Classification of Hate/Sarcasm in Memes

INFO 5082

JAMMULA KEERTHI CHANDANA
UNIVERSITY OF NORTH TEXAS

UNDER THE GUIDANCE OF
JUNHUA DING

Introduction:

Since the advent of Facebook, the number of social media sites has grown, making communication easier for billions of internet users. Anyone may influence anyone through different means of communication, such as posting a text, a picture, or a meme to express their thoughts, thanks to the internet's new medium. These tools are helping a lot of people to increase their freedom of speech, which can be good and bad as well due to access that they have on the platform. For example, Facebook has billions of users, which makes Facebook vulnerable to things like hate speech, abusive content, and fake news. Facebook is attempting to monitor the dissemination of hate speech using Artificial Intelligence, which assists in detecting hate speech, primarily in memes, by training a machine learning model with a large amount of meme data. Sarcasm is often misinterpreted as hate speech. We created a classifier to detect hate/sarcasm in memes in order to assist companies such as Facebook and Twitter in maintaining a clean social media site.

This report aims to define key terms or features that aid in the detection of hateful/sarcastic material, as well as demonstrate how to create a classifier that can recognize both text and image content at the same time, as humans do when viewing a meme. Even humans have a difficult time deciphering a meme because they must understand the meaning of the meme through the text or image, which is usually based on internet slang. In terms of computers, detecting sarcasm can be difficult because recognizing sarcasm also requires meaning, which is difficult to obtain with current technology. For this project, we used a Facebook dataset with 6000 memes and a Twitter dataset with 1000+ memes to evaluate the classification model. We divided the memes into five categories and labeled them as sarcastic or non-sarcastic (racial, gender, nationality, religion, others).

Keywords like Muslim, goat, girl, government, kill, fuck, and trans were extracted using topic modeling. These keywords appear in a large number of memes in our dataset, and they represent a sample of how hateful memes are spread on social media. Trigrams and named entity recognition can be used as layers of filters in flagging memes for further approval, according to our research.

Problem Definition and Significance:

Because of the abusive usage of social media sites, public questions about social media platforms have grown in recent years. Facebook, for example, is actively investing in its services in order to keep them useful and insightful. Facebook has faced several problems as a result of promoting hatred during recent social movements in the United States. For all social media platforms, detecting hate/sarcasm has become a necessity. Furthermore, according to 2019 figures, 38% of social media users follow meme accounts, 55% of people aged 13 to 35 send memes every week, and 75% of people send memes to others to respond to something. Because of the internet's and social media's accessibility, cyberbullying and trolling have risen dramatically. Previous research has linked increased social media use to improved mental wellbeing. Cyberbullying and trolling have a huge impact on large number of teenagers. Cyberbullying also takes the form of memes with a sarcastic tone, which can be harmful to the victim's mental health. To keep this under control, existing technologies should be able to detect sarcastic memes that might have a negative impact on social media users. Sarcasm is often misinterpreted as hate speech and detecting sarcasm may assist social media moderators in accurately identifying hate speech. As a result, developing a sarcasm/hate classifier can be extremely beneficial to companies like Facebook and Twitter in terms of maintaining a platform.

The current efficient models are unimodal, which works well with only text or image as a mode, making solving this problem difficult and interesting. However, several memes are made up of two or more modals; the textual and image combination is interpretable, and either one alone does not produce the desired results. This multimodal nature, combined with benign confounders, makes meme identification much more difficult.

The aim of this study is to see how well a classifier can detect sarcasm/hateful memes using text and image features in both unimodal (text only) and multimodal (text and image data) fashion.

Review on Prior Literature:

Detecting hate speech in multimodal memes (Facebook):

By building a dataset from scratch where multimodal prevails, the Facebook AI research team has created a multimodal machine learning model to detect multimodal hate memes rather than unimodal hate memes in this paper. Using a pre-trained model on the COCO dataset, they were able to achieve an AUC of 0.71. We learned about the different types of models that can be used for multimodal datasets.

