modal incongruity also finds its use in [9].

2.1 DATASETS USED

A small account of various datasets picked or compiled by the authors in their research is provided in this section along with the details of any new dataset being compiled by the researchers for their work.

2.1.1 Instagram, Tumblr and Twitter Dataset

In [5] ,authors compiled the dataset using the publicly available APIs for the three platforms : twitter, instagram and tumblr. To create this dataset, 10000 posts were chosen at random from all three platforms considering positive samples, or those having sarcasm related tags or content. Tags as well as the links of any kind were deleted from the entries that included both visual as well as text components. In the textual section, hashtags along with emojis were considered in addition to regular words since they lead to the better understanding of every message by revealing the genuine sentiment engulfed in the post, whereas conventional words simply tell only the semantic meaning. Cues from the context play an important role in decoding sarcastic parts. Because there is a word restriction for messages on Twitter and Instagram, shorter sentences or phrases are commonly seen on them, combined with usage of emojis as well as hashtags, which constitute the actual core of sarcasm. Meanwhile, Tumblr favors lengthier texts over emoticons. Similarly, 10000 negative samples were taken into account for the same.

2.1.2 Yahoo Flickr Sarcasm[YFS] Dataset

The authors of [12] used the self-annotated data notion. Samples are collected using the tags that individuals use on Flickr. Besides the keyword "sarcasm" , some similar keywords are also considered like 'satire', 'irony' to name a few. An image is somehow having both the sarcasm as well as non-sarcasm tags, that is straightaway discarded. In the practical world, sarcasm is rare and this is evident from this dataset too. Merely 443(23.99 %) out of the total 1846 images collected here belonged to the sarcastic category while 1403(76.01 %) belonged to the opposite one.This dataset is broken into 90: 10 respectively for the training and validation. The authors called it Yahoo Flickr Sarcasm Dataset or YFS.

2.1.3 Twitter + Reddit Dataset

The training dataset used in [13] comprises 5000 tweets, 4400 reddit comments labeled automatically(Ghosh[24], Khodak[23]). There is some context for each response or comment, in the form of a list, an ordered list, consisting of all the previous comments in the whole conversation, as well as a label designating the response into sarcastic or non-sarcastic category. The test data contains 1,800 tweets and 1,800 Reddit comments. Similar to training data, the test replies too have their very own conversational context and are balanced.

2.1.4 Facebook+Twitter Dataset

In [25] authors used Twitter for collecting data over a 5-month period as well as from Facebook between 2015 to 2017. Popular sarcastic facebook pages such as 'sarcasmLOL' and 'sarcasmBro' were referred to, while tweets having hashtags 'sarcasm' as well as 'sarcastic' were evaluated for collection of data on Twitter. Web scraping was used to gather the data, which was then preprocessed before being reduced to text and emoji. Special characters, URLs, hashtags, and retweets were among the items removed from the data during preprocessing.

2.1.5 MUStARD Dataset

The authors compiled a new dataset which they called MUStARD in [17], which contains short clips tagged with sarcastic labels manually. The clips are picked from TV shows famous for their sarcastic nature like Big Bang theory ,Friends, Sarcasmaholics Anonymous and The Golden Girls. To get non-sarcastic clips, we made use of a collection of 400 videos from MELD, which is nothing but a multimodal emotion recognition dataset derived from Friends which was a popular TV series of the 90s and early 2000s, courtesy of Poria. So, from the 6,365 annotated videos, a balanced sample is obtained. The dataset contains 690 clips having sarcastic and non-sarcastic tags in the same numbers. Every utterance, including its context, has three modalities: audio, video and transcription (text).

2.1.6 Instagram Dataset

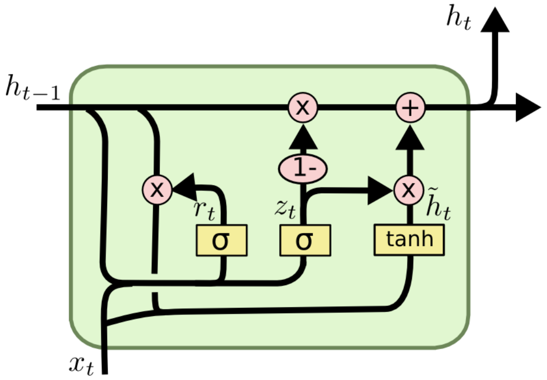
The authors in [19] enhanced the silver dataset available to them to what they refer to as gold-standard dataset courtesy of randomly picking positive cases (having sarcasm) counting to

1600. Then they asked the different annotators to label the instances by just text modality, later the examples were finally labeled based on the majority. In the upcoming round, they showed both the image and also the corresponding text and asked them to relabel them as sarcastic or non-sarcastic.This way a more accurate dataset with correct labels is thus compiled.

2.2 METHODS USED

2.2.1 RNN

For text, in [19] authors adopted an RNN (Recurrent Neural Network) based structure. From an input, they suggest a model in which multimodal data (image, text, as well as image's transcripts) is processed. From the text, every word along with emoji is stored after preprocessing in the form of a feature vector of 100 dimensions using Glove. Each word from the completely preprocessed text is then passed through a bidirectional GRU or Gated Recurrent Unit. Bi-GRU captures the context based information and finally only a single output representation is obtained for all the input words. If some text is present in the image, it is extracted by making use of Optical Character Recognition and later fed to bidirectional GRUs for every word extracted leading to only a single output representation or else for that modality a zero vector having dimensions (1*1000) is used. With the help of VGG-16 which used ImageNet[11], feature vectors with dimensions (1*4096) are obtained corresponding to every input image. In the end, all the three modalities' features are then fed on to a fully connected layer which keeps the output dimensions the same.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Fig 2.1 Gated Recurrent Unit

2.2.2 Automated methods

Two automated approaches for identifying multimodal sarcasm are investigated by the authors of [5]. One of them is a linear SVM or Support Vector Machine approach that is widely employed in earlier research, but it is heavily dependent on the text extracted features from the post or a set having many posts. Visual features are first extracted from the images and then fused with a number of natural language processing elements which they call Fusion with SVM. The other approach makes use of unigrams obtained from the combination of textual input and visual representation based on deep network adaptation , with the technique being called Fusion with Deep Network Adaption. In both of these techniques, they evaluate the separate contributions from visual and textual features, as well as their combination.

2.2.3 CNN based method

In [12], authors treated the task of sarcasm detection similar to the task of binary classification. They didn't do the feature detection part as feature extraction from pictures is difficult, and picking excellent features that are hand-crafted might result in a bad classifier. Instead they used a convolutional neural network to solve this problem (CNN). They suggested that feeding visual components or cues into CNN could properly identify sarcasm. A CNN contains many convolution layers, generally separated by a sub-sampling step, along with one or more totally connected layers. This group of layers, which contains an activation layer with a relu function, a 2-dimensional convolution layer, along with a max-pooling layer which has pool size = (2,2) will be referred to as "Layer group A". The authors' neural network [12] consists of one flatten layer, 2 dense networks having relu as well as sigmoid functions, and three Group A layer. A tradeoff is inevitable when it comes to deciding the size of a batch for the purpose of training CNN. In the finalized model, a batch size of 16 was used. Also an additional relevant parameter in training of CNN is epoch count, which refers to the count for which samples for training pass through a cycle of backward as well as forward passes. In [12], the CNN model was trained using 50 epochs.

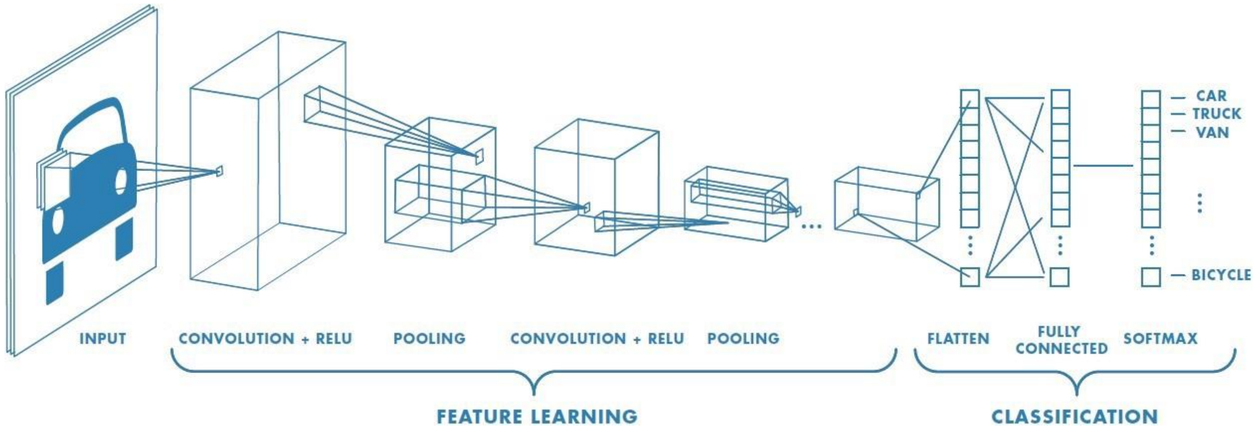


Fig 2.2 Convolution neural network

2.2.4 Ensemble Approach

In [13], for training the ensemble, from four models which are considered as components, expected probabilities of sarcasm are used along with other factors like length of the comment, sentiment associated and also the actual source so that component wise importance of input can be found. The component models include an LSTM having hashtag as well as emoji representation, a CNN-LSTM having punctuation, sentiment, casing and stop words representation, an Infersent embeddings based MLP, along with features from emotions and an SVM that is stylometrically trained. All component models incorporate 2 conversational turns before the response as context, with the exception of the SVM, which uses only the response's features. The ensemble contains an adaboost classifier that leverages the decision tree from the base classifier.

