Memotion 2.0 - Sentiment Analysis and Emotion classification of Memes

MSc Research Project
Data Analytics

Nilam Bhapkar
Student ID: X20145331

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nilam Bhapkar
Student ID:	X20145331
Programme:	Data Analytics
Year:	2021-2022
Module:	MSc Research Project
Supervisor:	Mohammed Hasanuzzaman
Submission Due Date:	16/12/2021
Project Title:	Memotion 2.0 - Sentiment Analysis andEmotion classification of Memes
Word Count:	6778
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Nilam Bhapkar
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Memotion 2.0 - Sentiment Analysis andEmotion classification of Memes

Nilam Bhapkar

x20145331

Abstract

Modern scientific advances in the Internet and media adoption have contributed to the emergence of more adequate methods for communicating. These platforms, which comprise visual, textual, and voice mediums, have given rise to a distinct social phenomenon known as Internet memes. Internet memes are photographs with humorous, eye-catching, or satirical text captions attached. Nowadays, memes are a very widespread way of expression on social networks. Their multi-modal feature, as a result of a combination of text and pictures, makes them a difficult research item for machine analysis. By categorizing memes based on their emotional content, we can better understand what they are about and avoid the proliferation of sarcasm or negative attitudes. It is the primary goal of this scientific report to shed light on the segmentation of memes into three distinct categories by utilizing the Memotion analysis dataset available through Google Colab. The F1 score, Weighted Avg score, Precision, Recall, and computational duration for execution are all used to evaluate the approach's overall performance and effectiveness. The Deep neural network achieves the highest F1 score possible, which is 64%.

1 Introduction

Memotion analysis is a challenge that is performed in order to better perceive the emotions expressed by memes. A meme is an idea, behavior, or skill that can be transferred from one person to another through imitation Guo et al. (2020). Examples of memes include stories, styles, innovations, formulas, music, and methods of plowing a field, throwing a baseball, or creating a statue. Memes can be transmitted through social media Blackmore and Blackmore (2000). The widespread use of the Internet and immediate messaging applications has resulted in the proliferation of Internet memes on social media platforms such as Facebook, Instagram, and Twitter, which have proven to be a successful form of communication Guo et al. (2020). However, the spread of hate speech in social networking sites has been aided by the most recent Internet memes, making this research topic even more important to address Shifman (2014). Due to the fact that inflammatory memes require visual and language comprehension, identifying insulting memes is more difficult than sensing offensive text. Therefore, it is important to develop a hybrid model for the computerized processing of Internet memes in order to achieve success Guo et al. (2020).

Motivation. Internet memes are available in a variety of genres and layouts. Several of them are mainly entertaining, whereas others, hidden behind a humourous look, are intended to express subtle nuances such as facetiousness, unbelief in a given idea, or an inspirational goal. The majority of them can be found in the online world and provide insight into the viewpoints of some societies on specific issues Williams et al. (2016). Furthermore, they can be utilized for the purpose of acquiring crucial data that will be used to make further positive changes to the online content extraction system in the future.

Challenges. The mining challenge becomes exceedingly challenging because the clearness of both the picture and the message appears to vary significantly based on a user or the territory from which the post was made, making the task extremely hard. Once an evaluation operation is carried out, these circumstances can consequence in adverse outcomes, and in some situations, they can even result in the transmission of a different concept than designed. It is possible that when the level of confusion exceeds a certain point, it will be impossible to distinguish between the two primary ways memes deliver information: message and picture. Aside from that, because the message is engrained within the photo, a poor image can cause distortion into the information, which can bring down the entire meme's integrity. The opposite is also true: a text that is totally vague will result in a disparity between it and the appearance of the post, which will nullify the core message.

Proposed Approach. By presenting a neural network based on multi-task learning, we hope to be able to solve the subtask that was mentioned Memotion analysis challenge. The structure will include components devoted solely to image processing as well as modules that are specifically geared toward text analysis. The architectural style produces a solution for the task that is considered necessary.

1.1 Research Question

RQ : *Is it possible to prevent the spread of unpleasantness in society by enhancing the identification and classification of meme attitudes using machine learning algorithms (Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Multilayer perception (MLP)) using the Google Colab platform?*

1.2 Research Objectives

The research project goals outlined below are presented in relation to the research question posed in this research work, and they include a detailed clear roadmap for consummating the project with success on time.

Objective 1: How accurate is the created meme classification model (RQ) in distinguishing positive, negative, and neutral memes from one other?

Objective 2: How well the newly created model worked on an imbalanced dataset and successfully categorized memes according to category?

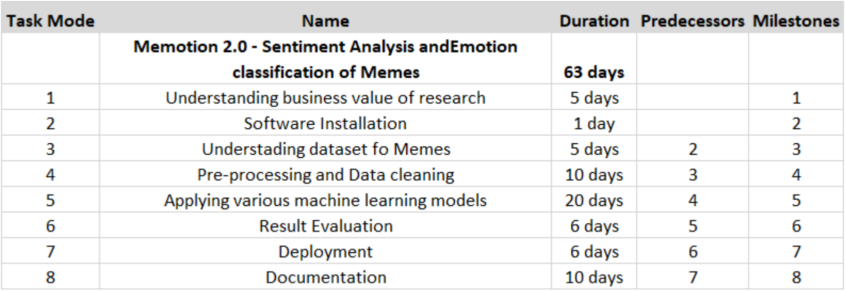
Objective 3: Identify the appropriate machine learning algorithm for the ongoing study based on their efficiency.

1.3 Contributions

The research study participate to the existing area of study in two ways: by examining the differences in the literature review and by emphasizing the significance of the research topic areas that have been posed for consideration. The involvement of the work are described below in accordance with the criteria.

Machine learning and deep learning techniques are used in conjunction with Google Colab to develop, assess, and forecast sentiment of memes in the memotion 2.0 dataset. In order to prevent the spread of a negative attitude in the community, this research study implemented the RNN, MLP and CNN models using Google Colab.

The study has contributed the best outcome-providing technique for the specified dataset by employing a state-of-the-art method and applying it to the selected dataset. The author attempted to acquire the best classification performance possible under the circumstances that were chosen. Timeline of the project as shown in the Figure 1.



Task Mode	Name	Duration	Predecessors	Milestones
	Memotion 2.0 - Sentiment Analysis andEmotion classification of Memes	**63 days**		
1	Understanding business value of research	5 days		1
2	Software Installation	1 day		2
3	Understading dataset fo Memes	5 days	2	3
4	Pre-processing and Data cleaning	10 days	3	4
5	Applying various machine learning models	20 days	4	5
6	Result Evaluation	6 days	5	6
7	Deployment	6 days	6	7
8	Documentation	10 days	7	8

Figure 1: Project Plan

The remaining sections of the document are outlined as follows. Literature review and the previous work done on the challenge are topics covered in Chapter 2, as well as the related projects of emotion prediction of memes employing Deep Learning, Sentiment Classification using text information, and Content Pre-processing, which is typically used throughout Data Analysis and Sentiment Classification. The CRISP-DM technique, which was used in the implemented project, is described in detail in Chapter 3. The structure and architectural style of the research work are discussed in detail in Chapter 4. Detailed instructions on how to complete the project's step-by-step deployment are provided in Chapter 5, which covers data gathering, data pre-processing, sentiment rating estimation, correlation analysis, prediction models development, The assessment of the approach, Error analysis, the Model comparison as well as visualizations. Chapter 7 contains the outcome of the work that has been completed as well as the scope of the project that will be undertaken in the coming years.

