ISM 6930: Text Analytics

# Meme Classification

## Hate and Sarcasm


Submitted by
Nagarjuna Kanneganti
Sai Suraj Argula
Ashrit Kulkarni


Submission Date
9th November 2020

Project Guide
Dr. Anol Bhattacherjee

# Contents

# Executive summary

Social media platforms have increased since the inception of Facebook, which has made the communication easier to billions of users who have access to the internet. This new internet medium has become so powerful that anyone can influence anyone through various forms of communication like posting a text or an image or a meme to convey their thoughts. 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 trying very hard to control the spread of hate speech through the help of Artificial Intelligence which helps in detecting hate speech, mainly in memes by training a machine learning model with a lot of memes data. Oftentimes, sarcasm can be misunderstood as hate speech. We built a classifier to detect hate/sarcasm in memes to help companies like Facebook and twitter to keep their social media platform clean.

This report aims to identify the key words or features that helps in flagging the hateful/sarcastic content and also shows how to build a classifier to understand text content and image content simultaneously like humans do when they see a meme. Interpreting a meme is a difficult job even for humans as they must understand the context of a meme through the text or image which is generally built on internet slang. As for the machines, detecting sarcasm can be very hard because oftentimes understanding sarcasm needs context and it is very difficult to get the context using the current technology. For this project, we have taken a dataset from Facebook containing 6000 memes and extracted 1000+ memes from twitter for testing the classification model. We have labelled the memes as sarcastic/non-sarcastic and segregated the memes into 5 categories(racial, gender, nationality, religion, others).

We used topic modelling to extract keywords like Muslim, goat, girl, government, kill, fuck, trans. These keywords represent a lot of memes in our dataset and reflects the sample of how hateful memes are on social media. Our analysis found that trigrams and named entity recognition can be used as layers of filters in flagging memes for further approval.

# Problem definition & significance

In recent years, societal concerns of social media platforms have increased due to the abusive use of social media. Companies like Facebook are investing heavily into their platforms to keep them useful and informative. Facebook has faced a lot of challenges from spreading hate during some social movements that happened in the USA. Detecting hate/sarcasm has become the need of the hour, for all the social media platforms. In addition, 2019 statistics show that 38% of the social media users follow meme accounts, 55% of ages between 13-35 send memes every week and 75% people send memes to others to react to something. Cyberbullying and trolling have increased exponentially due to the accessibility of the internet as well as social media tools. Prior studies suggest that increase in usage of social media is proportional to mental health. Lot of teens are getting affected by cyberbullying or trolling. Oftentimes, cyberbullying happens in the form of memes with a lot of sarcastic tone which can hurt them mentally. To control this, current technology should be able to detect sarcastic memes which could have a potential effect on social media users in a negative way. Most of the time, sarcasm can be mistaken as hate speech and detecting sarcasm can help social media moderators to correctly identify hate speech. Hence, building a sarcasm/hate classifier can really help companies like Facebook or twitter in maintaining the platform.

Solving this problem is difficult and interesting as the current efficient models are unimodal which works well with only text or image as a mode. But often memes are the combination of two or more modal, the textual and Image combination is interpretable and either of one alone could not give appropriate results. This Multimodal nature, along with benign confounders makes detection of memes even more difficult.

This study aims to detect sarcasm/hateful memes by using text and image features in unimodal(using only text) and multimodal(using text and image data) fashion to see the performance of the classifier in detecting sarcasm/hate.

# Prior Literature Review:

We studied multiple research papers and reviewed multiple literature to get a good understanding of how multimodal machine learning works. Here are the summaries of the paperer's we reviewed.

### Detecting hate speech in multimodal memes(Facebook):

In this paper, Facebook AI research team has developed a multimodal machine learning model to detect multimodal hate memes rather than unimodal hate memes by creating a dataset from scratch where multimodal prevails. They were able to achieve an AUC of 0.71 using a pre-trained model on COCO dataset. We learned what types of models can be applied for multimodal datasets.

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

This paper focuses on extracting insights from memes which has text superimposed on the meme. They were able to analyze the memes using optical character recognition to extract the text and pre-trained models to identify objects in the image. We learned how to use pretrained models to extract useful features from images and OCR to extract text.