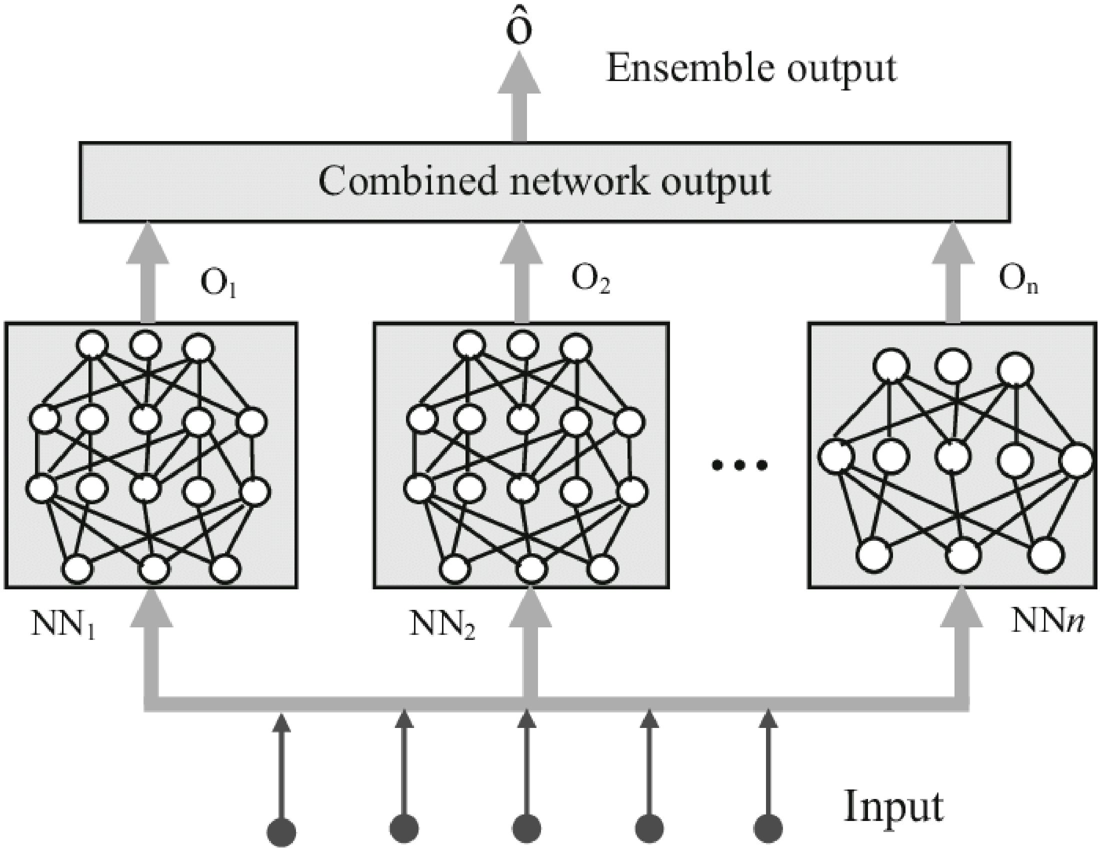


Fig. 2.3 Ensemble Neural Network

2.2.5 ESD Framework

In [25], a new model is proposed by the authors called ESD framework or the Emoji Sarcasm Detection framework for sarcasm detection. The model contains three main components namely an emoji encoder, one encoder for text and also a module that predicts sarcasm. The task of a text encoder is mapping words to their latent representations, here emoji encoder is responsible for extracting representations related to emojis and the sarcasm prediction component makes use of a classification function for the purpose of detection on social media posts. This system takes both the emojis and words as inputs in the form of vectors and converts them into corresponding emoji and word embeddings. The concatenated vectors then produced are fed to bidirectional GRU. The attention vector's appropriate weights are then multiplied, summed up using the vectors to produce the context vector, and then finally passed for classification to the sigmoid function.

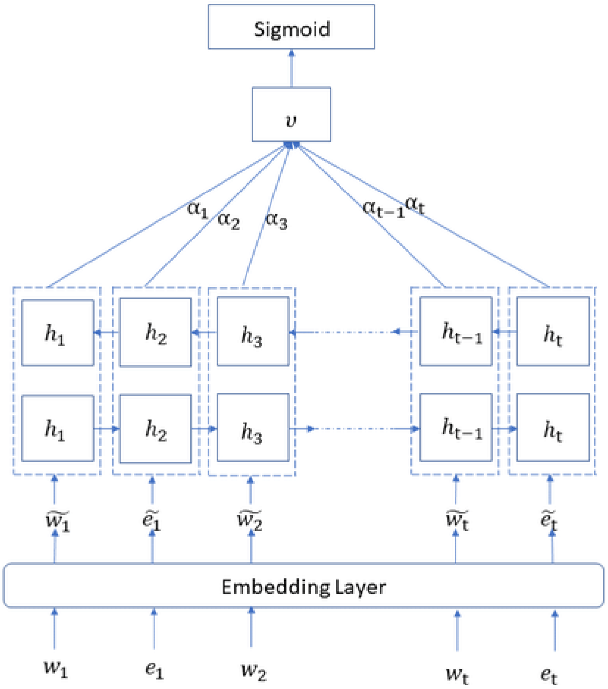


Fig. 2.4 ESD Framework

2.2.6 Fusion Method

In [8], the authors considered text, image features along with the attributes from image as different modalities thereby suggesting a hierarchy based multimodal fusion model. First of all features,attributes from the image are extracted and then text features are extracted using a bidirectional LSTM network. These features from all 3 different modalities are combined to a just one feature vector after they are recreated from each modality. The vectors corresponding to the text are created using Bi-LSTM in which raw text vectors are created by concatenating forward as well as backward hidden states corresponding to each time unit of Bi-LSTM, whereas the guidance vector for texts is obtained from the raw vectors' average. Then non-linear adjustments are applied for the attribute guidance vector, then output is passed to the Bi-LSTM in the hidden state. This occurrence is referred to as early fusion. Representation fusion is offered as a means to leverage data from several modes for refinement of representation from all the modalities, wherein feature vectors of all 3 modalities are reconstructed using guiding as well as raw vectors. In fusion of modality, the refined vectors from all the three different modalities are made into a single vector through fusion by weighted average instead of just basic concatenation.

At last, the fused vector is fed as an input into a two-layer fully-connected neural network for the purpose of classification.