2 Related Work

2.1 Introduction

The experiment performed in classifying sentiment of content in text and pictures is enclosed in the related section of this document. It also discusses the work that has been conducted in the areas of meme analysis and multimodality in general. This section is broken into several sub-sections-i) Sentiment Identification using Text Data ii) Experiment with Text and Image-Based meme iii) Image-based Sentiment Classification iv) Detection of hate speech in Internet memes v) Combining Textual and Visual Input Approaches vi) Humor Recognition vii) Others

2.2 Sentiment Identification using Text Data

Due to the multimodal nature of the internet meme, it presents an immensely difficult research topic for automatic detection. Nonetheless, internet memes are a major part of users' online appearance and thus provide an income of potential data for determining user emotion. Automatic emotion recognition in memes may aid in the advancement of studies into the viral transmission of internet phenomena, particularly with regard to the concept of emotional expression Guadagno et al. (2013). Regardless of the recent paucity of research in the automatic sentimental analysis of memes, except perhaps french2017image, who investigated the retrieval of its intrinsic emotion by connecting the memes to the relevant reader comments, the emphasis of evolving sentiment identification study is on multimodal categorization, which has the potential to be very effective Verma et al. (2020). While significant progress has been made in the identification of sentiments in text data joshi2017survey, continued attempts to implement image-based characteristics from the area of Computer Vision, such as optical character recognition (OCR) and facial detection, have revealed the need for enhancements in the textual data of memes, due to the short and complicated nature of memes Verma et al. (2020). It has been proposed to apply sentimental analysis to picture data with the aim of automatic tag forecasts for pictures uploaded to social networking sites Gajarla and Gupta (2015), that also, in turn, will aid in the optimization of image search methodologies by producing a wide collection of labeled image files.

2.3 Experiment with Text and Image-Based meme

When we examine the complication that multimodal channels of communication add to challenges such as the automatic identification of abusive speech online Williams et al. (2016);Lee et al. (2018), the necessity of combining the knowledge gained from the research of the two modalities of memes becomes apparent Zampieri et al. (2019). The SemEval Memotion work intends to make a contribution to the field of automatic emotion identification in memes, as well as to the collection of fine-grained data such as irony, comedy, and rudeness from memes. The distinction between these three sorts of humor is difficult to make sense even the type of comedy identified in non-offensive memes seems to straddle the line between what is appropriate and what is not. These individuals' ability to work together is manifested in this, sometimes politically wrong, way Procházka (2016), 2016). In view of the arbitrary nature of the subcategories to be recognized, the task has been expanded to include the determination of the degree to which the comedy type discovered is present (not, slightly, moderately, very) in the meme under investigation. As the meme creation studies carried out by Oliveira et al. (2016) demonstrated, the observed hilarious quality of memes is frequently depending on the macros that are utilized and the meaning that they already bring with them. These macros are multimodal, which means that their meaning must be automatically determined by using both textual and image-based elements in conjunction. Traditionally held hypotheses of sense of humor, such as Raskin (2012) three sources of humor: incongruity, arousal-safety, and depreciation, serve as the foundation for the attributes of facetiousness which are used in automatic investigative techniques, such as inconsistency, information sharing, believability, and derision Joshi, Bhattacharyya and Carman (2017). Sarcasm identification on information that is not exclusively textual has only been attempted on typographical memes (memes comprising solely of the word, but often structured in a number of potential typefaces) using a Multi-Layer perceptron, with an average accuracy of 88 percent, according to the authors Kumar and Garg (2019). While significant advances have been made in the identification of text-based sarcasm, including the use of semi-supervised sequence retrieval, lexico-semantic knowledge representation, and data-driven process to identify implied emotion Joshi, Prajapati, Shaikh and Vala (2017);Van Hee et al. (2018), the sensing of sarcastic in multimodal contexts will require consideration of additional features for sarcasm generated by the picture, as well as the interplay among both text and pictures

2.4 Image-based Sentiment Classification

Unlike text-based sentiment classification, which was invented by Davidov et al. (2010) and Bollen et al. (2011), picture-based emotion recognition has received far less priority. This is due to the fact that image-based sentiment classification is more difficult to perform than word-based sentiment classification. Because Internet memes are a comparatively recent study area, current work is widely connected to investigations on forecasting the emotion from visual imagery in web-based commentaries Truong and Lauw (2017), which are also relevant to current research. A system for collecting mid-level lexical characteristics and photos was developed by Borth et al. (2013) to do sentiment classification on a massive visual sentiment ontology. This method is termed SentiBank. Chen et al. (2015) established the Visual Emotional Latent Dirichlet Allocation (VELDA) prototype to collect the text-image similarity from different modalities: the text, the visual view of the image, and the sentimental point of view of the photo. This model was used to examine the picture uploading behavior of people on social media. And, perhaps most surprisingly, they discovered that 66 percent of users include a picture in their social network posts.

2.5 Detection of hate speech in Internet memes

Lately, Hu and Flaxman (2018) created a multimodal sentiment classification strategy that uses textual and image to predict the sentiment keyword tags that users have applied to their Tumblr posts, which was published in Nature Communications. Author and colleagues (2015) created a robust Image Sentiment Classification approach by integrating a CNN on Flickr with domain transfer from Twitter for binary sentiment classification. They used a CNN on Flickr with information is transferred from Twitter for binary sentiment analysis. By utilizing visual data, Sabat et al. (2019) and colleague attempted to solve the problem of automatically detecting hate speech in Internet memes. A Multimodal Social Media system created by Blandfort et al. (2019) was used to investigate how public tweets with photographs made by youths who reference gang crime on Twitter may be used to instinctively detect psycho-social characteristics and concerns.

2.6 Combining Textual and Visual Input Approaches

A small number of investigators have looked into automating the process of creating Internet memes, while a smaller number have attempted to retrieve the emotion contained inside them. Williams et al. (2016) suggested a non-paranormal strategy for generating famous meme descriptions, incorporating visual and verbal data. This technique was successful in generating famous meme characterizations. Peirson V and Tolunay (2018) proposed an encoder-decoder meme producing system, which consisted of Google's pretrained Inception-v3 network to build an image encoding, accompanied by an LSTM model with a focus to create the meme descriptions. Wang and Hua (2014) developed a technique for combining textual and visual input, which was then scaled up utilizing efficient dropout regularization. This technique is discussed in detail in this publication.

2.7 Humor Recognition

By merging a Convolutional Neural Network (CNN) Fukushima and Miyake (1982) and a Long Short-term Memory Network (LSTMN),Yoshida et al. (2018) suggested a multimodal technique for humor recognition Hochreiter and Schmidhuber (1997). It should be noted that the researchers also used the ResNet-152 He et al. (2016) model and created a bespoke loss function that takes into consideration a humorous score. They were using the comments from some other comedy website, where users rate the posts using stars, to calculate the humor of a particular post. The loss function is centered around a specific threshold, which the authors determined to be 100. In addition, employing a feature-based approach, the problem of humor recognition has been addressed Chandrasekaran et al. (2016).A number of various levels of analysis were used to produce the characteristics by the researchers, including cardinality and location, as well as object and instance-level features. For the purpose of predicting comedy score, a Support Vector Regression is applied. As a result of their work, the authors developed a novel approach for enhancing humor ratings by adjusting how amusing or unamusing a scenario is. For starters, they determined the elements that add to the humor of a specific scene before identifying substitutes that can change the level of amusement provided by that specific scene. They employed a multi-layer perceptron (MLP) for the first half of the experiment, which produces a boolean class for each object in the situation. Later, in order to change the tone of the humor, they utilized another MLP that had been practiced to recognize alternative substitutes for the original.

2.8 Others

Because of the diverse variety of conversational roles that memes perform, they have piqued the curiosity of researchers across a wide range of fields. The linguistic and cultural study of the unifying influence of online memes on societies Procházka (2016), as well as their capacity to raise consciousness for controversial topics Phillips and Milner

(2017), has examined the impact of internet memes on societies. It is particularly interesting in the field of Communication Sciences to measure the influence of specific types of humor on the likelihood of a meme becoming viral Taecharungroj and Nueangjamnong (2014);Mina (2019).There has also been researching into memes in the fields of computational inventiveness, with efforts to optimize the meme creation process Peirson V and Tolunay (2018);Oliveira et al. (2016), and information extraction, with involvement in the potential of individualized search queries for memes Peirson V and Tolunay (2018);Oliveira et al. (2016);Milo et al. (2019).