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

In this study, they focused on how to combine both modalities in a meme(text and image) to classify a meme using early fusion technique and were able to get good baseline scores. From this study, we got the idea of early fusion and late fusion techniques which are commonly used to understand multiple modalities

### Exploring Hate Speech Detection in Multimodal Publications

This paper suggests that text content itself is outperforming current multimodals and it is little counterintuitive when compared to the Facebook paper. They annotated a large dataset to do this task and they found out that image content is useful in getting the context right but was unable to give better results than text+image combined. They have used inception V3 pre trained model for image and LSTM for text in multimodal analysisFrom this paper, we were inspired to work on textual models and tried to compare it with standard multimodals.

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

This is a fun study done by IIT jodhpur, as they have tried to generate semantics by identifying the objects in an image and text. They were able to improve the detection of context of an image using multi-channel visual semantics and textual content to get richer representations of an image. This work inspired us to build on future work on the current work we have done for this project. Generating semantic representation of an image can be used as a modality for a classifier due to its richer representation.

# Data Sourcing:

We gathered the memes data from Facebook AI research hateful memes dataset and extracted 1000+ memes from twitter. We labelled the memes as sarcastic/non-sarcastic and categorized into race, religion, nationality, gender, and others to see what kinds of memes are used generally to spread hate. Sarcasm is often sarcastic but it heavily context dependent in our memes. To exemplify it, text on the meme itself could be funny without any relation to the background image but on the contrary, image could be funny with text being normal(very few cases)

**Sarcasm labelling technique:** When the annotator laughs or thinks it is funny to him, he labelled them as sarcastic or non-sarcastic. This technique could introduce a little bias in our labelling technique, but we will try to average the decision of sarcastic/non sarcastic by having multiple annotators.

Our dataset has true text, extracted text, labels and objects representing the image and hate/sarcasm labels.

| Image_id | Hate_label | FB_text | google_raw_text | label_list | object_list | Semi_cleaned_text | Sarcasm_label | Category | label_objects |
|---|---|---|---|---|---|---|---|---|---|
| 42953.png | 0 | its their character not their color that matters | its their character\nnot their color\nthat mat... | Photo caption, Forehead, Internet meme, Font, ... | Person, Person, Clothing | its their character not their color that matters | 0 | 3 | Photo caption, Forehead, Internet meme, Font, ... |

## Data Preparation/cleaning:

Analyzing memes is very hard due to its complex representation in the format of image and text. To analyze text, we used Tesseract to extract text but the results were not satisfactory. To get more clearer text, we used Google VISION API to extract text and objects of an image. Even Google API struggled to achieve 100% accuracy in extracting text. So, we had to apply multiple pre-processing techniques to clean the text for text analysis. We listed each technique that has been used to address why it is needed and how is it done with usage.

- Lot of memes has words like(af, asf, I'm)--> we used custom contractions & language model(pycontractions- glove twitter model) to fix the contractions as there are so many internet slang words, it is better to use glove twitter to resolve contractions

  Usage: I'd -> I would, or I had -> to use the right contraction, language model can understand the context and fix it accordingly rather than a dictionary based approach.

- Word segmentation must be done as there were few memes with confusing text with mix up of words adjacent to each other(we have used sym spell dictionary method to tackle this)

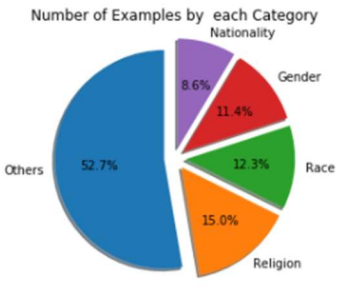
  Usage: “Icanthink ofthis” → “I can think of this”
- A lot of memes generally have spelling mistakes, to restore the context, we have used language model to fix it, which can understand the context better than a dictionary-based approach
- We also cleaned the text by removing punctuations, numbers, non ascii characters and tokenizing.

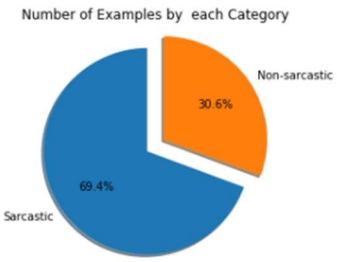
We built out a cleaned corpus after all the cleaning and we still feel that text still needs cleaning for some memes as Google API didn’t do well for memes which have complex fonts or words from different languages.