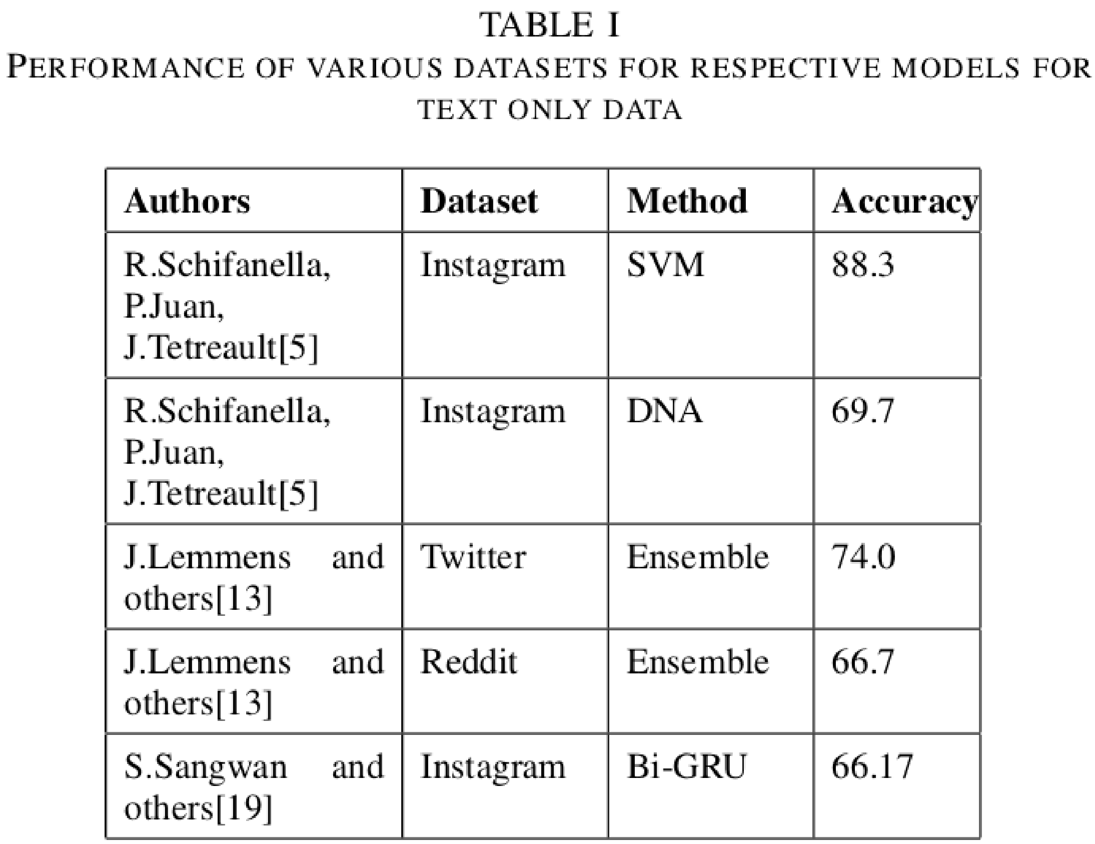


TABLE I
PERFORMANCE OF VARIOUS DATASETS FOR RESPECTIVE MODELS FOR TEXT ONLY DATA

Authors	**Dataset**	**Method**	**Accuracy**
R.Schifanella, P.Juan, J.Tetreault[5]	Instagram	SVM	88.3
R.Schifanella, P.Juan, J.Tetreault[5]	Instagram	DNA	69.7
J.Lemmens and others[13]	Twitter	Ensemble	74.0
J.Lemmens and others[13]	Reddit	Ensemble	66.7
S.Sangwan and others[19]	Instagram	Bi-GRU	66.17

TABLE II
PERFORMANCE OF VARIOUS DATASETS FOR RESPECTIVE MODELS FOR MULTIMODAL DATA

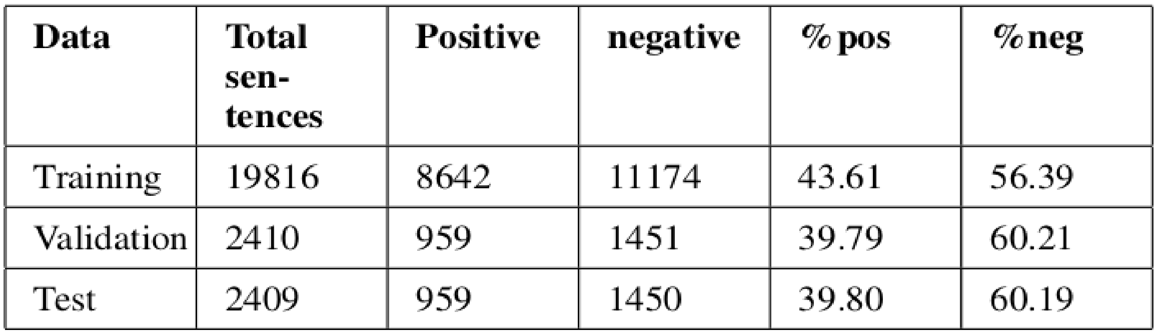
Authors	**Modality**	**Dataset**	**Method**	**Acc.**
R.Schifanella, P.Juan, J.Tetreault[5]	Text,Image	Instagram	SVM	88.5
R.Schifanella, P.Juan, J.Tetreault[5]	Text,Image	Instagram	DNA	77.8
D.Das and A.J.Clark [12]	Image	YFS	CNN	84
S.Castro and others[17]	Text,Aud,Vid	MUStARD	BERT,Deep Learning	71.5
Yitao Cai, Huiyu Cai and Xiaojun Wan[18]	Image	Twitter	Fusion	80.49
Yitao Cai, Huiyu Cai and Xiaojun Wan[18]	Text,Img,Img Attributes	Twitter	Fusion	83.44
S.Sangwan and others[19]	Text+Img	Instagram	Bi-GRU	70.0
S.Sangwan and others[19]	Text+Img+Img Transcripts	Instagram	Bi-GRU	71.5

CHAPTER3: PROPOSED METHODOLOGY

3.1 DATASET USED

We utilize the same publicly accessible Twitter dataset that the authors prepared and used in [8]. The dataset includes images, image attributes as well as tweets totaling 24k items. The table below shows the instance wise breakdown. The authors in [8] divided the dataset in the ratio 80:10:10 respectively to be used for training, validation, as well as test. We opt to use the dataset with the same distribution since it seemed to be a comparable situation to [8]. In the phase where data is preprocessed, by making use of NLTK toolkit, hashtags, emoticons and the words are segregated from the dataset. Because this appears to be the lone wolf when it comes to manual validation of the attributes of image , the findings have been reported solely on a single dataset that contains text, images, and image attributes. Even the images need to be resized to 224 so as to maintain uniformity, which was achieved by normalizing and center cropping the images. The emoji contribution is neglected after pre-processing. A variety of data augmentation techniques, like random contrast, brightness, image saturation as well as random center crop, were also utilized.

TABLE III Twitter Dataset At a glance



Data	Total sentences	Positive	negative	%pos	%neg
Training	19816	8642	11174	43.61	56.39
Validation	2410	959	1451	39.79	60.21
Test	2409	959	1450	39.80	60.19

3.2 ARCHITECTURE

Here a new model based on RoBERTa , a facebook customized version of a famous NLP model , BERT. RoBERTa or Robustly Optimized BERT pre-training approach is used here for detecting sarcasm. RoBERTa is far more advanced than simple BERT as it has been trained on larger datasets in comparison and also made use of very long training sequences. Before we get into the model details, let's look at FilM first. FiLM or Feature Wise linear modulation which is a method used for conditioning of neural nets[20]. Film layers are mainly used for visual reasoning i.e even the complex questions needing high level process are now being answered. Based on the conditioning information, FiLM layers influence the neural network using affine transformation done feature-wise. In the model that we propose, the representation for the text as well as image attributes are first obtained along with the image's representation which is conditioned on input text using FiLM ,co-attention process is being used here to find multimodal incongruity and all the representations are finally concatenated using CLS token from RoBERTa. It is then passed through a connected neural etc and later on through a sigmoid function for the purpose of classification. Text can be considered as a continuous sequence of words for representation.

$$L = \{[CLS], L_1, L_2, ..., L_n, [SEP]\}, \text{ here } L_i \in R^d \tag{3.1}$$

n denotes embedding's maximum length , d is size of embedding features, A from the text are extracted from RoBERTa's first encoder layer and the text is represented.

$$A \in R^{N \times d}$$

d denotes RoBERTa's size and N denotes length of set L. In a similar manner, attributes of images can be represented using I and features are represented by B

$$I = \{[CLS], I_1, I_2, ..., I_n, [SEP]\}, \text{ here } I_i \in R^d \tag{3.2}$$

$$B \in R^{M \times d}$$

M denotes length of set I . Drawing inspiration from [20],the image is conditioned on input text after every feature's affine transformation is applied on it. ResNet-50 which is a pre-trained residual network is used for extracting the features from image while text is processed using gated Recurrent unit[27] whose input is 100-dimensional GloVe[28] embeddings. Final layer from GRU gives nth block FiLM parameters. (γ_i^n , β_i^n). f and g are two arbitrary functions which for input x_i give output γ_i and β_i which are learned by FiLM. 4 FiLMed Residual blocks having a linear layer up top have been used resulting in the final result B_{film} .