2.9 Conclusion

Many methods were used to examine the meme content. Initially, some researchers used various strategies to classify emotions in text data. Later, several experiments on text and image-based memes were conducted in order to discern between the comic and objectionable memes. Some of the linked work is quite relevant to the current investigation.

3 Methodology

The KDD or CRISP-DM experiment procedures are typically used for data analysis study approaches. In this particular instance, a revised version of the CRISP-DM technique has been used instead. It has been changed and will be executed in the following phases: i) Gaining a knowledge of the business. It was decided to use the semEval Memotion dataset for this study. ii) Data selection The data is basically composed of four different categories of emotions. iii) data is analyzed and preprocessed using methods such as denoising and normalizing. The preprocessed data is then used to obtain the features and labeling that are needed. After that, the data is changed in order to fit the categorization models. 4. Implementation - sentiment categorization is carried out in Google Colab utilizing RNN, CNN, MLP amongst other techniques. v) The weighted avg, accuracy, F1 score, Precision, and recall of the system are used to assess the model's overall results. vi) The completed strategy is then applied to a variety of data sets to determine its performance. The study technique that was employed is depicted in Figure 2.

3.1 Business Understanding

In addition to the amusing visual, the Online culture, or the attitude that is spread, memes are distinguished by the variety and individuality of their language: they are frequently highly organized, with an uncommon writing style, and they are frequently shared on social media. The Internet Memes frequently incorporate a layered written description with poor spelling and grammatical inconsistencies, as well as other elements. Furthermore, Online memes can be found in practically all types of media, with new and innovative formats being introduced on a regular basis. Originally, they serve as a vehicle for the dissemination of humor through the use of cultural topics. The use of social networks can, though, be misused in order to support political beliefs, commercial advertising, and social media advertising campaigns.

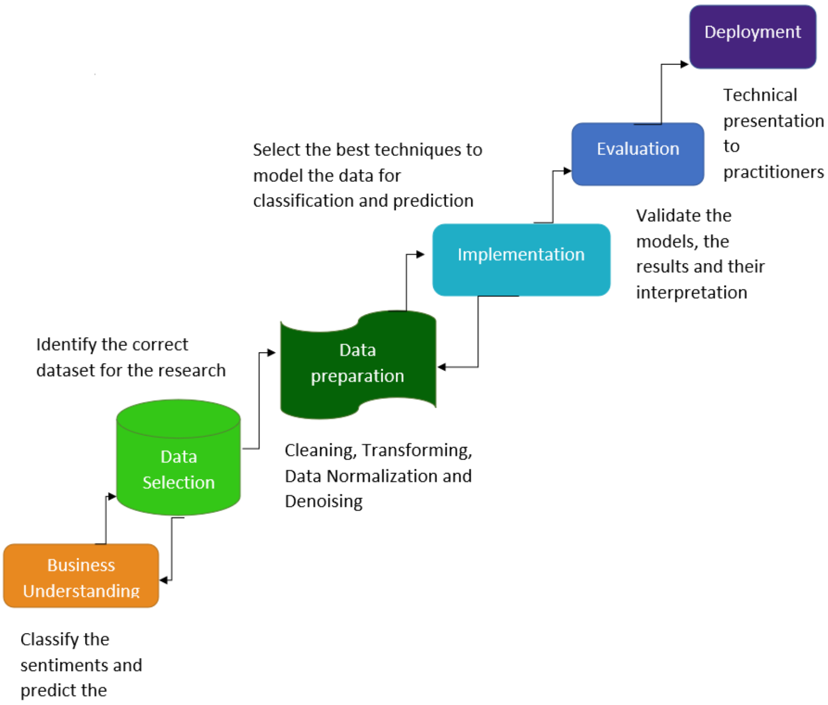

Deployment
Technical presentation to practitioners
Select the best techniques to model the data for classification and prediction
Evaluation
Validate the models, the results and their interpretation
Implementation
Identify the correct dataset for the research
Data preparation
Cleaning, Transforming, Data Normalization and Denoising
Data Selection
Business Understanding
Classify the sentiments and predict the

Figure 2: Research Methodology

3.2 Dataset Selection

The Memotion 2.0 dataset [1], which is the second iteration of the Memotion task completed at Semeval 2020 [2] and is publicly accessible, has been chosen for this investigation. The task is broken down into three subtasks: Assignment A- Sentiment Classification: Provided an Ironic joke, the very first job is to categorize it as either a good, bad, or neutral meme, depending on how it makes you feel. B- Sentiment Categorisation: the algorithm must determine the sort of emotion that is communicated. The types are distinguished: hilarious, caustic, provocative, and inspirational statements. A meme can be classified into more than one genre. Using scales or intensities of sentiment classes, the third job attempts to measure the degree to which a specific feeling is being displayed. In total, there are roughly 7000 memes in this set of data. All of the 7000 memes will be used in this investigation.

3.3 Data Processing and Transformation

Memes are composed of two types of modalities, which are visuals and words. In terms of images, we use bilinear interpolation to reformat images of varying sizes to 224 224 pixels in dimension. We can see that there are various tags, HTML elements, non-alphabets, and any other form of symbols which may not be a fit in the language used in phrases when we look at the text information. As a result, we eliminate all keywords, weblinks, and figures from the text that do not appear to be relevant for memotion analysis. Next, all of the characters in writing are converted to lowercase, and phrases are lengthened to

[1] https://drive.google.com/drive/folders/1fC0EaJtwcfF28GnQBecIZJh$_{3}F0XNwHA$

[2] https://competitions.codalab.org/competitions/35688

match the size of the lengthiest sentence in the database.

3.4 Modelling

Following the processing and conversion of the dataset to the desired format, the data modeling program commences. In this phase, the models were built, trained, verified, and then assessed to determine their effectiveness. A total of five unique tests are carried out throughout the modeling step, with the data from each trial being processed and saved in various formats from the other. The algorithms are performed on Google Colab using the Pyhton library.

3.5 Evaluation Metrics

The following are some of the most frequently served terms, as well as their definitions -

- Precision: This column denotes the percentage of affirmative identifications that were perfectly right from the classified category of the data. A accuracy of 1.0 is achieved by a classifier that delivers no false positives.

- Recall: The percentage of correct positive that were successfully categorised out of the overall number of positives in that specific class is shown by the value of this variable. A recall of 1.0 is achieved by an algorithm that generates no false negatives.

- F1 score : The F1 score is a composite metric that represents the summation of prediction accuracy (precision and recall) into a single statistic. The harmonic mean is used to determine the F1 score. The F1 score for a successful model is one hundred percent.

- Support: The sample size on which each measure was computed.

- Macro avg: This measure generates the F1 by class, however, it does not employ weights in the aggregate of the scores. If the technique fails to perform effectively in classes with small sample sizes, the model will be penalized more severely as a result.

- Accuracy : Specifically, it is specified as ”the ratio of right assumptions made by a classifier to the overall number of estimates provided.”

$$\frac{TN + TP}{FN + FP + TN + TP}$$

- Weighted F1 Score: Using the weighted F1 score, each class’ F1 score is calculated separately, but when they are added collectively, the weight used is determined by how many true labels that class has.

- True Positive (TP) : The model successfully anticipates a positive class when it is in the appropriate category.

- True Negative (TN) : It occurs when the model properly identifies a negative class.

- False Positive (FP) : It occurs when the model suggests erroneously that a positive class exists.
- False Negative (FN) : It occurs when the algorithm wrongly estimates a negative class.

For the purpose of measuring model efficiency as a result of classification, the parameters of F1 Score, Precision, Recall, Weighted Avg are most crucially,Also computation time are explored.