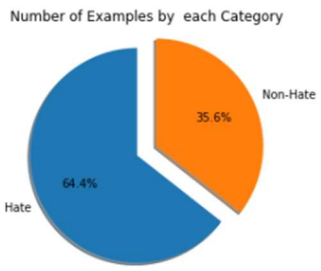
# Exploratory Data Analysis/Visualizations:

To get a general idea of the data, we did an extensive exploratory analysis on cleaned corpus to understand what kinds of memes, words or word combinations that people are using to spread in sarcastic and hateful memes.

### Distribution of Dataset Labels by category



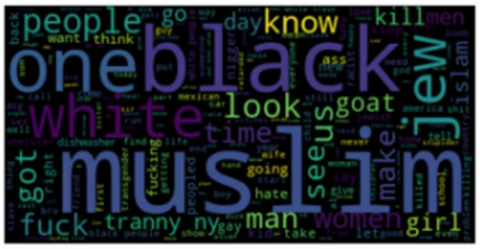




The pie charts show the distribution of the dataset across categories - Hate/Non-Hate, Sarcasm / Non-Sarcasm and Categories. The dataset is imbalanced across each of the categories. The ratio of hate to Non-hate and sarcastic/non-sarcastic is little high in our dataset, respectively. We

have a greater number of memes in religious/racial categories. Most of the others, memes are related to politics , children ,adults etc.

### Word Clouds/N-grams/Topic Modelling:

Hate- 1

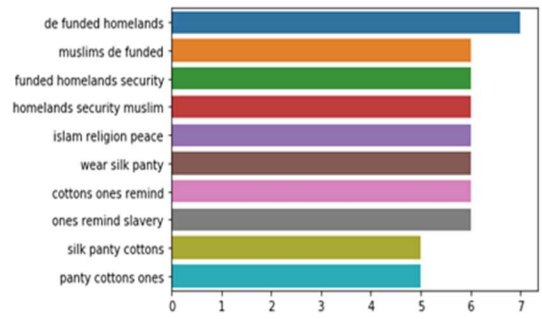



Sarcastic- 1

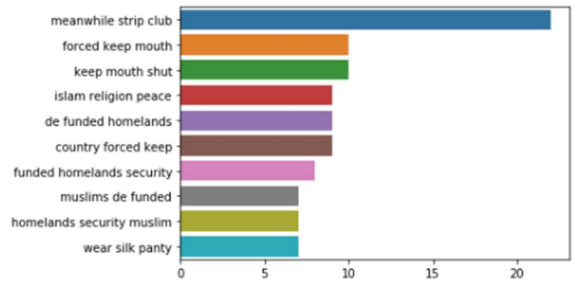

From Above Word clouds we can see the difference in words being used in sarcastic/hate memes. We can see that words like Muslim, black, goat are frequently used in both kinds of memes. This makes it very hard for the model to separate sarcastic from hate content

**N-grams:** We analyzed Bi-grams and trigrams across Hateful/ Not-hateful, sarcastic/non-sarcastic memes and category categorization. We found that there are different bigrams and trigrams used in categories, also we found that the trigrams were more representative and interpretable compared to bigrams. These could become another layer of filtering in detecting the meme categories.

Hate- 1




Sarcastic- 1

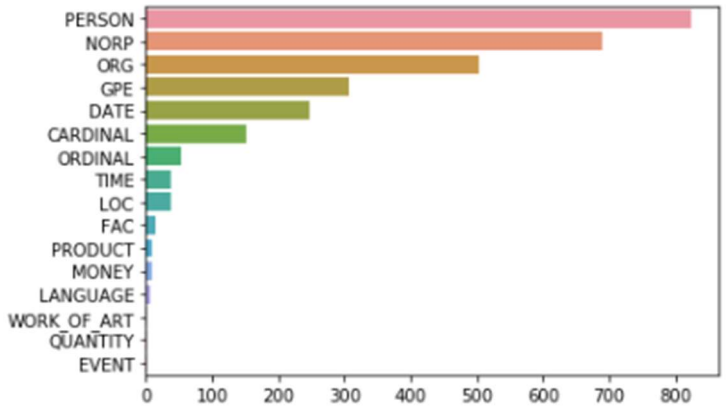



**Topic Modelling:** We analyzed topics that were most discussed across memes with help of Latent Discriminatory Analysis - an unsupervised Topic Modelling technique and were able to approximately categorize the memes into 4 topics based on the most frequent Words used.