An Empirical Analysis of Text Superimposed on Memes Shared on Twitter:

The aim of this paper is to draw insights from memes that have text superimposed on them. They were able to analyze the memes by extracting the text using optical character recognition and identifying objects in the image using pre-trained models. We learned how to extract useful features from images using pretrained models and how to extract text using OCR.

Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text:

They concentrated on how to use the early fusion technique to combine all modalities in a meme (text and image) to identify a meme and were able to achieve good baseline scores. We learned about early fusion and late fusion techniques from this research, which are widely used to understand different modalities.

Exploring Hate Speech Detection in Multimodal Publications:

This paper claims that text content outperforms current multimodal, which is counterintuitive when compared to the Facebook study. To complete this mission, they annotated a broad dataset and discovered that image content is helpful in determining meaning, but it was unable to outperform text + image combined performance. In multimodal research, they used the Inception V3 pre-trained model for images and the LSTM for text. We were encouraged by this paper to work on textual models and compare them to the traditional multimodal.

Beyond Visual Semantics: Exploring the Role of Scene Text in Image Understanding:

IIT Jodhpur conducted a fun study in which they attempted to produce semantics by defining the objects in an image and text. They were able to improve image context detection using multi-channel visual semantics and textual information, resulting in richer image representations. This work has motivated us to expand on the existing work we have done for this project. Because of its richer representation, generating semantic representation of an image can be used as a modality for a classifier.

Data Sourcing:

We extracted memes dataset from Twitter and collected memes data from Facebook AI study hateful memes dataset. We categorized the memes into sarcastic/non-sarcastic and race, religion, nationality, gender, and other categories to see what types of memes are commonly used to spread hate. In our memes, sarcasm is often sarcastic, but it is highly context based. To give an example, the text on the meme could be funny even if it has nothing to do with the background image, but the image could be funny even if the text is usual (very few cases).

Data Preparation/cleaning:

Memes are difficult to analyze because of their complex representation of image and text formats. To clean the text for text analysis, we had to use a multiple pre-processing techniques. We compiled a list of any method that has been used to address why it is required and how it is implemented.

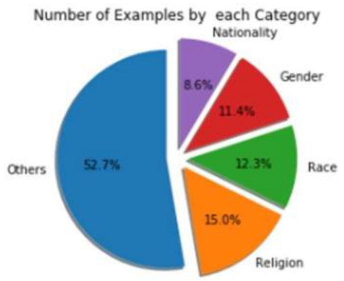
- Because many memes contain spelling mistakes, we used a language model to fix them, which can explain the context better than a dictionary-based approach.
- We also cleaned the text and removed punctuation, numbers, and non-ascii characters.

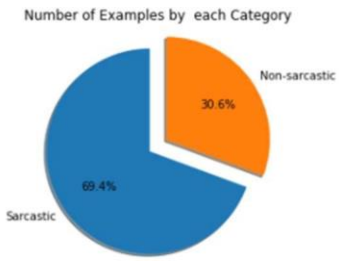
After the cleaning process, we have a clean corpus, but we still believe that some memes' text needs to be cleaned.

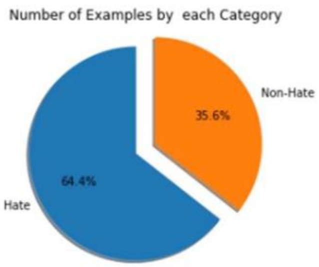
Exploratory Data Analysis/Visualizations:

We conducted an extensive exploratory analysis on the cleaned corpus to learn what types of memes, words or word combinations people use to propagate sarcastic and hateful memes.