3.6 Deployment

At this point, summaries of the evaluated results as well as a visual representation of the conclusion will be displayed. At this phase, we will acquire a picture of the model's efficiency based on selected data.

4 Design Specification

A full specification of the system design and technical design of the models executed will be provided in this part.

The design phase is divided into three primary tiers: the Database Tier, the Application Tier, and the Presentation Tier. This stage involves collecting and processing large amounts of data in order to eliminate unnecessary information and distortion. The information is then saved in a comma-separated excel sheet, which will be imported into Google colab later.

The application layer (logic layer) is primarily concerned with the business values that can aid in the achievement of the model's purpose. When dealing with memes, the business logic is to categorize and forecast the mood expressed by each meme based on its classification. The algorithms that are developed must accomplish that goal of precisely classifying the feelings and appropriately predicting the intensity of the meme by including business logic into their design. The attempts are carried out in the Google Colab environment. This dataset contains a variety of properties that are recognized, retrieved, and then evaluated using the models described above.

The findings of the deployment are reviewed with the consumers or clients during the presentation layer. The design is then published if satisfied with the findings.

5 Implementation, Evaluation and Results of Sentiment classification of Memes

5.1 Introduction

This segment will present a full study of the methods applied to estimate the emotion of memes, including their execution, assessment, and findings. In order to test the architecture, the development will be segmented into several trials.

5.2 Data Selection

In the Defactify-workshop hosted by codelab[3], we made use of the corpus given by memotion 2.0. It is referred to as Memotion Analysis in this assignment. The dataset includes of Memes and the textual information that goes along with them. Positive, neutral, and negative emotions are all represented by sub-emotions such as humor, sarcasm, offensiveness, and motivation. All memes are divided into three sentiment classes: positive, neutral, and negative. In Table 1 an example table is provided.

Table 1: Count of records on the basis of sentiment

Sentiment	**Train**	**Validation**
Positive	1517	564
Neutral	4510	279
Negative	973	71

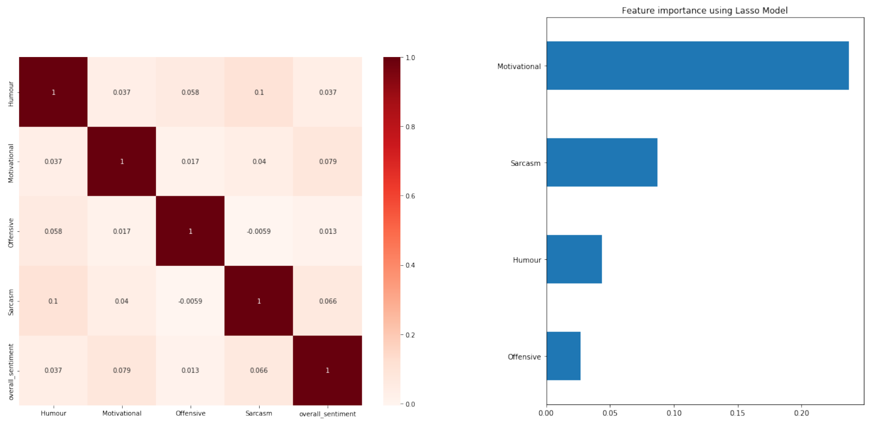
5.3 Data Processing, Data Transformation and Feature Extraction

Specifically, the database utilized for this challenge is the one supplied by the Defacify competition. It comprises around 7000 records, each of which has the following characteristics: image URL, OCR extracted text, corrected text, the amount to which it is humorous, insulting, and sarcastic in fuzzy words, whether this is inspirational or not, and the whole emotion of the picture. Many difficulties were encountered when working with a meme dataset, including words or sentences with different wording variants, small phrases with confusing grammar and syntax, images with no visuals and simple text, Links or HTML tags that were not properly formatted, and an unbalanced dataset. In order to address some of the concerns listed above, we went through the below preprocessing stages.

1. **Removal of Non-OCR text records:** Initially, any instance that did not comprise OCR extracted text and/or corrected text was deemed ineffective and so eliminated. Due to an anomaly in the data, several occurrences that did not have the Overall Sentiment scores experienced a shifting issue, with each value being incorrectly inserted in the column to the left of the aspect level sentiment value in each example. As a result, these occurrences were moved to the right once.

2. **Removal of HTML Tags, URLs, Punctuation Marks, Digits, and Non-ASCII Glyphs:** It was discovered that the revised text had a lot of noise, including repetitive Links, names, and irrelevant symbols. These background elements were

[3] https://competitions.codalab.org/competitions/35688



(a) Co-relation matrix of Training data

(b) Bar plot of sentiment feature

Figure 3: Feature importance of Training data

also eliminated from the context of the meme.

3. **Elimination of Stopwords:** Then, with the help of the nltk library, all of the stop words (such as am, is, and are) were eliminated.Then, the ambiguous adjectives used to define how funny, sardonic, insulting, and inspiring the memes are were standardized to a range between 0 and 1.

4. **Handling of Imbalanced Datasets:** Finally, when the feature selection technique of the data has been completed, We discovered that two of the four classified data points included in the study (Motivational and Sarcasm) are not as significant and conclusive as the other two. As a result, we did not include them in our input data. The Figure 3 are the graphics that are related .

5.4 Model Implementation

Keras's functional API is used to build the system that we developed. Depending upon the type of our dataset, three distinct types of input levels

- Meme's Picture
- Meme's Textual data
- Meme's Categorical data (humor, irony, rudeness, motivational)

are provided into the network. So we now have three distinct branches in our distribution network. Figure 4 shows a high-level overview of the deployment, analysis, and outcomes process.

5.4.1 Convolutional Neural Network(CNN)

Text may be expressed in a trainable vector form, making CNN techniques viable for text recognition. Considering that text may be expressed in such a way, CNN is also capable of

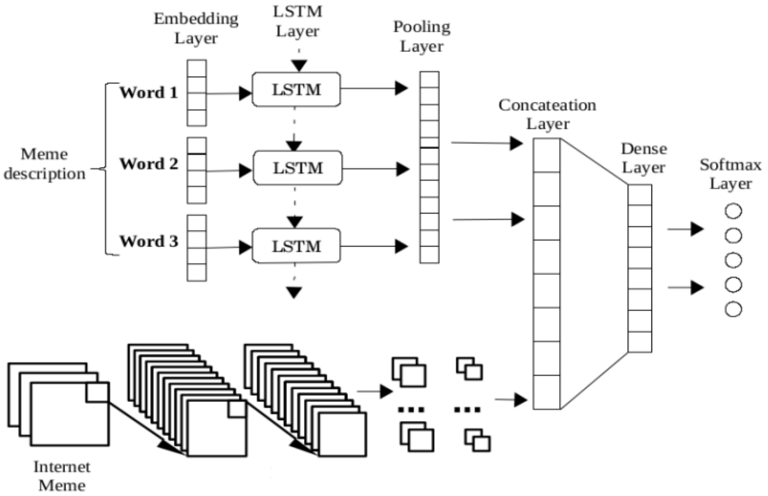


Figure 4: Architecture of Implementation

classifying text-based information. Specifically, two fundamental building pieces of CNN are utilized in this network. It has been utilized the convolutional layer with the max-pooling layer, with the outcome of the former linked to the source of the latter and vice versa. A total of four of these convolutional blocks were used prior to the classification stage. The flattening layer, which comes prior to the classification model, turns the vector into a single dimension, which is then used by the deep network that is fully connected. Last but not least, the outcome of this level has been concatenated to the last layer, with relu as the activation function of the first option.