| 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 |

From the above table, we can see that the Topic-1 is related to Gender, Topic-2 is related to Religion, Topic-3 is related to a mix of Gender / Religion , Topic-4 is related to generally used Hate words.

**Named Entity Recognition:** We used a pre-built named entity recognition from spacy library to check what are the most frequently used named entities in the memes. We observed that memes mostly talked about persons, followed by NORP(Nationality or religious or political groups),ORG, GPE. For example, Hitler, Obama and Hillary Clinton were used to construct political memes.



# Text/Image Classification & Results:

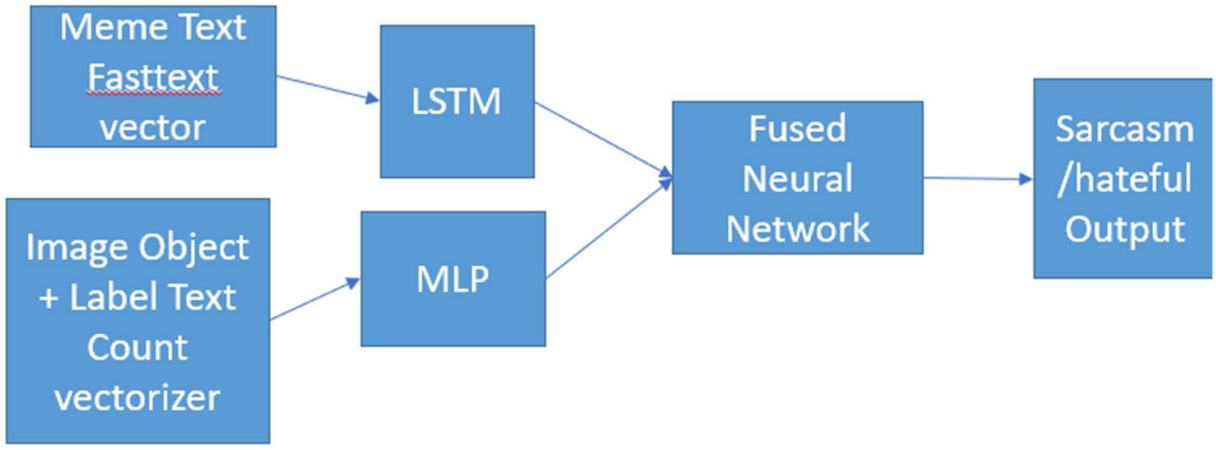
We built four models to classify a meme as sarcastic/hateful, where we used two unimodal (input: text only) and two multimodal(input: Image + text). We found out that multimodal machine learning models outperformed textual models by a little. We have transformed the text and Image into vectors as explained below

**Text vectorization:** We used a pre-trained fasttext model on urban dictionary to get a Word2Vec representation from cleaned corpus as urban dictionary contains Internet slang words and representations are little better than Google's word2vec and facebook fasttext.

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

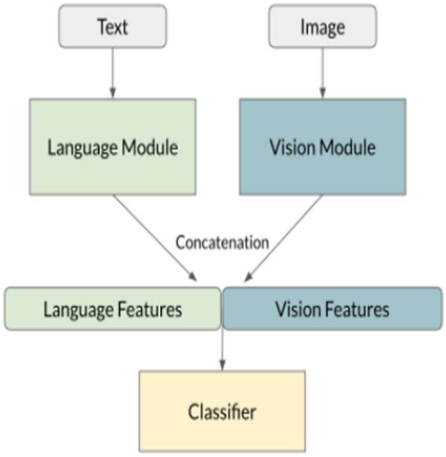
**Models:**

**Multimodal-1:**




We have designed the first multimodal-1 in a late fusion technique(combining features after training the classifier with text and image features separately) using LSTM for text and labels+objects as Image features without the image as input, which reduces the training time of the neural network. We observed the improvement in accuracy/f1 score by a little bit when we compared with the text only model.