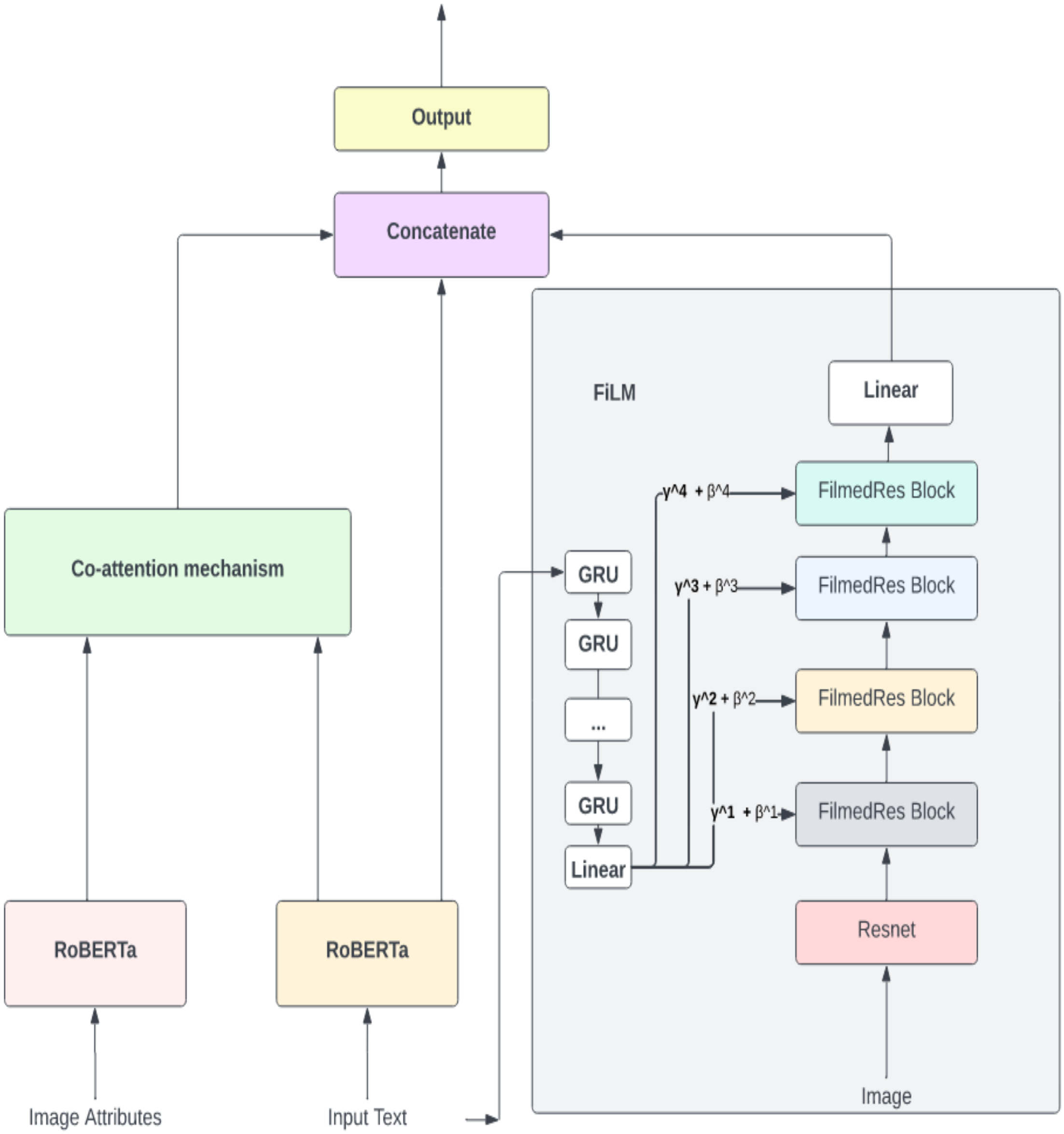


Fig. 3.1. A Diagram of the Proposed Method

For conditioning the visual pipeline, in between every residual block, FiLM layers are inserted with help from these parameters γand β which perform affine transformation corresponding to every feature extracted by ResNet. $F_{i,c}$ refers to feature map c's i^{th} input, motive is to extract the visual feature similar in meaning to the text.

$$FiLm(F_{i,c}) = \gamma_{i,c} * F_{i,c} + \beta_{i,c} \qquad (3.3)$$

Inter-modal incongruity is captured by conditioning of textual representation on image representation. Drawing inspiration from[29], we make use of a mechanism based on co-attention to find incongruity between attributes of the image and the input text. Both output for image as well as text from the RoBERTa model are fed to the co-attention block. The interaction between the attributes of the image and the text is then calculated by affinity matrix, C by making use of bilinear transformation.

$$C = \tanh(AW B^{T}) \qquad (3.4)$$

$$A \in R^{N \times d}, W \in R^{d \times d}, B \in R^{M \times d}$$

W is the weight parameter , A, B denote features of input text and image attribute respectively. With C,the affinity matrix, the attention space of text is now changed to that of the attribute of image. Using a kernel (N x 1) ,2D max-pooling is done over C to calculate α, which is known as attention weight.

$$\alpha = MaxPool(C) \qquad (3.5)$$

At last, the factor measuring how contradictory images' features are with respect to text, called attention matrix of the image attribute B_{att} is computed:

$$B_{att} = \alpha B \qquad (3.6)$$

As a final step we fuse B film obtained from FiLM, [CLS] token from RoBERTa's representation of input text and B_{att} derived above from co-attention mechanism. A fusion vector is obtained on concatenation:

$$V_{fusion} = concat(B_{film}, [CLS], B_{att}) \qquad (3.7)$$

Using a completely connected layer and then a sigmoid function, this fusion vector is passed to obtain the prediction results.

CHAPTER4: RESULTS

To have a comparative analysis with other best performing models proposed so far, we will refer to TABLE IV. Our model has been stacked against various proposed models like image only based ResNet model proposed in [4]. Another text only CNN based model is proposed in [6] which performs decently on the problems like classification of the text. The real competition is provided by the models based on both text and images like in [7] which is BERT based model relying on both inter and intra-modal incongruity for sarcasm detection.Another model using ResNet and BERT is proposed by the same authors from which we draw heavy inspiration. Similarly in [8], a fusion based model is proposed which takes image ,text and image attributes as different modalities and fuses their proper representations for prediction.

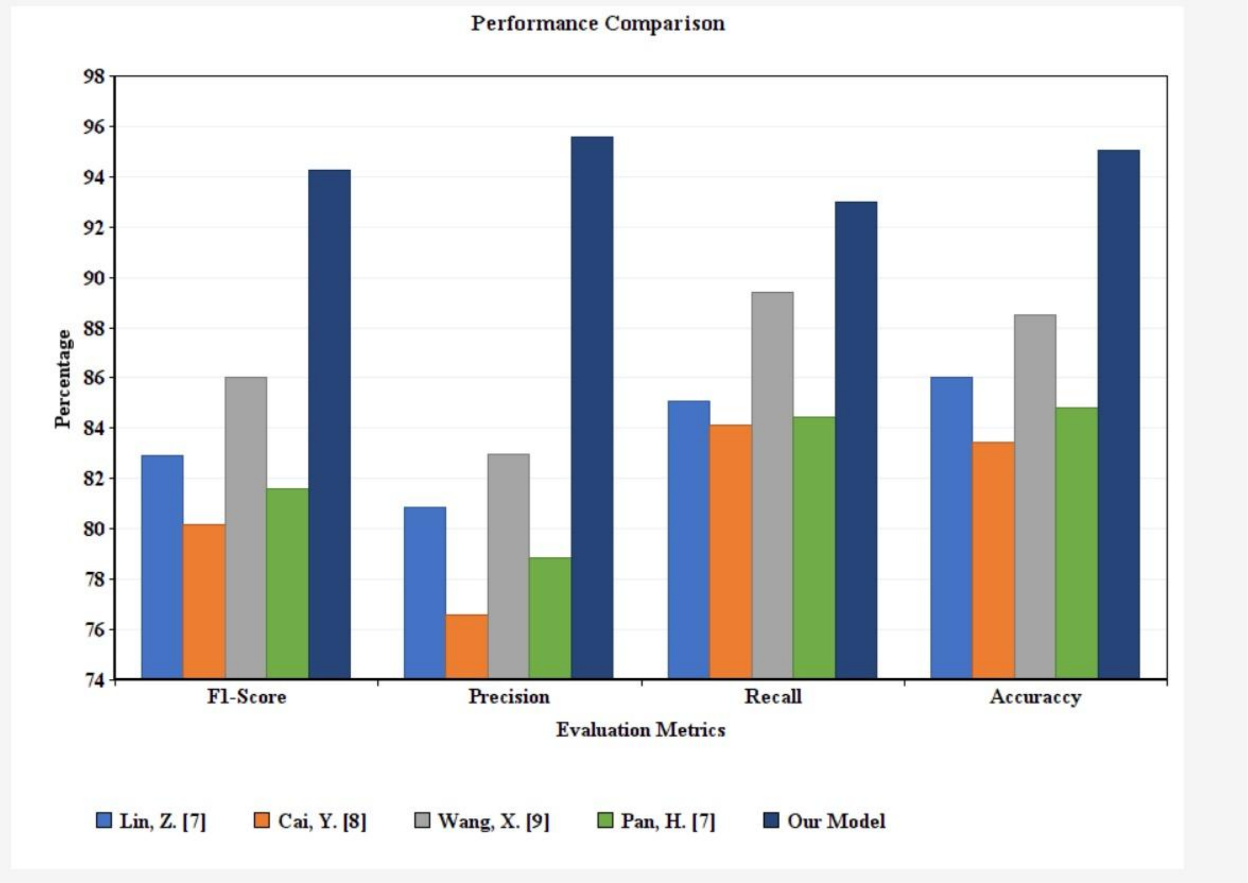


Fig.4.1 Performance Comparison when stacked against other models

The best performing model we found in our research is the one proposed in [9] which uses RoBERTa, ResNet and a bridge layer; an attention layer is also used to find the relationship between the image and text. Our model performed better than this best performing model by 8.21 % in F1-Score as well as 6.51 % in accuracy thus proving that our model surpasses the other models. It can also be concluded from the table that unimodal models suffer a bit due to the absence of other modality, especially only image based models as they lack the required information for the detection of sarcasm. Thus it can also be concluded that the adding another modality leads to performance enhancement when it comes to sarcasm detection.