Distribution of Dataset Labels by category:

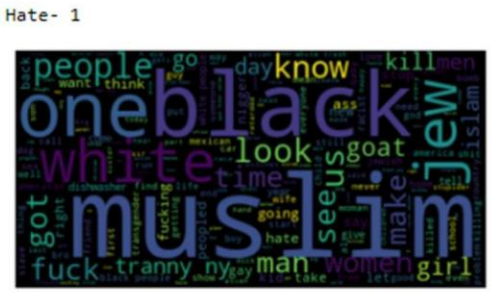


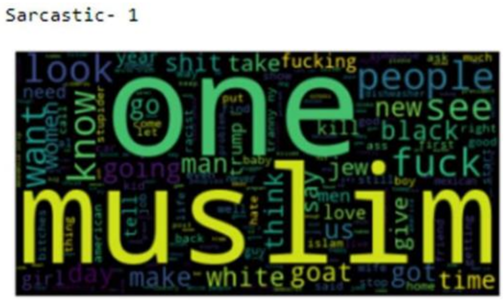




The pie charts give us the dataset's distribution across three categories: hate/non-hate, sarcasm/non-sarcasm, and categories. Each of the categories in the dataset is imbalanced. In our dataset, the ratios of hate to non-hate and sarcastic/non-sarcastic are a little high. Religious/racial groups have a higher number of memes. The majority of the other memes are about politics, children, adults, etc.

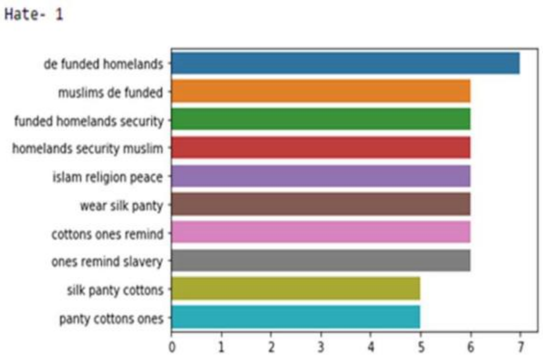
Word Clouds/N-grams/Topic Modelling:

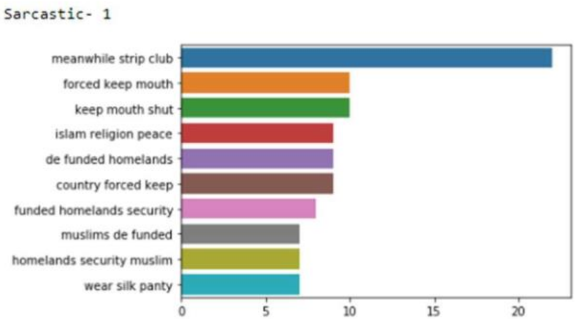




The variation in words used in sarcastic/hate memes can be seen in the word clouds above. Words like Muslim, black, and goat can be found commonly in both types of memes. This makes it difficult for the model to distinguish between sarcastic and hateful material.

N-grams: Bi-grams and trigrams were analyzed through Hateful/Not-Hateful, sarcastic/non-sarcastic memes, and category categorization. We discovered that various bigrams and trigrams are used in different categories, and that trigrams are more representative and interpretable than bigrams. These may be used as another layer of filtering when it comes to detecting meme categories.



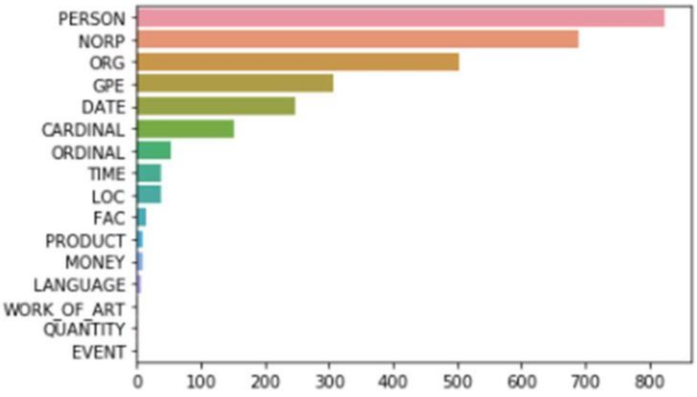


Topic Modelling: We used Latent Discriminatory Analysis, an unsupervised Topic Modelling technique, to analyze the most frequently discussed topics through memes and were able to roughly categorize the memes into four topics based on the most frequently used Words.