### Multimodal-2(Facebook's pre-built model):




Multimodal-2 is inspired from Facebook's multi modal framework, **we have used the required code from Facebook AI research community to see the difference between our own model(Multimodal-1) and Facebook's model(Multimodal-2) performance on detecting sarcasm/hateful memes.** We used fasttext for extracting text features, ResNet(Residual Network) with 101 layers for Image features and combined the features using torch modules from PyTorch.

### Results:

| **Model Type** | **F1 score(weighted)** |
|---|---|
| Naive Bayes(unimodal) - baseline model | Sarcasm: 0.58<br>Hateful: 0.51 |
| LSTM(Text only) | Sarcasm: 0.61<br>Hateful: 0.68 |
| LSTM(Text) + MLP(Image features(Objects+lables)) | Sarcasm: 0.61<br>Hateful: 0.63 |
| Fasttext(text) + ResNet101(Image) | Sarcasm: 0.56<br>Hateful: 0.55 |

F1 score interpretation: F1 score is a metric to evaluate classification models. As f1 score increases, the ability of a model will improve in detecting the memes correctly. Our best model for our dataset is LSTM(text only) which has high f1 scores.

# Insights & Recommendations/future work:

Our classification model is pretty much like any other neural network and interpreting a blackbox model is very challenging as we don't have much control on how it works. We don't have many business insights/recommendations on our analysis. Our goal was to build a better sarcastic/hateful classifier. Some of the insights from our exploratory data analysis are as follows

1. Our analysis suggests that bigrams and trigrams are quite different for hateful and non-hateful memes as well as for sarcastic and non-sarcastic memes. ing the memes. Trigrams are more representative of meme nature and these can be used to find groups, religions and specific gender which are targeted(black lives matter)
2. Named entity recognition/topic modelling tells us about what kind of words/ who is referred mostly in memes to represent hate/sarcastic content. This can be a second layer of filter to flag the memes
3. **Recommendation:** Trigrams can be used as a filter to flag memes(hateful or non-hateful or sarcastic/non-sarcastic ) on social media platforms
4. Based on our classification results, our best model was able to give high f1 scores of 0.61 and 0.68 for sarcastic and hateful classifiers respectively improving this can help detecting memes accurately.

**Future work:** Using better OCR techniques will improve the text quality, which helps in improving the accuracy in flagging the memes. Increasing data points can increase the model's robustness in predicting hate/sarcasm labels. This analysis helps decrease the societal impact of hate /sarcastic content on social media.

# References:


1. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. https://arxiv.org/abs/2005.04790
2. Early detection of promoted campaigns on social media. https://doi.org/10.1140/epjds/s13688-017-0111-y
3. 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
4. Exploring Hate Speech Detection in Multimodal Publications - https://arxiv.org/abs/1910.03814
5. Beyond Visual Semantics: Exploring the Role of Scene Text in Image Understanding - https://arxiv.org/abs/1905.10622
6. MuSe 2020 – The First International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop - https://arxiv.org/abs/2004.14858
7. Urban Dictionary Embeddings for Slang NLP Applications https://www.aclweb.org/anthology/2020.lrec-1.586.pdf
8. Citation for Facebook pre-built model: https://www.drivendata.co/blog/hateful-memes-benchmark/
9. Citation for Keras Multimodal Implementation: https://www.pyimagesearch.com/2019/02/04/keras-multiple-inputs-and-mixed-data/
10. MMF: A multimodal framework for vision and language research https://github.com/facebookresearch/mmf


# Appendix

1. **GitHub Link**- The code used for this project is stored at GitHub repository. https://github.com/ARGULASAISURAJ/Meme-Classification
2. Google Vision API is used to extract Labels, Objects, and text. https://cloud.google.com/vision/docs/drag-and-drop
3. Google Collab is used to run the Facebook's multimodal PyTorch implementation on GPU https://colab.research.google.com/notebooks/intro.ipynb
4. Images Link in Google drive - https://drive.google.com/drive/folders/1QCbkMCfQGm1_-kYKiI7LdJ1Hzv932N0a?usp=sharing