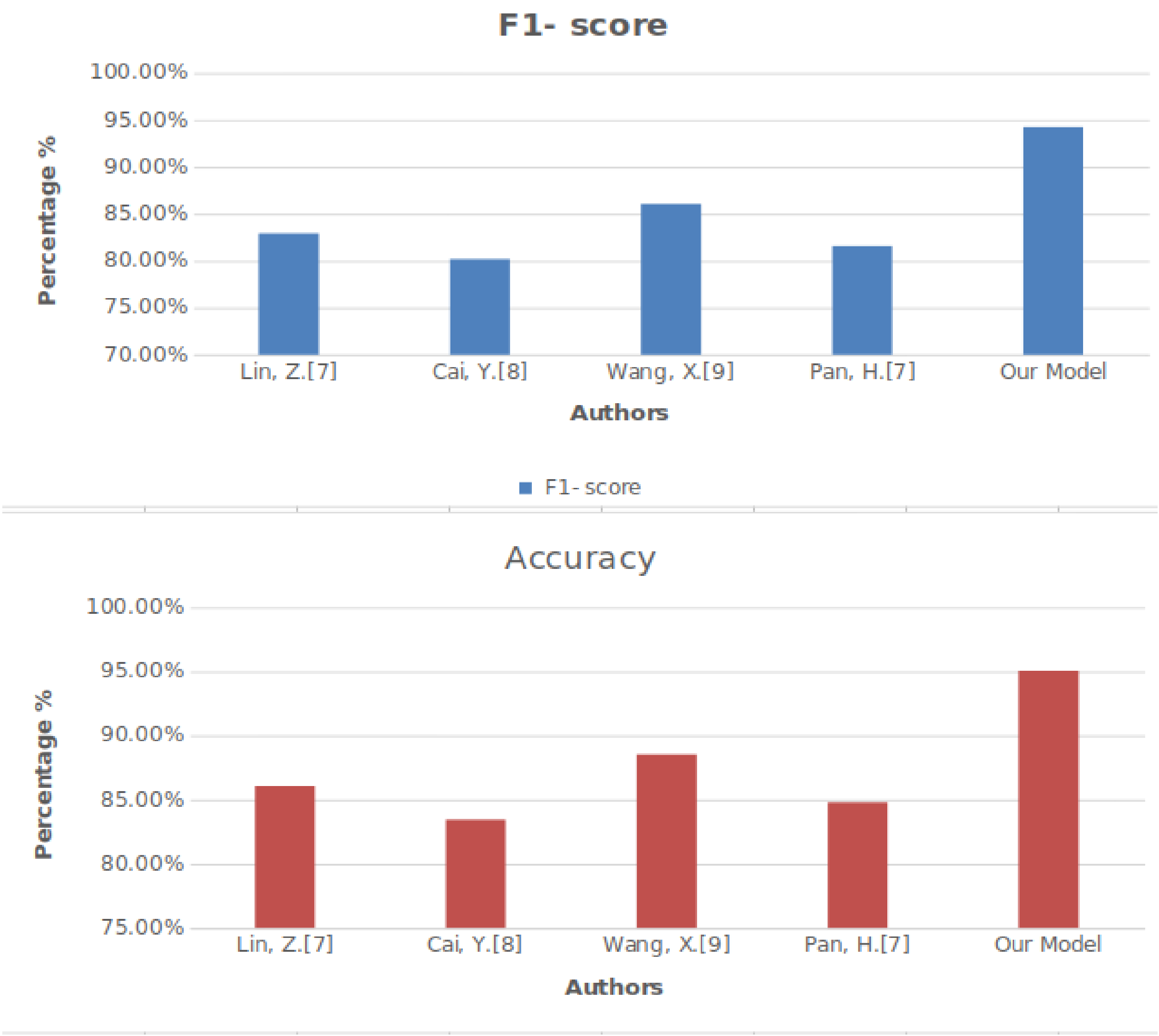


Fig.4.2. F1-Score and Accuracy comparison when stacked against other models

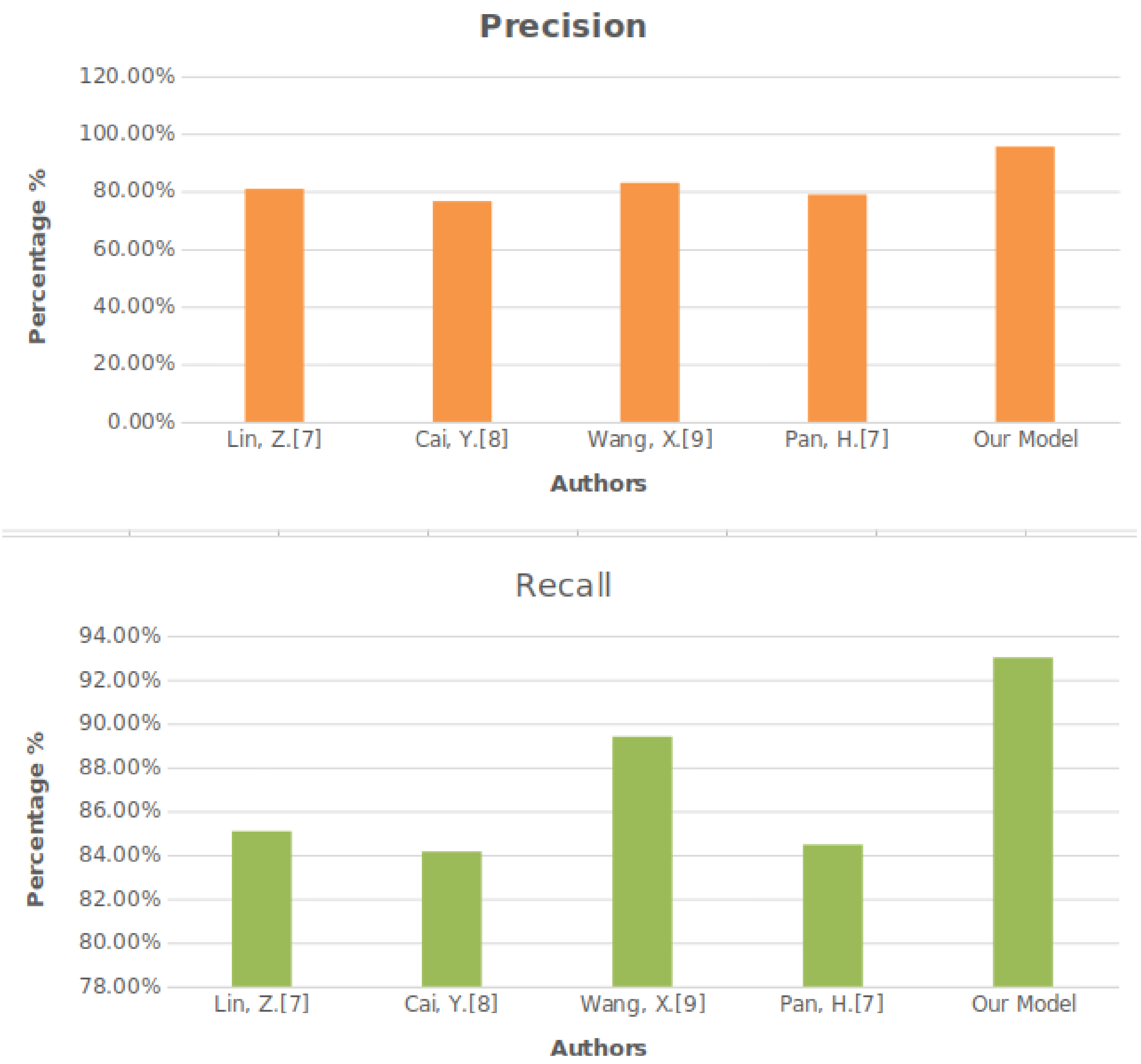
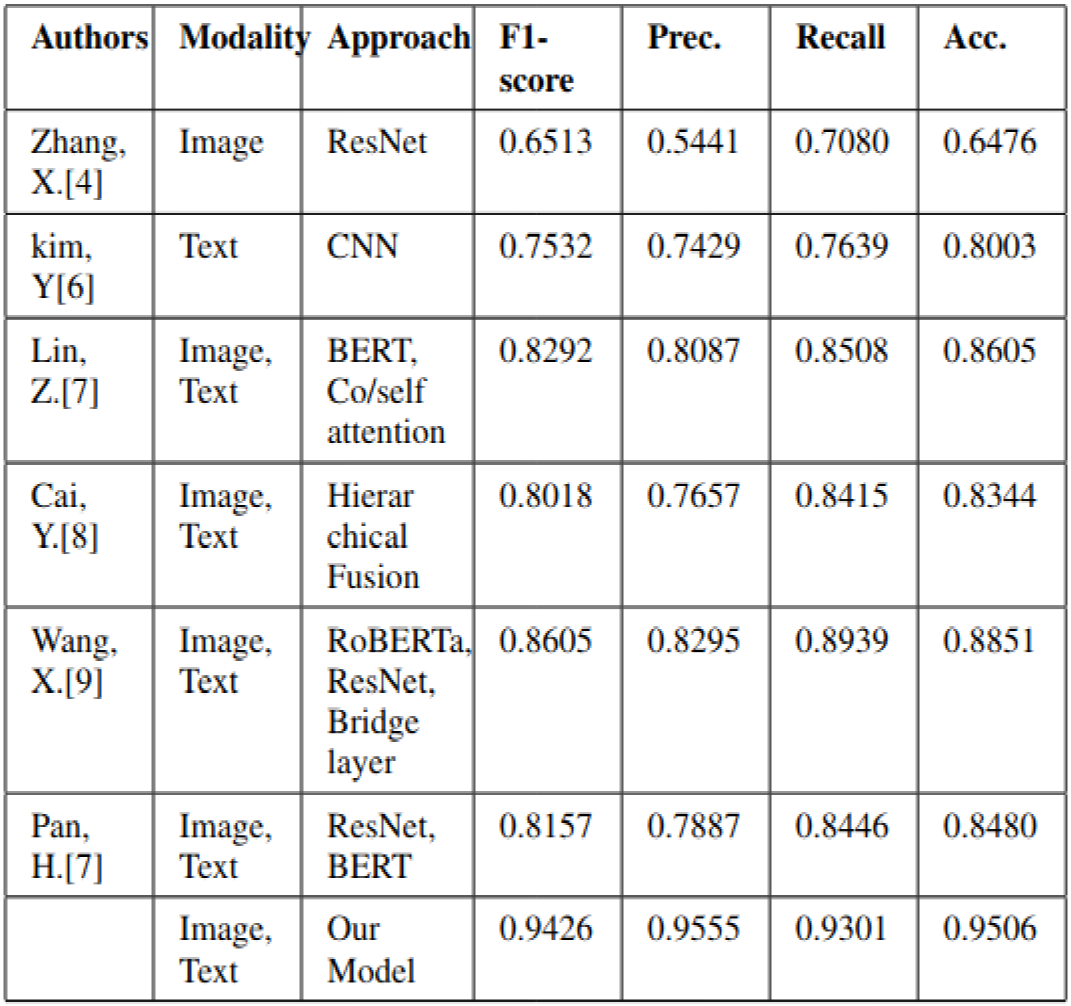