5.4.2 Recurrent Neural Network(RNN)

The bag of words technique, in which each term is treated as an individual entity, does not retain the circumstances in which the term is used. By analyzing textual information as though it were a temporal sequence, Long short-term memory (LSTM) is a neural network that keeps the meaning of the phrase intact. The textual characteristic has been extracted with the help of LSTM. It preserves the necessary meaning from the sentences so that they can be retrieved and utilized afterward without having to deal with the problem of vanishing gradient descent (VGD). Three LSTMs are piled on top of each other in this strategy. It is possible to generate a more accurate depiction of the data using a layered LSTM algorithm. As a result, the result of one LSTM layer has been used to feed into another. In this architecture, stacked LSTMs are utilized as feature separators prior to the data being passed on to the classification stage, which is where the classification is done. Word embeddings are generated employing a Global Vectors for Word Representation (GloVe) sample that has already been trained. The application of previously trained word embedding maximizes the massive impact of the context-sensitive meaning of the phrase.

5.4.3 Multi-Layer Perceptron(MLP)

Following that, the category text data was supplied as input into an MLP, which was constructed as two dense layers with a Rectified Linear Unit (ReLU) activation function and one dropout layer. In this stage, the output is passed on as a parameter to the concatenated layer.

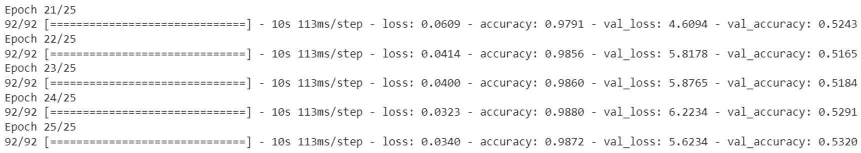
5.4.4 Final Dense Layer

Once the findings from each of these layers have been collected, they are aggregated and fed into a final MLP, which is capable of identifying the meme's overall emotion based on the information gathered from the linked sub-branches of the meme. A Dense layer serves as the output layer of our model; it comprises three neurons, one for each class count, which have been combined, and it is triggered using a Softmax activation method.

5.5 Experiment 1 – Baseline Model of Classification

Implementation: The baseline model is organized in the manner described in the preceding section. It is a product of three different layers, which encompass the LSTM, the RNN, and the MLP. The embedded sequence parameter is used as an input to the LSTM during the operation. In addition, there are two additional LSTM layers introduced to the branch. In order to gain a better understanding of the categorical data, we added an MLP layer, which is composed of four layers using Relu as an activation function. MLPs are made up of one or more levels of interconnected neurons together. Data is sent into the input nodes, and there could be one or more hidden layers that provide different abstraction levels. Forecasts are generated on the output nodes, also known as the visible layer, which is fed into the intake layer and hidden layers. CNN's have the ability to generate an internal structure of a two-dimensional picture from its external depiction. This enables the models to understand the placement and dimension of data structures that are different from one another, which is critical when dealing with pictures. We are using CNN to study from the meme's image and recover the relevant insight for the situation at hand. We utilized four Con2D layers with the Relu activation function in our design. As the last layer, the flattening layer has been placed. The textual data of the dataset was used to construct the RNN model that was used to fit the classifier. The stack of four dense layers is included in the merged branch. The first three layers make use of the Relu activation function, with the final layer making use of the softmax activation function. The model is applied using a batch size of 64 and 25 epochs in each iteration. The Adam optimizer was utilized to achieve the best possible outcomes.

Evaluation and Results: The total amount of training records we have includes 233 positive subgroups, 649 neutral classes, and 148 negative subgroups. Because of this, the baseline model for this research would be to label each meme as neutral; this would result in a validation accuracy of 53.20% when solving the problem. As seen in the Figure 5, the validation accuracy is 53.20%, which is significantly lower than the model performance of 98.72%. Consequently, the baseline model is overfitted to the dataset. The model, on the other hand, took 113 milliseconds to run. The operation on this model took a reasonable amount of time to complete.

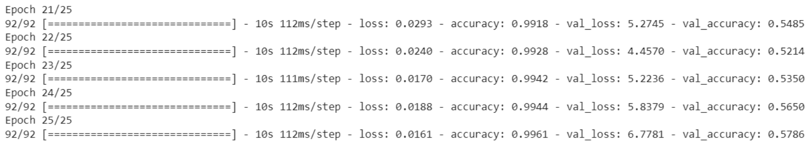


```
Epoch 21/25
92/92 [==============================] - 10s 113ms/step - loss: 0.0609 - accuracy: 0.9791 - val_loss: 4.6094 - val_accuracy: 0.5243
Epoch 22/25
92/92 [==============================] - 10s 113ms/step - loss: 0.0414 - accuracy: 0.9856 - val_loss: 5.8178 - val_accuracy: 0.5165
Epoch 23/25
92/92 [==============================] - 10s 113ms/step - loss: 0.0400 - accuracy: 0.9860 - val_loss: 5.8765 - val_accuracy: 0.5184
Epoch 24/25
92/92 [==============================] - 10s 113ms/step - loss: 0.0323 - accuracy: 0.9880 - val_loss: 6.2234 - val_accuracy: 0.5291
Epoch 25/25
92/92 [==============================] - 10s 113ms/step - loss: 0.0340 - accuracy: 0.9872 - val_loss: 5.6234 - val_accuracy: 0.5320
```

Figure 5: Baseline Model Result

5.6 Experiment 2 – Reduced Model of Classification

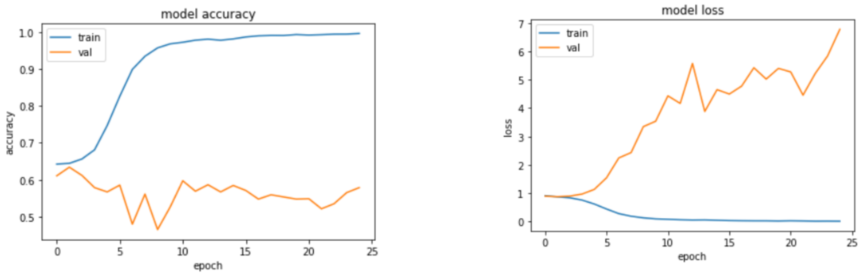
Implementation: When addressing overfitting, the first step we performed was to reduce the complexness of the algorithm we were working with. As shown in Figure 6, In the earlier basic model, there are four hidden Layers, one of which contains 64 neurons and the other three of which include 128 neurons. Furthermore, the output layer incorporates a softmax activation function. In aim is to lower the intricacy of our system, we simply deleted layers in terms of making it simpler in size. In terms of how much should be removed or how large our network should be, there is no basic guideline. To achieve a more compact model, we deleted the two hidden layers from the MLP model, as well as one Con2D and one max pooling from the CNN model. We eliminated two hidden layers from the ultimate combined layer in order to improve the model's quality. Figure 7, shows the accuracy and loss curves for the model under consideration.



```
Epoch 21/25
92/92 [==============================] - 10s 112ms/step - loss: 0.0293 - accuracy: 0.9918 - val_loss: 5.2745 - val_accuracy: 0.5485
Epoch 22/25
92/92 [==============================] - 10s 112ms/step - loss: 0.0240 - accuracy: 0.9928 - val_loss: 4.4570 - val_accuracy: 0.5214
Epoch 23/25
92/92 [==============================] - 10s 111ms/step - loss: 0.0170 - accuracy: 0.9942 - val_loss: 5.2236 - val_accuracy: 0.5350
Epoch 24/25
92/92 [==============================] - 10s 112ms/step - loss: 0.0188 - accuracy: 0.9944 - val_loss: 5.8379 - val_accuracy: 0.5650
Epoch 25/25
92/92 [==============================] - 10s 112ms/step - loss: 0.0161 - accuracy: 0.9961 - val_loss: 6.7781 - val_accuracy: 0.5786
```