Topic	10 most used words in order of frequency
Topic-1	man, women, home, men, country, trump, way, bitch, goat, guy, shit
Topic-2	muslim, day, time, look, people, problem, racist, peace, women, religion, goat
Topic-3	people, child, kid, tranny, life, everyone, fucker, race, society, jew, thing, tell, goat
Topic-4	friend, girl, fuck, islam, car, need, strip, part, news, hand, club, obama, bomb, life, attack

Topic-1 is related to Gender, Topic-2 is related to Religion, Topic-3 is related to a mix of Gender and Religion, and Topic-4 is related to commonly used Hate words, as seen in the table above.

Named Entity Recognition: We used the most commonly used named entities in the memes using a pre-built named entity recognition from the spacy library. We discovered that the majority of memes were about persons, led by NORP (Nationality or religion, or Political Groups), ORG, and GPE. Hitler, Obama, and Hillary Clinton, for example, have all been used to create political memes.



Text and Image Classification & Results:

We used two unimodal (input: text only) and two multimodal (input: Image + text) models to classify a meme as sarcastic/hateful. Multimodal machine learning models outperformed textual models by a small margin. As discussed below, we converted the text and image into vectors.

Text vectorization:

We used an urban dictionary pre-trained fasttext model to get a Word2Vec representation from a cleaned corpus because urban dictionary includes Internet slang words and representations are slightly better than Google's word2vec and Facebook fasttext.

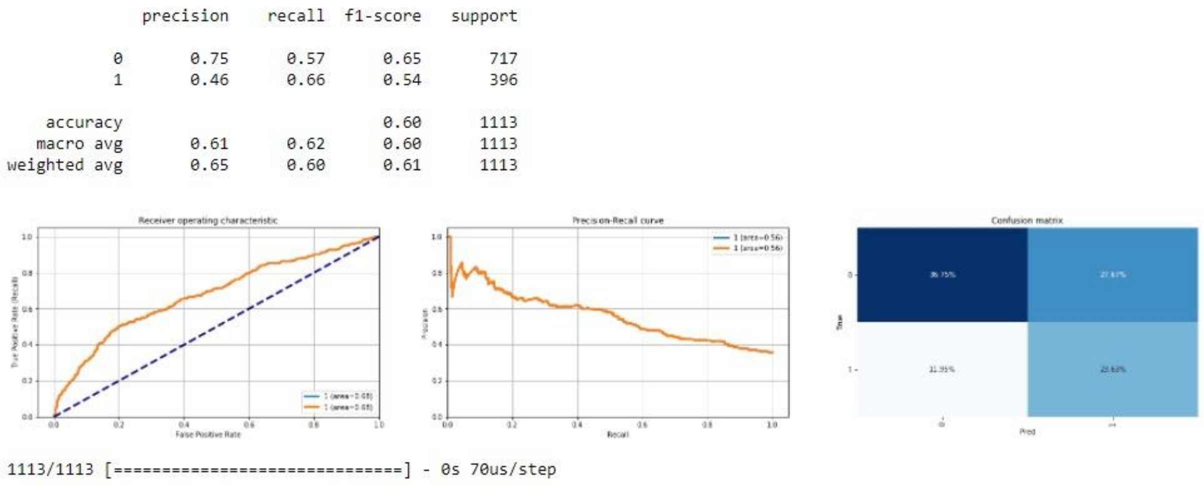
Image input:

We transformed all of the images into one image and trained one of the multimodal models with a pre-trained model (Resnet101). In the other model, along with the text, we used high-level representations of an Image (objects + labels) as a modality to the model.

Models:

LSTM on Text-Only:

Sarcasm Model:

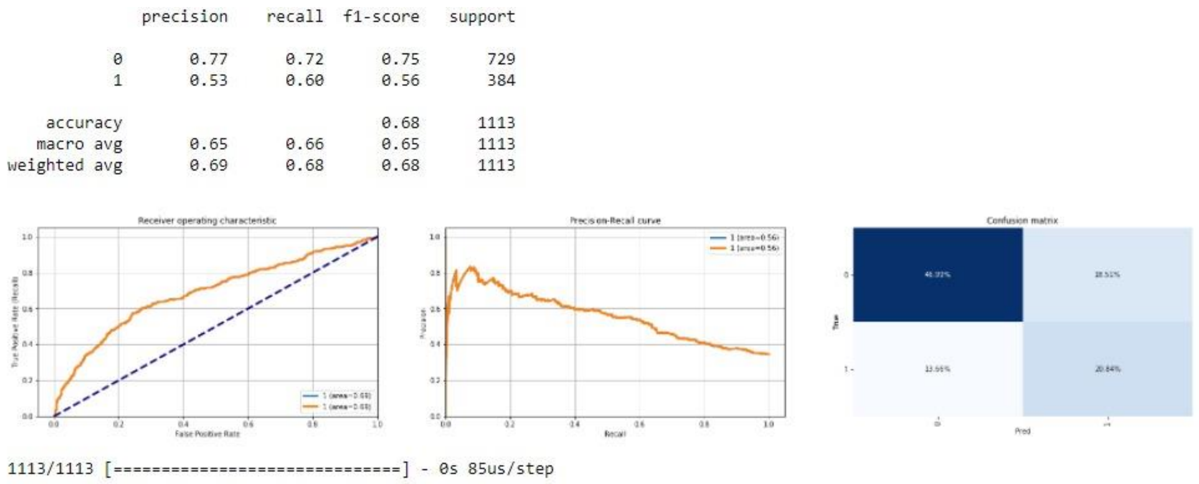


```
              precision    recall  f1-score   support

           0       0.75      0.57      0.65       717
           1       0.46      0.66      0.54       396

    accuracy                           0.60      1113
   macro avg       0.61      0.62      0.60      1113
weighted avg       0.65      0.60      0.61      1113
```

```
1113/1113 [==============================] - 0s 70us/step
```

From the above image, we can observe the classification report and confusion matrix for LSTM text-only on Sarcasm model.

Hate Model:

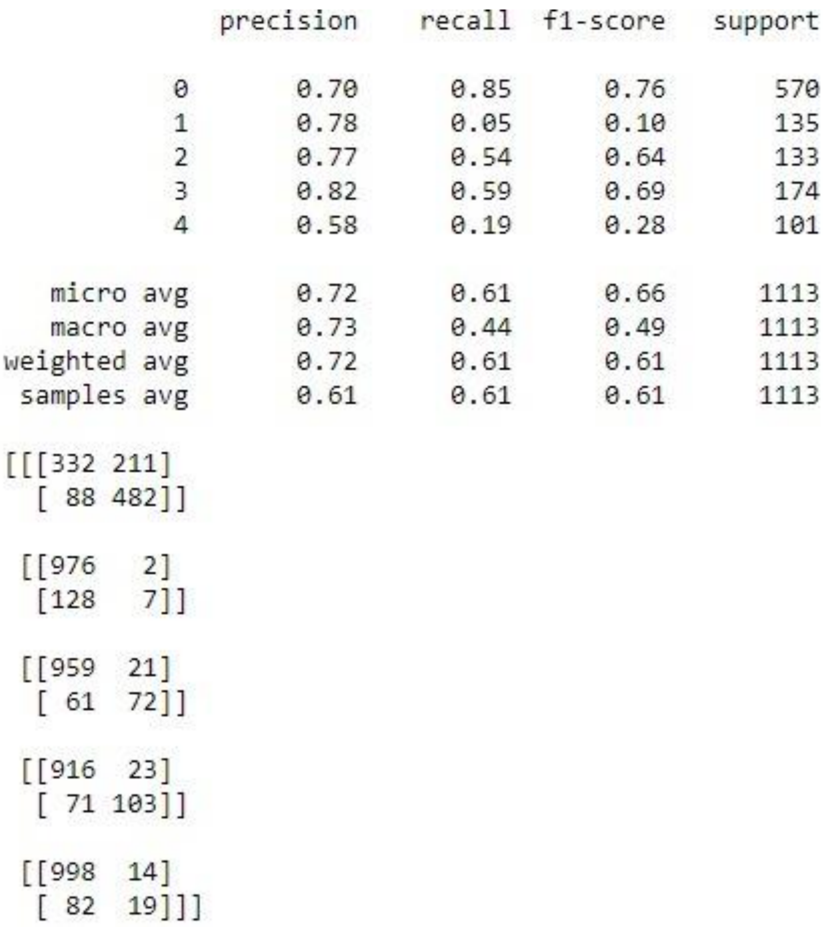


```
              precision    recall  f1-score   support

           0       0.77      0.72      0.75       729
           1       0.53      0.60      0.56       384

    accuracy                           0.68      1113
   macro avg       0.65      0.66      0.65      1113
weighted avg       0.69      0.68      0.68      1113
```

```
1113/1113 [==============================] - 0s 85us/step
```