Fig.4.3. Precision and Recall when stacked against other models

Our model is able to perform better as it better captures the incongruity between the images and text which we used as the basis for sarcasm detection. Another reason for the improvement is the capture of incongruity between text and image attributes and also between the image features and text which is achieved with the help of Film Layers. Also the co-attention process allows every word from the input text to be attended by attributes of the image which facilitates better high level representation of features of images conditioned over input text.Thus inter-modal incongruity between the images and the text can be better captured between the attributes of the image and input text. Finally the CLS token from the RoBERTa model is used while concatenating the representations obtained so far. Thus making use of inter-modal incongruity is effective for doing sarcasm detection.

TABLE IV Stacking up against other best performing models



Authors	Modality	Approach	F1-score	Prec.	Recall	Acc.
Zhang, X.[4]	Image	ResNet	0.6513	0.5441	0.7080	0.6476
kim, Y[6]	Text	CNN	0.7532	0.7429	0.7639	0.8003
Lin, Z.[7]	Image, Text	BERT, Co/self attention	0.8292	0.8087	0.8508	0.8605
Cai, Y.[8]	Image, Text	Hierarchical Fusion	0.8018	0.7657	0.8415	0.8344
Wang, X.[9]	Image, Text	RoBERTa, ResNet, Bridge layer	0.8605	0.8295	0.8939	0.8851
Pan, H.[7]	Image, Text	ResNet, BERT	0.8157	0.7887	0.8446	0.8480
	Image, Text	Our Model	0.9426	0.9555	0.9301	0.9506

CHAPTER5: CONCLUSION & FUTURE WORK

In the work done by us, we suggested a new model for sarcasm detection based on finding between image and the text, the inter-modal incongruity. The incongruity is being handled in two ways by taking into consideration both image features and image attributes and their incongruity with the input text. One way is by using the co-attention mechanism which makes sure that attributes of the image attend every word from the input text which helps in getting better representation of image features after being conditioned with the input text.The other one finds incongruity between the features of an image and the text by making use of Film parameters. The model surpasses other proposed models because it has been able to capture the incongruity in a better way owing to the two ways described above.

Also from the analysis done comparatively with the metrics from the other models, it can be easily seen that unimodal models take a hit in the performance due to the absence of other modality with only image based unimodal models performing even worse due to inadequate amount of information captured for sarcasm detection. Thus we can conclude that adding other modality helps in making the models perform better. Also for our future endeavors, we would explore the case where adding extra modality leads to performance degradation. Also this work can be taken a step further to distinguish between the good sarcasm or healthy humor as well as bad sarcasm which can take drastic forms like bullying, harassment etc. Sarcasm is highly common today mainly on social media platforms and it finds its applications in a lot of cases mainly sentiment analysis, cyberbully detection etc. The scope of this research is limited only to the data having 2 modalities namely text and images but in today’s fast growing word, audios and videos are most common, in some case even more than the text and images like on social media platforms like Reels by Instagram, shorts by Youtube and tik-toks from Tik-Tok. In near future we would like to dive more into the multimodal data containing videos for sarcasm detection.

REFERENCES

[1] D. Bamman and N. A. Smith. Contextualized sarcasm detection on twitter. In M. Cha, C. Mascolo, and C. Sandvig, editors, Proc. of the Ninth Int. Conference on Web and Social Media, ICWSM, pages 574–577. AAAI Press, 2015.

[2] A. Joshi, V. Sharma, and P. Bhattacharyya. Harnessing context incongruity for sarcasm detection. In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), volume 2, pages 757–762. The Association for Computer Linguistics, 2015.

[3] A. Khattri, A. Joshi, P. Bhattacharyya, and M. Carman. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In Proc. of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 25–30, Lisboa, Portugal, September 2015.Association for Computational Linguistics.

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In:Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[5] Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In Proceedings of the 2016 ACM on Multimedia Conference. ACM, 1136–1145.

[6] Kim, Y.: Convolutional neural networks for sentence classification. CoRR abs/1408.5882 (2014), http://arxiv.org/abs/1408.5882

[7] Pan, H., Lin, Z., Fu, P., Qi, Y., Wang, W.: Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. pp. 1383–1392 (2020)

[8] Yitao Cai, Huiyu Cai and Xiaojun Wan ,"Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model" , Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, August 2019

[9] Wang, X., Sun, X., Yang, T., Wang, H.: Building a bridge: A method for image- text sarcasm detection without pre-training on image-text data. In: Proceedings of the First International Workshop on Natural Language Processing Beyond Text. pp. 19–29 (2020)

[10] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018a. Hybrid attention based multimodal network for spoken language classification. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2379–2390. Association for Computational Linguistics

[11] Vasavi Gajarla and Aditi Gupta. 2015. Emotion Detection and Sentiment Analysis of Images. Georgia Institute of Technology (2015).

[12] Dipto Das ,Anthony J. Clark . "Sarcasm Detection on Flickr Using a CNN" , ICCBD '18: Proceedings of the 2018 International Conference on Computing and Big Data September 2018

[13] Jens Lemmens , Ben Burtenshaw , Ehsan Lotfi ,Ilia Markov ,Walter Daelemans ."Sarcasm Detection Using an Ensemble Approach" , Second Workshop on Figurative Language Processing ,July 2020

[14] A. Mishra, K. Dey, P. Bhattacharyya (2017). Learning Cognitive Features from Gaze Data for Sentiment and Sarcasm Classification using Convolutional Neural Network. ACL.