Figure 6: Last Epochs of Reduced Model

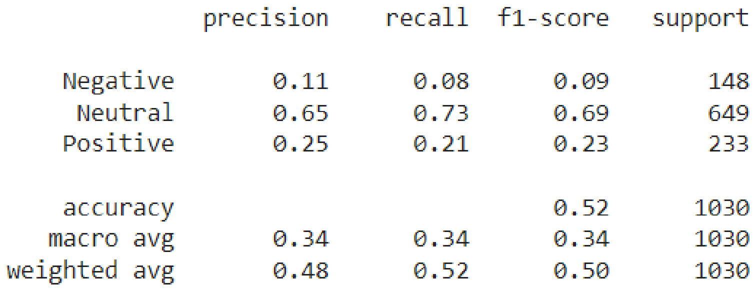
Evaluation and Results: As a starting point, we used a baseline model that was overfitting and had an accuracy of 98.72% at the time. We can see the outcomes of the revised model, which was trained on the input, in the section below. It's still not ideal, but as the graph shows, the classifier is overfitting significantly less than before. The new model receives F1 score of roughly 52%, which is much better than the baseline model. The Neutral category has a high recall score as well as a high precision value, according



(a) Accuracy graph of Reduced model

(b) Loss graph of Reduced model feature

Figure 7: Accuracy and Loss graph of Reduced model



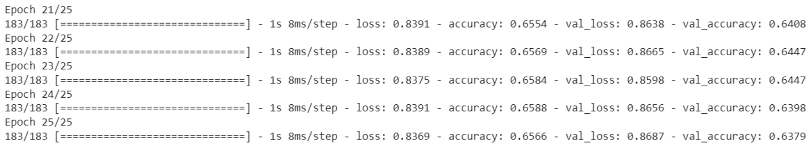
```
              precision    recall  f1-score   support

    Negative       0.11      0.08      0.09       148
     Neutral       0.65      0.73      0.69       649
    Positive       0.25      0.21      0.23       233

    accuracy                           0.52      1030
   macro avg       0.34      0.34      0.34      1030
weighted avg       0.48      0.52      0.50      1030
```

Figure 8: Reduced Model Classification Report

to the classification report Figure 8 . As the dataset includes a large number of memes, the bulk of them has a neutral sentiment. As a result, that metric cannot be used for evaluating the model. The weighted average value provided a highly balanced result in terms of both precision and recall.

5.7 Experiment 3 – Model with L2 Regularization

Implementation: We attempted to use certain regularization approaches in order to avoid the overfitting issue we encountered with our earlier approach. We employed L2 regularization techniques to CNN and RNN layers of our model.As a result of the inclusion of this regularization component, the number of weight matrices decreases, which is supported by the hypothesis that a neural network with a reduced weight matrix corresponds to linear models in general. Figure 9 As a result, it will drastically reduce overfitting to a statistically meaningful level. Ensemble models often outperform single models because they collect a greater amount of unpredictability than single models.

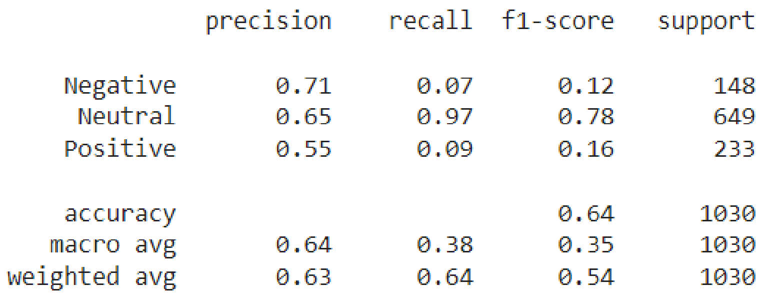


```
Epoch 21/25
183/183 [==============================] - 1s 8ms/step - loss: 0.8391 - accuracy: 0.6554 - val_loss: 0.8638 - val_accuracy: 0.6408
Epoch 22/25
183/183 [==============================] - 1s 8ms/step - loss: 0.8389 - accuracy: 0.6569 - val_loss: 0.8665 - val_accuracy: 0.6447
Epoch 23/25
183/183 [==============================] - 1s 8ms/step - loss: 0.8375 - accuracy: 0.6584 - val_loss: 0.8598 - val_accuracy: 0.6447
Epoch 24/25
183/183 [==============================] - 1s 8ms/step - loss: 0.8391 - accuracy: 0.6588 - val_loss: 0.8656 - val_accuracy: 0.6398
Epoch 25/25
183/183 [==============================] - 1s 8ms/step - loss: 0.8369 - accuracy: 0.6566 - val_loss: 0.8687 - val_accuracy: 0.6379
```

Figure 9: Last Epochs of L2 Regularized Model

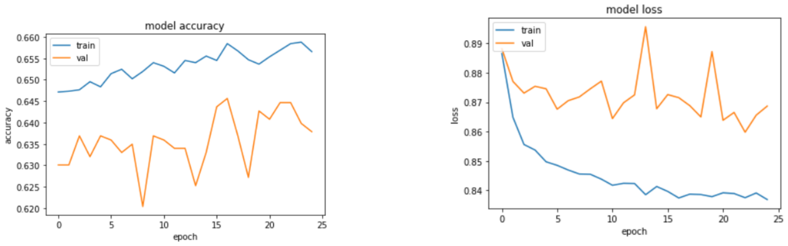
Evaluation and Results:

The result Figure 10 shows that 7% of the samples in the Negative class were correctly identified out of a total of 148 samples in the same category. Additionally, it shows that 71% of the samples in the Negative category are accurately classified by the classifier. The positive class has a precision score of 55%, which indicates that 55% of the samples are correctly classified as positive. Over a total of 1030 samples of all forms of sentiment, the classifier achieved an accuracy of 64% as measured by the F1 score, the highest accuracy achieved yet. Because our dataset is unbalanced, we must consider the weighted average for evaluation. i.e. 54%. Furthermore, the weighted average for both precision and recall is nearly identical.



```
              precision    recall  f1-score   support

    Negative       0.71      0.07      0.12       148
     Neutral       0.65      0.97      0.78       649
    Positive       0.55      0.09      0.16       233

    accuracy                           0.64      1030
   macro avg       0.64      0.38      0.35      1030
weighted avg       0.63      0.64      0.54      1030
```

Figure 10: Classification Report of L2 Regularized Model



(a) Accuracy graph of L2 Regularized Model

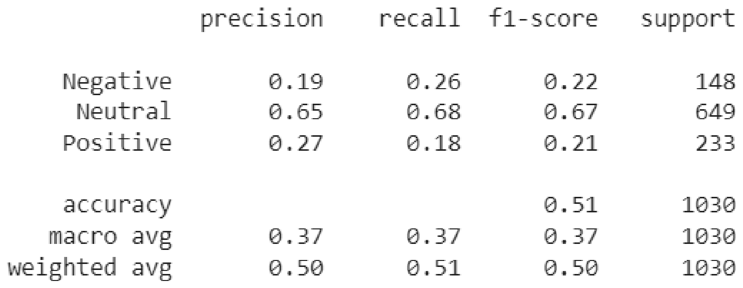
(b) Loss graph of L2 Regularized Model feature

Figure 11: Accuracy and Loss graph of L2 Regularized Model

5.8 Experiment 4 – Model with Dropout Layer

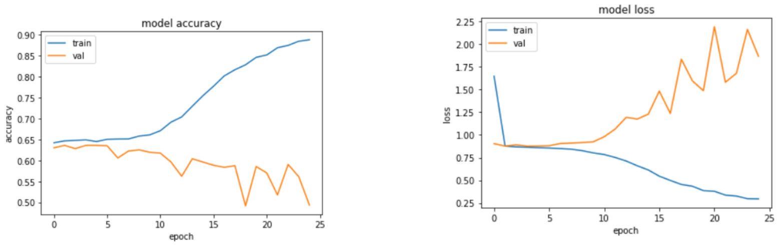
Implementation: we also included Dropout levels to make our prototype more realistic. Dropout outperforms a normal neural network model in terms of performance. The likelihood of deciding on the number of nodes to lose is the hyperparameter of the dropout function, which is defined as.