From the above image, we can observe the classification report and confusion matrix for LSTM text-only on Hate model.

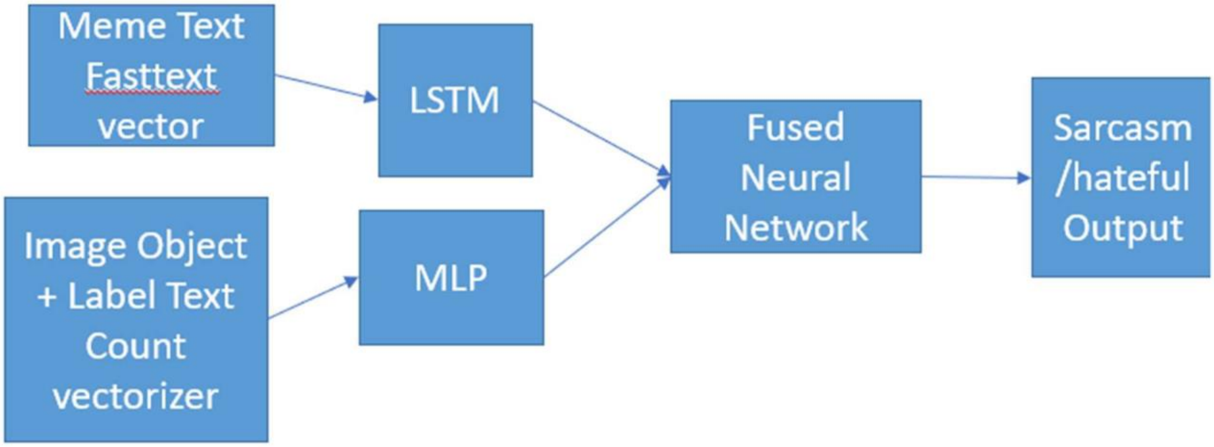
Category:



```
              precision    recall  f1-score   support

           0       0.70      0.85      0.76       570
           1       0.78      0.05      0.10       135
           2       0.77      0.54      0.64       133
           3       0.82      0.59      0.69       174
           4       0.58      0.19      0.28       101

   micro avg       0.72      0.61      0.66      1113
   macro avg       0.73      0.44      0.49      1113
weighted avg       0.72      0.61      0.61      1113
 samples avg       0.61      0.61      0.61      1113

[[[332 211]
  [ 88 482]]

 [[976   2]
  [128   7]]

 [[959  21]
  [ 61  72]]

 [[916  23]
  [ 71 103]]

 [[998  14]
  [ 82  19]]]
```

From the above image, we can observe the classification report and confusion matrix for LSTM text-only on Category model.