[15] Joshi, A., Sharma, V., Bhattacharyya, P.: Harnessing context incongruity for sar- casm detection. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 757–762 (2015)

[16] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., Huang, R.: Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 704–714 (2013)

[17] Santiago Castro, Devamanyu Hazarika, Ver ́onica P ́erez-Rosas†,Roger Zimmermann, Rada Mihalcea†, Soujanya Poria . “Towards Multimodal Sarcasm Detection” , Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, August 2019

[18] Yitao Cai, Huiyu Cai and Xiaojun Wan ,”Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model” , Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, August 2019

[19] S. Sangwan, M. S. Akhtar, P. Behera and A. Ekbal, ”I didn’t mean what I wrote! Exploring Multimodality for Sarcasm Detection,” 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206905.

[20] Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual rea- soning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

[21] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 2539–2544.

[22] Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. Thu ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 51–56

[23] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

[24] Debanjan Ghosh, Alexander R. Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. Computational Linguistics, 44(4):755–792.

[25] Jayashree Subramanian , Varun Sridharan , Kai ShuHuan Liu .”Exploiting Emojis for Sarcasm Detection”, International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation , June 2019

[26] Xu, N., Zeng, Z., Mao, W.: Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3777–3786 (2020)

[27] Chung, J., G ul cehre, C ., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555 (2014), http://arxiv.org/abs/1412.3555

[28] Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word repre- sentation. In: Proceedings of the 2014 Conference on Empirical Methods in Nat- ural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, https://www.aclweb.org/anthology/D14-1162

[29] Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 289–297. NIPS’16, Curran Associates Inc., Red Hook, NY, USA (2016)

[30] Z. Wang, Z. Wu, R. Wang, and Y. Ren. Twitter sarcasm detection exploiting a context-based model.In J. Wang, W. Cellary, D. Wang, H.Wang, S. Chen, T. Li, and Y. Zhang, editors, Web Information Systems Engineering - WISE 2015 - 16th Int. Conference, Miami, FL, USA, November 1-3, 2015, Proc., Part I, volume 9418 of Lecture Notes in Computer Science, pages 77–91. Springer, 2015.

[31] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P.Xing. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. CoRR, abs/1609.05244.

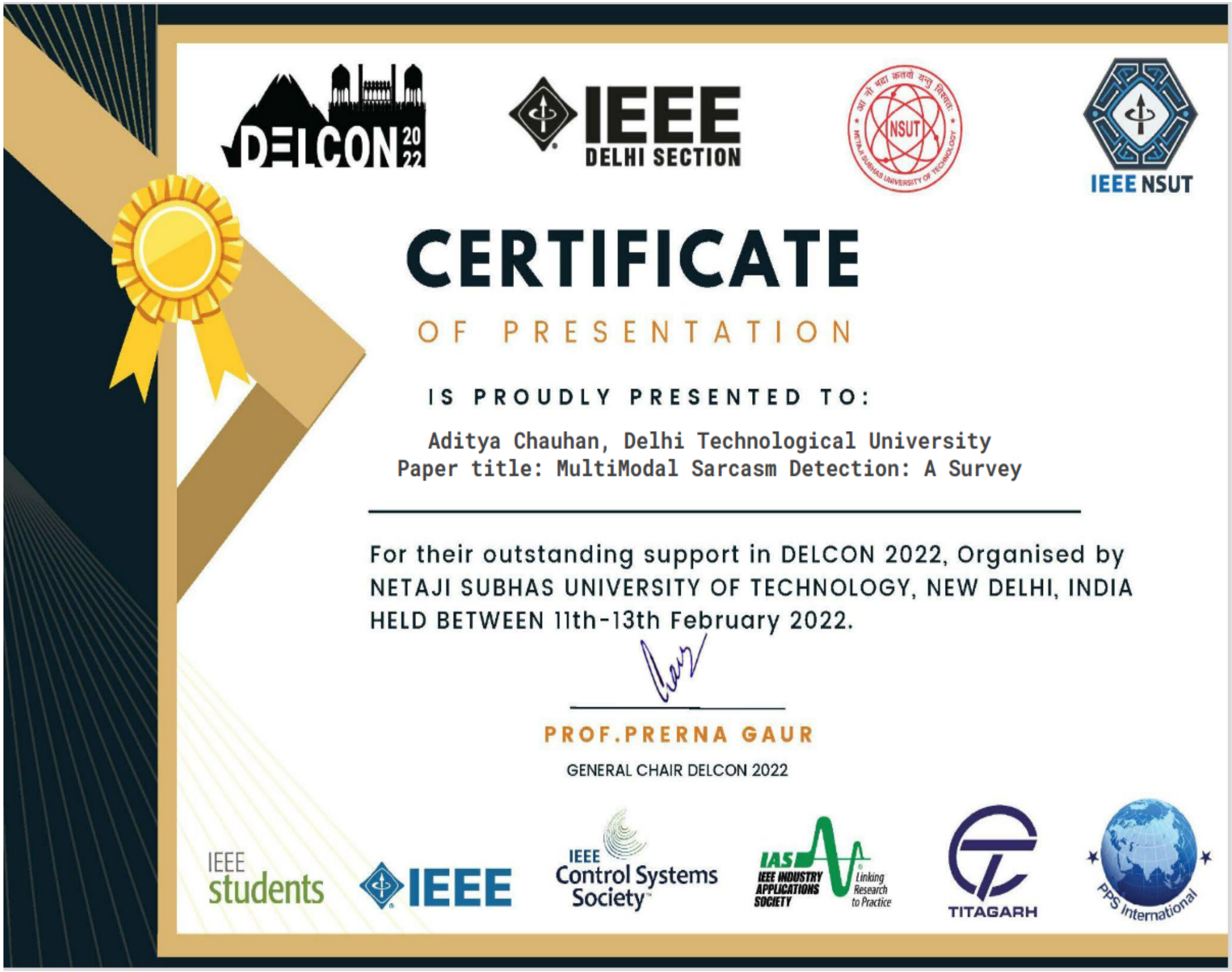
[32] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. CoRR, abs/1707.07250.

[33] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018b. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2225–2235. Association for Computational Linguistics

[34] Tay, Y., Luu, A.T., Hui, S.C., Su, J.: Reasoning with sarcasm by reading in-between. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp.1010–1020. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1093, https://www.aclweb.org/anthology/P18-1093

LIST OF PUBLICATIONS

[1] A. Bhat and A. Chauhan, "MultiModal Sarcasm Detection: A Survey," 2022 IEEE Delhi Section Conference (DELCON), 2022, pp. 1-7, doi: 10.1109/DELCON54057.2022.9753058.
Indexed by Scopus and Google Scholar
URL: https://ieeexplore.ieee.org/document/9753058


DELCON 2022
IEEE DELHI SECTION
NSUT
IEEE NSUT
IEEE students
IEEE
IEEE Control Systems Society
IAS IEEE INDUSTRY APPLICATIONS SOCIETY Linking Research to Practice
TITAGARH
PPS International

CERTIFICATE

OF PRESENTATION

IS PROUDLY PRESENTED TO:

Aditya Chauhan, Delhi Technological University
Paper title: MultiModal Sarcasm Detection: A Survey

For their outstanding support in DELCON 2022, Organised by NETAJI SUBHAS UNIVERSITY OF TECHNOLOGY, NEW DELHI, INDIA HELD BETWEEN 11th-13th February 2022.