Evaluation and Results: Comparing the results from the previous report, the Model with Dropout layer correctly identified the Negative and Positive memes in 26% and 18% of the total number of samples in the corresponding class, respectively. Additionally, the report Figure 12 indicates that the F1 score has increased for both the negative and positive categories, but that it has declined for the neutral category. Classier received a



```
              precision    recall  f1-score   support

    Negative       0.19      0.26      0.22       148
     Neutral       0.65      0.68      0.67       649
    Positive       0.27      0.18      0.21       233

    accuracy                           0.51      1030
   macro avg       0.37      0.37      0.37      1030
weighted avg       0.50      0.51      0.50      1030
```

Figure 12: Classification Report of Dropout Layer Model



(a) Accuracy graph of Dropout Layer Model

(b) Loss graph of Dropout Layer Model feature

Figure 13: Accuracy and Loss graph of Dropout Layer Model

51% overall F1 score. In this case, it appears that the inclusion of a dropout layer has made no discernible impact on the final outcome.

5.9 Experiment 5 – Model with Early stopping, Learning Rate scheduler, Checkpoints

Implementation: First and foremost, because the number of cases in the negative category is significantly lower than the count of occurrences in the remaining two classes, we calculated and applied class weights in terms of balancing the dataset. After that, we made advantage of some of the attributes provided by Keras' callbacks package. Consider the following example: we built a function that decreased the methodologies rate of learning on every 30 epochs. Checkpoints were also employed to prevent us from increasing our weights when the model began to overfit.
Early stopping is a type of cross-validation approach in which we preserve a portion of the training sample as the validation data while discarding the rest of the training dataset. When we notice that the model's performance on the validation set is deteriorating, we instantly cease the model's learning. This is referred to as "early stopping."

Evaluation and Results: According to the classification report Figure 14 , memes with Neutral emotions were most accurately categorized by the algorithm. The algorithm correctly detected 36% of Memes with a positive attitude, which is much better than the previous results. The recall weighted average is slightly lower than the precision for the

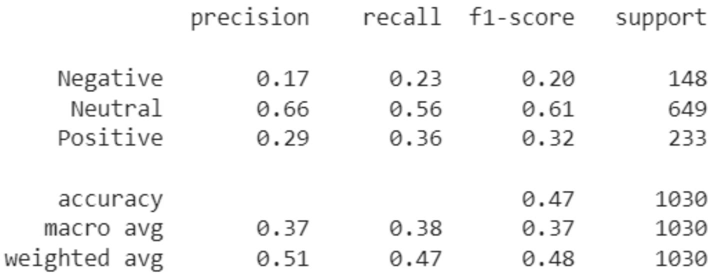
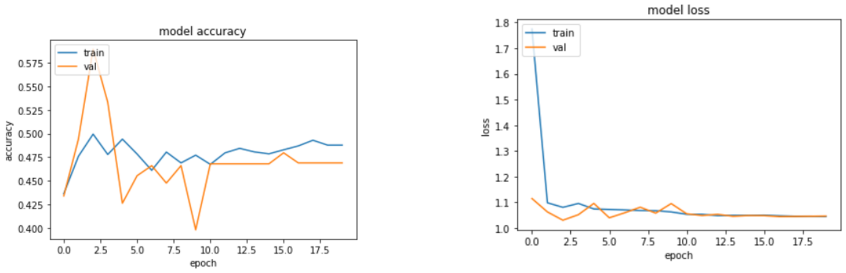

```
              precision    recall  f1-score   support

    Negative       0.17      0.23      0.20       148
     Neutral       0.66      0.56      0.61       649
    Positive       0.29      0.36      0.32       233

    accuracy                           0.47      1030
   macro avg       0.37      0.38      0.37      1030
weighted avg       0.51      0.47      0.48      1030
```

Figure 14: Classification Report of Model with Early stopping, Learning Ratescheduler, Checkpoints



(a) Accuracy graph of Model with Early stopping, Learning Ratescheduler, Checkpoints

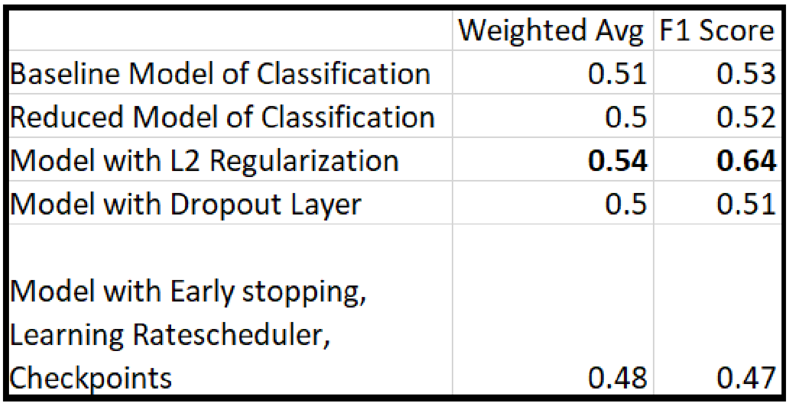
(b) Loss graph of Model with Early stopping, Learning Ratescheduler, Checkpoints feature

Figure 15: Accuracy and Loss graph of Model with Early stopping, Learning Ratescheduler, Checkpoints

provided model. On the basis of the weighted average F1 score, the model correctly determines the emotion of the meme 48% of the time.

5.10 Model Comparison

The numbers in Figure 16 shows a condensed view of the weighted average and F1 score for all five experiments. Except for the last model, all of the models exhibited reasonable accuracies (50% or more). The model with early stopping, learning rate scheduler, and checkpoints, on the other hand, exhibited a remarkably low accuracy and weighted avg. For the given dataset, the model using L2 regularization produces the best results.



	Weighted Avg	F1 Score
Baseline Model of Classification	0.51	0.53
Reduced Model of Classification	0.5	0.52
Model with L2 Regularization	**0.54**	**0.64**
Model with Dropout Layer	0.5	0.51
Model with Early stopping, Learning Ratescheduler, Checkpoints	0.48	0.47

Figure 16: Models comparison

Figure 17 depicts a graphical depiction of all generated model performance, with the weighted average and F1 score as performance evaluation metrics for each model produced.

5.11 Error Analysis

There are numerous aspects that influence the classification model's overall outcomes. The chosen dataset is significantly imbalanced, which may be the root reason for the

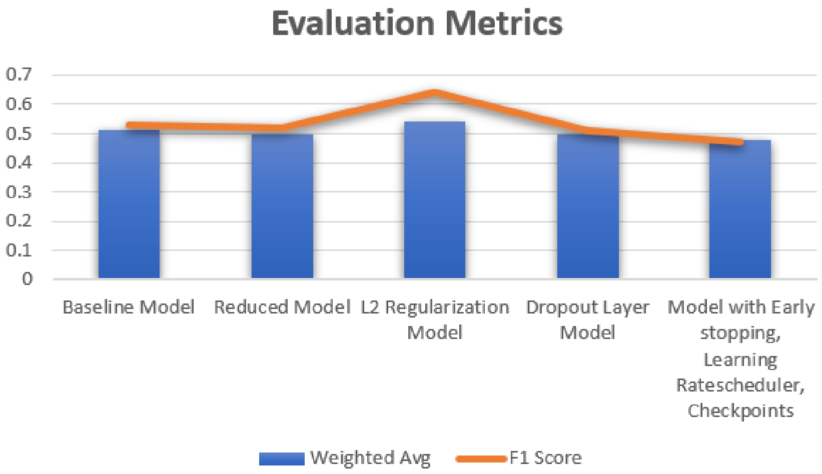


Figure 17: Performance Evaluation using Bar Plot

reduced accuracy. Furthermore, as we have seen in our tests, unbalanced data can lead to overfitting. Because of the environmental setting, the outcome may vary. After preprocessing and feature extraction, the dataset's size was reduced. As a result, the amount of training data can have a negative impact on the output. Large training datasets can help enhance results by allowing the model to learn more.