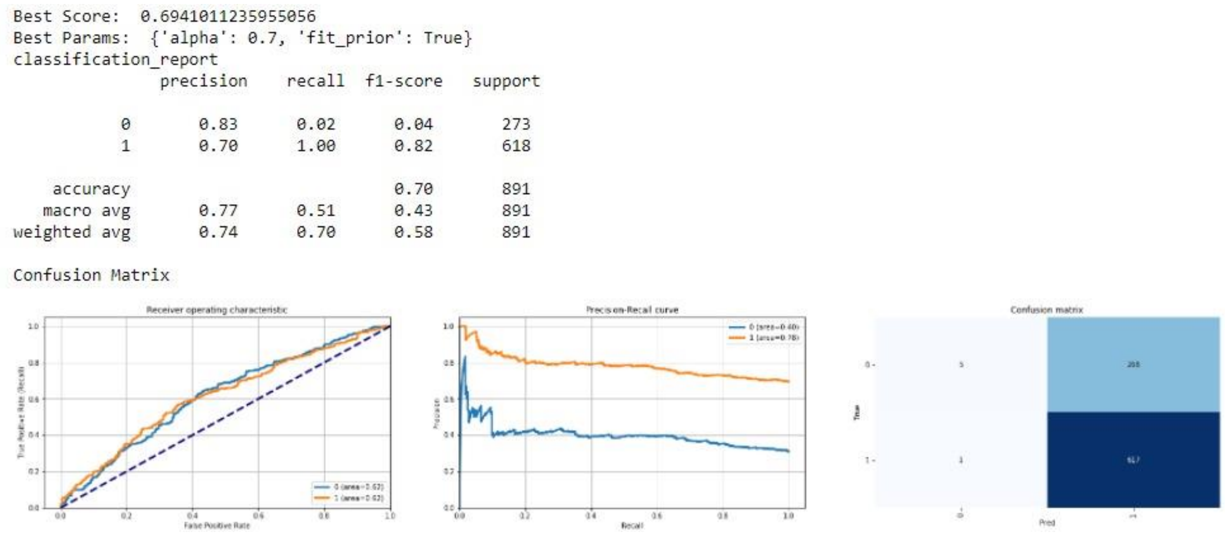
Multimodal-1:


Meme Text Fasttext vector
LSTM
Image Object + Label Text Count vectorizer
MLP
Fused Neural Network
Sarcasm /hateful Output

We created the first multimodal-1 using LSTM for text and labels+objects as Image features without the image as input in a late fusion technique (combining features after training the classifier with text and image features separately), which decreases the neural network's training time. When compared to the text-only model, we saw a slight improvement in accuracy/f1 score.

Naïve Bayes - Base Model:

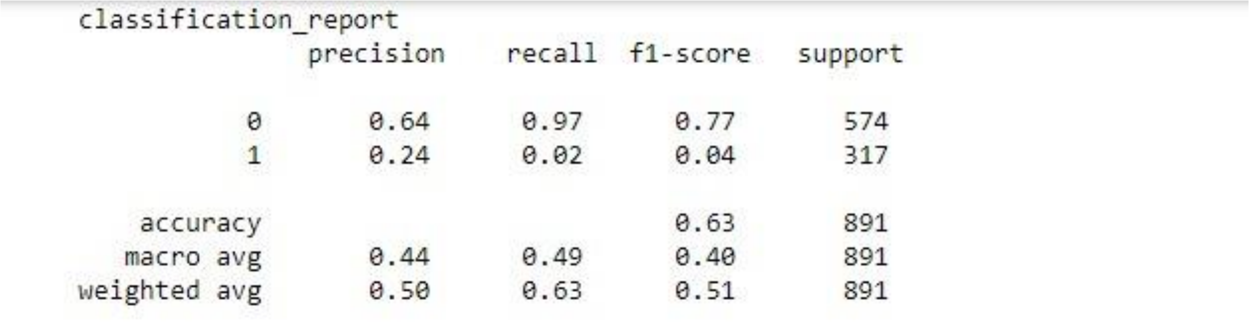
Sarcasm label:



```
Best Score:  0.6941011235955056
Best Params:  {'alpha': 0.7, 'fit_prior': True}
classification_report
              precision    recall  f1-score   support

           0       0.83      0.02      0.04       273
           1       0.70      1.00      0.82       618

    accuracy                           0.70       891
   macro avg       0.77      0.51      0.43       891
weighted avg       0.74      0.70      0.58       891

Confusion Matrix
```