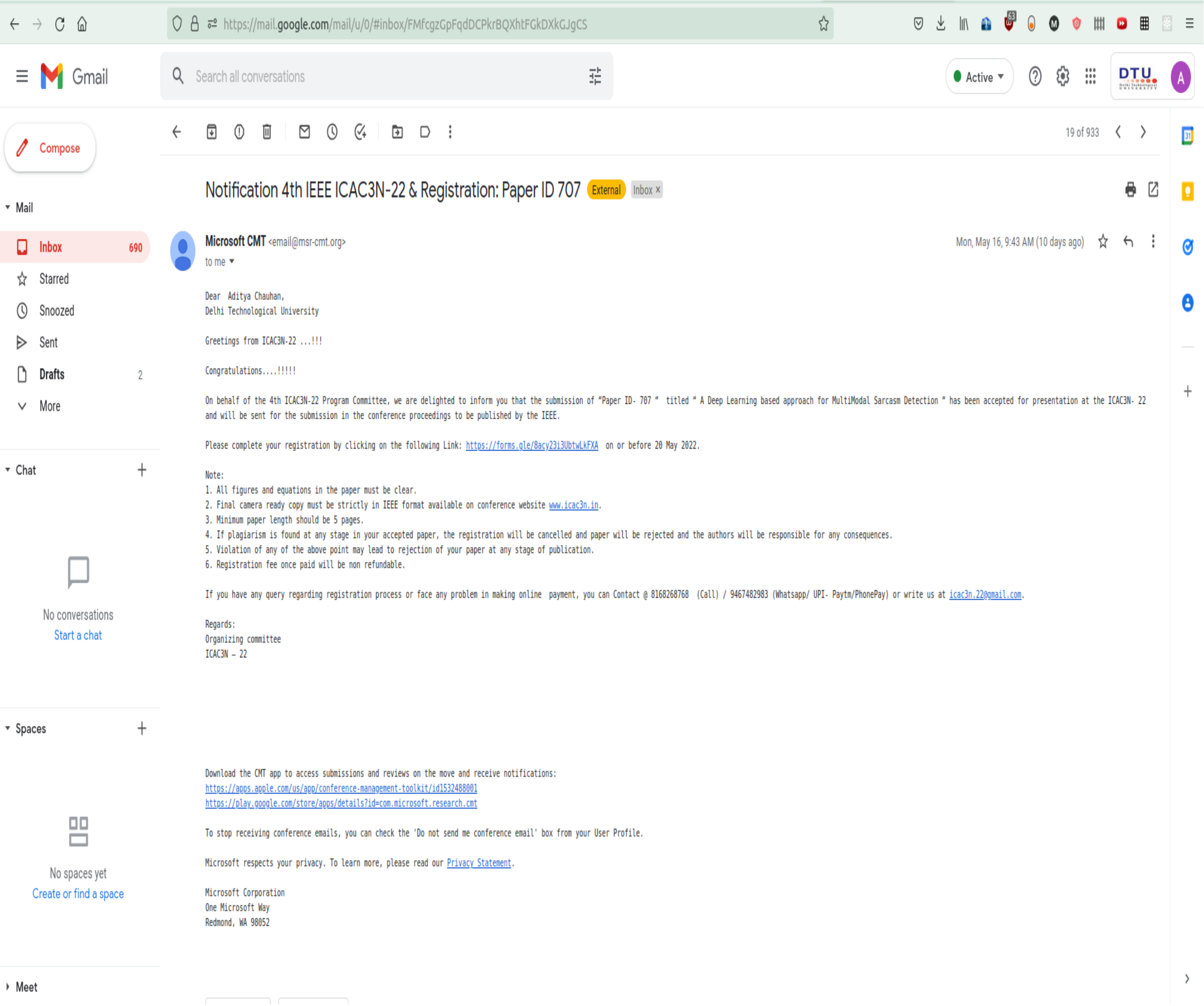
PROF.PRERNA GAUR
GENERAL CHAIR DELCON 2022

[2] A. Bhat and A. Chauhan, "A Deep Learning Based Approach for MultiModal Sarcasm Detection”.

Accepted at the 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N–22).

Indexed by Scopus and Google Scholar

Proof of Acceptance:



Notification 4th IEEE ICAC3N-22 & Registration: Paper ID 707

Microsoft CMT <email@msr-cmt.org>

Mon, May 16, 9:43 AM (10 days ago)

Dear Aditya Chauhan,
Delhi Technological University

Greetings from ICAC3N-22 ...!!!

Congratulations....!!!!!

On behalf of the 4th ICAC3N-22 Program Committee, we are delighted to inform you that the submission of “Paper ID- 707 “ titled “ A Deep Learning based approach for MultiModal Sarcasm Detection “ has been accepted for presentation at the ICAC3N- 22 and will be sent for the submission in the conference proceedings to be published by the IEEE.

Please complete your registration by clicking on the following Link: https://forms.gle/8acy23i3UbtwLkFXA on or before 20 May 2022.

Note:
1. All figures and equations in the paper must be clear.
2. Final camera ready copy must be strictly in IEEE format available on conference website www.icac3n.in.
3. Minimum paper length should be 5 pages.
4. If plagiarism is found at any stage in your accepted paper, the registration will be cancelled and paper will be rejected and the authors will be responsible for any consequences.
5. Violation of any of the above point may lead to rejection of your paper at any stage of publication.
6. Registration fee once paid will be non refundable.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8168268768 (Call) / 9467482983 (Whatsapp/ UPI- Paytm/PhonePay) or write us at icac3n.22@gmail.com.

Regards:
Organizing committee
ICAC3N – 22

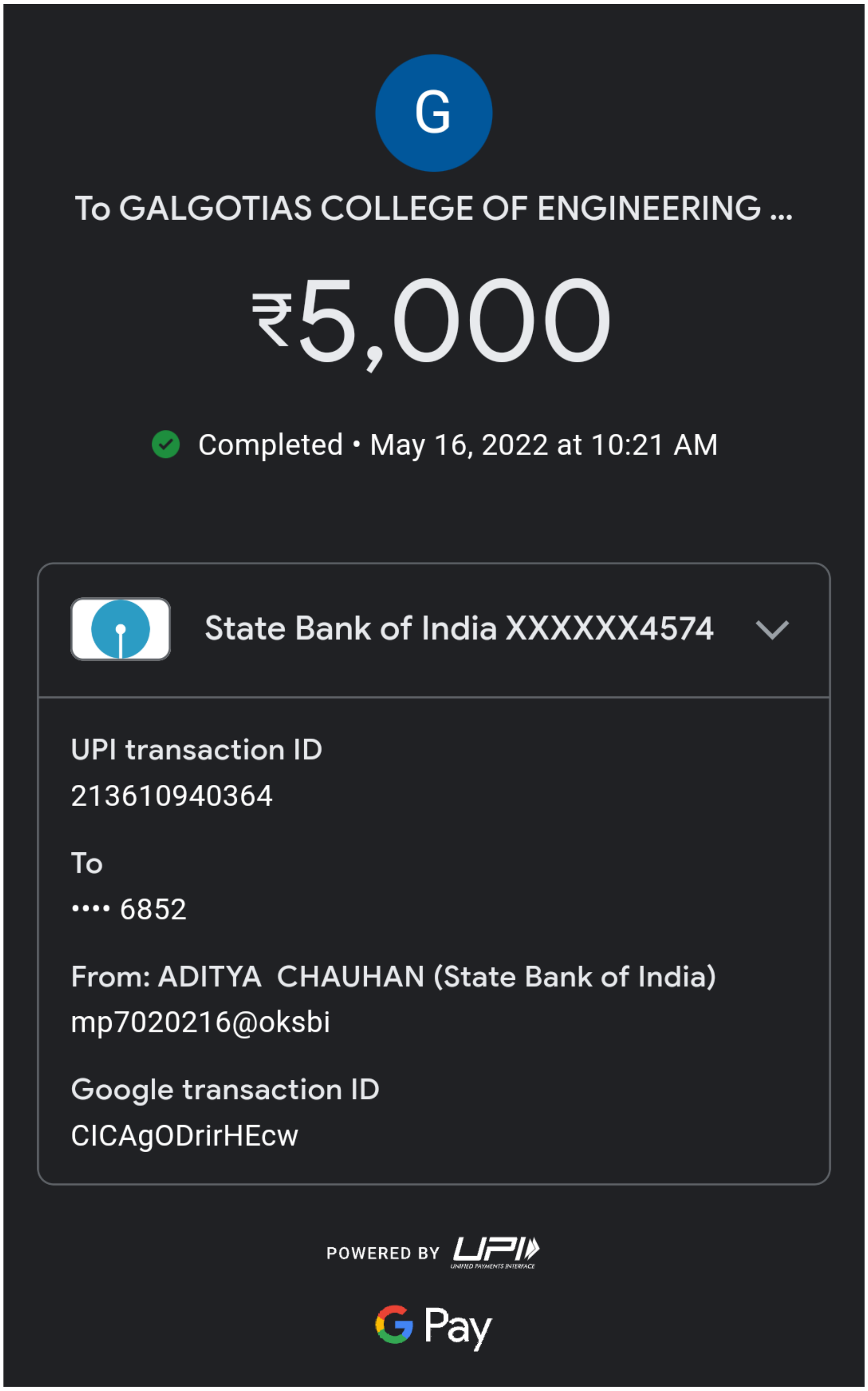
Download the CMT app to access submissions and reviews on the move and receive notifications:
https://apps.apple.com/us/app/conference-management-toolkit/id1532488001
https://play.google.com/store/apps/details?id=com.microsoft.research.cmt

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our Privacy Statement.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Payment Slip :



G

To GALGOTIAS COLLEGE OF ENGINEERING ...

₹5,000

✓ Completed • May 16, 2022 at 10:21 AM

State Bank of India XXXXXX4574

UPI transaction ID
213610940364

To
•••• 6852

From: ADITYA CHAUHAN (State Bank of India)
mp7020216@oksbi

Google transaction ID
CICAgODrirHEcw

POWERED BY UPI

G Pay