5.12 Result and Discussion

No matter how many efforts we made, our model was never able to attain an accuracy of more than 64%. However, the most significant enhancement over the prior technique is the eventual narrowing of the difference between training and validation precision; this is an indication that our network is no longer overfitting in the same way as before and that its prediction performance has increased. Figure 10 illustrates the accuracy and loss charts for this model. Figure 10 shows the classification report for the model, as well as the model's categorization outcome. This result provides us with significant information about how the system performed on each category. For instance, it is visible that hardly any of the values belonging to class 0 i.e. Negative were analyzed properly by the algorithm. Given the extremely limited amount of instances available in that subclass, this was a foreseeable bad situation. As previously stated, our correctness on the validation data is 64.37%, which is a significant improvement.

6 Conclusion and Future Work

As previously stated in the preceding subsection, this work is difficult due to the fact that it involves humor, which is strongly tied to how individuals see and comprehend. Even distinct persons may have differing perspectives on the subject of humor. It occurs frequently that something that one person finds amusing becomes insulting to the other. As a result, robots have a tough time learning and categorizing topics that are linked to them. The fact that we were able to obtain greater accuracy than the baseline approach in our research should be noted, though. This suggests that, in the case of a sufficient dataset, Deep Learning algorithms may be used to handle this issue; this is the principal cause why our approach could not obtain the improved outcomes. Despite the fact that we were able to significantly minimize overfitting, the fact that a number of parameters for a Deep Learning assignment are extremely limited means that it is unable to acquire more valuable frequent patterns and, as a result, overfits in the final. It is necessary to either expand the number of training data we have or utilize alternative ways and attempt more techniques in order to evade this, which we were unable to achieve.

Specifically, we will focus on difficulties such as shorter meme descriptors, free term sequencing in phrases, extra functionality to recognize sentiments in memes, the irony in depictions, and so on. We would like to investigate deeper neural network systems that are capable of capturing the comedy and sarcasm seen in Internet memes in the near future.

References

Blackmore, S. and Blackmore, S. J. (2000). *The meme machine*, Vol. 25, Oxford Paperbacks.

Blandfort, P., Patton, D. U., Frey, W. R., Karaman, S., Bhargava, S., Lee, F.-T., Varia, S., Kedzie, C., Gaskell, M. B., Schifanella, R. et al. (2019). Multimodal social media analysis for gang violence prevention, *Proceedings of the International AAAI conference on web and social media*, Vol. 13, pp. 114–124.

Bollen, J., Mao, H. and Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, *Proceedings of the international AAAI conference on web and social media*, Vol. 5.

Borth, D., Ji, R., Chen, T., Breuel, T. and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs, *Proceedings of the 21st ACM international conference on Multimedia*, pp. 223–232.

Chandrasekaran, A., Vijayakumar, A. K., Antol, S., Bansal, M., Batra, D., Zitnick, C. L. and Parikh, D. (2016). We are humor beings: Understanding and predicting visual humor, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4603–4612.

Chen, T., SalahEldeen, H. M., He, X., Kan, M.-Y. and Lu, D. (2015). Velda: Relating an image tweet's text and images, *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Davidov, D., Tsur, O. and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys, *Coling 2010: Posters*, pp. 241–249.

Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition, *Competition and cooperation in neural nets*, Springer, pp. 267–285.

Gajarla, V. and Gupta, A. (2015). Emotion detection and sentiment analysis of images, *Georgia Institute of Technology* pp. 1–4.

Guadagno, R. E., Rempala, D. M., Murphy, S. and Okdie, B. M. (2013). What makes a video go viral? an analysis of emotional contagion and internet memes, *Computers in Human Behavior* **29**(6): 2312–2319.

Guo, Y., Huang, J., Dong, Y. and Xu, M. (2020). Guoym at semeval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 1120–1125.

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

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural computation* **9**(8): 1735–1780.

Hu, A. and Flaxman, S. (2018). Multimodal sentiment analysis to explore the structure of emotions, *proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pp. 350–358.

Joshi, A., Bhattacharyya, P. and Carman, M. J. (2017). Automatic sarcasm detection: A survey, *ACM Computing Surveys (CSUR)* **50**(5): 1–22.

Joshi, M., Prajapati, P., Shaikh, A. and Vala, V. (2017). A survey on sentiment analysis, *International Journal of Computer Applications* **163**(6): 34–38.

Kumar, A. and Garg, G. (2019). Sarc-m: Sarcasm detection in typo-graphic memes, *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India.*

Lee, Y., Yoon, S. and Jung, K. (2018). Comparative studies of detecting abusive language on twitter, *arXiv preprint arXiv:1808.10245* .

Milo, T., Somech, A. and Youngmann, B. (2019). Simmeme: A search engine for internet memes, *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, IEEE, pp. 974–985.

Mina, A. X. (2019). *Memes to movements: How the world's most viral media is changing social protest and power*, Beacon Press.

Oliveira, H. G., Costa, D. and Pinto, A. M. (2016). One does not simply produce funny memes!–explorations on the automatic generation of internet humor, *Proceedings of 7th International Conference on Computational Creativity*, pp. 238–245.

Peirson V, A. L. and Tolunay, E. M. (2018). Dank learning: Generating memes using deep neural networks, *arXiv preprint arXiv:1806.04510* .

Phillips, W. and Milner, R. M. (2017). Decoding memes: Barthes' punctum, feminist standpoint theory, and the political significance of# yesallwomen, *Entertainment Values*, Springer, pp. 195–211.

Procházka, O. (2016). Cohesive aspects of humor in internet memes on facebook: A multimodal sociolinguistic analysis, *Ostrava Journal of English Philology* **8**(1): 7–38.

Raskin, V. (2012). *Semantic mechanisms of humor*, Vol. 24, Springer Science & Business Media.

Sabat, B. O., Ferrer, C. C. and Giro-i Nieto, X. (2019). Hate speech in pixels: Detection of offensive memes towards automatic moderation, *arXiv preprint arXiv:1910.02334* .

Shifman, L. (2014). *Memes in digital culture*, MIT press.

Taecharungroj, V. and Nueangjamnong, P. (2014). The effect of humour on virality: The study of internet memes on social media, *7th International Forum on Public Relations and Advertising Media Impacts on Culture and Social Communication. Bangkok, August*.

Truong, Q.-T. and Lauw, H. W. (2017). Visual sentiment analysis for review images with item-oriented and user-oriented cnn, *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1274–1282.

Van Hee, C., Lefever, E. and Hoste, V. (2018). We usually don't like going to the dentist: Using common sense to detect irony on twitter, *Computational Linguistics* **44**(4): 793–832.

Verma, D., Chandiramani, R., Jain, P., Chaudhari, C., Khandelwal, A., Bhattacharjee, K., ShivaKarthik, S., Mithran, S., Mehta, S. and Kumar, A. (2020). Sentiment extraction from image-based memes using natural language processing and machine learning, *ICT Analysis and Applications*, Springer, pp. 285–293.

Wang, W. Y. and Hua, Z. (2014). A semiparametric gaussian copula regression model for predicting financial risks from earnings calls, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1155–1165.

Williams, A., Oliver, C., Aumer, K. and Meyers, C. (2016). Racial microaggressions and perceptions of internet memes, *Computers in Human Behavior* **63**: 424–432.

Yoshida, K., Minoguchi, M., Wani, K., Nakamura, A. and Kataoka, H. (2018). Neural joking machine: Humorous image captioning, *arXiv preprint arXiv:1805.11850* .

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), *arXiv preprint arXiv:1903.08983* .