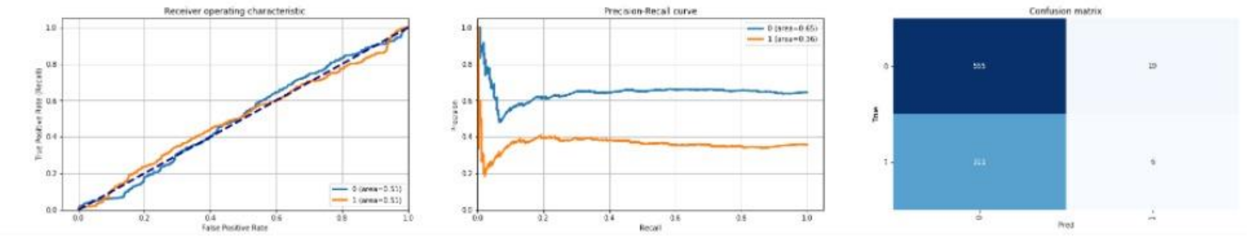
From the above image, we can observe the classification report and confusion matrix for Naïve Bayes Model on Sarcasm model.

Hate Label:



```
classification_report
               precision    recall  f1-score   support

           0       0.64      0.97      0.77       574
           1       0.24      0.02      0.04       317

    accuracy                           0.63       891
   macro avg       0.44      0.49      0.40       891
weighted avg       0.50      0.63      0.51       891
```

Confusion Matrix



From the above image, we can observe the classification report and confusion matrix for Naïve Bayes Model on Hate model.

Results:

Model Type	**F1 score(weighted)**
Naive Bayes(unimodal) - baseline model	Sarcasm: 0.58 Hateful: 0.51
LSTM (Text only)	Sarcasm: 0.61 Hateful: 0.68
LSTM(Text) + MLP (Imagefeatures (Objects+lables))	Sarcasm: 0.61 Hateful: 0.63

F1 score interpretation: F1 score is a metric used to assess classification models. The ability of a model to correctly detect memes will improve as the f1 score increases. For our dataset, the best model is LSTM (text only), which has a high f1 score.

Conclusion, Future work, and Recommendations:

Our classification model is similar to any other neural network and understanding a black box model is difficult because we have little control about how it operates. Our research yielded few business insights/recommendations. Our aim was to create a more accurate sarcastic/hateful classification system. The following are some of the main findings from our exploratory data analysis:

- According to our findings, bigrams and trigrams vary significantly between hateful and non-hateful memes, as well as sarcastic and non-sarcastic memes. Trigrams are more representative of meme itself, and they can be used to find groups, religions, and specific genders (black lives matter)
- Named entity recognition/topic modeling tells us which words/people are most commonly used in memes to portray hate/sarcastic content. This can be used as a second layer of filtering to identify memes.
- **Recommendation:** On social media platforms, trigrams can be used as a filter to flag memes (hateful or non-hateful, sarcastic, or non-sarcastic).
- Our best model was able to offer high f1 scores of 0.61 and 0.68 for sarcastic and hateful classifiers, respectively, based on our classification results, which can aid in accurately detecting memes.
- **Future work:** Using better OCR techniques will improve text quality, which will aid in improving meme detection accuracy. The model's robustness in predicting hate/sarcasm labels can be improved by adding more data points. This research contributes to a reduction in the societal impact of hate/sarcastic content on social media.

References:

- The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. https://arxiv.org/abs/2005.04790
- Early detection of promoted campaigns on social media. https://doi.org/10.1140/epjds/s13688-017-0111-y
- SESAM at SemEval-2020 Task 8: Investigating the relationship between image and text in sentiment analysis of memes." (2020). https://www.cs.kent.ac.uk/people/staff/mg483/documents/bonheme20SemEval2020.pdf
- Exploring Hate Speech Detection in Multimodal Publications - https://arxiv.org/abs/1910.03814
- Beyond Visual Semantics: Exploring the Role of Scene Text in Image Understanding - https://arxiv.org/abs/1905.10622
- MuSe 2020 – The First International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop - https://arxiv.org/abs/2004.14858
- Urban Dictionary Embeddings for Slang NLP Applications https://www.aclweb.org/anthology/2020.lrec-1.586.pdf
- Citation for Facebook pre-built model: https://www.drivendata.co/blog/hateful-memes-benchmark/