Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 32–41
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

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

Shardul Suryawanshi, Bharathi Raja Chakravarthi,
Mihael Arcan, Paul Buitelaar
Insight SFI Research Centre for Data Analytics
Data Science Institute, National University of Ireland Galway
{shardul.suryawanshi, bharathi.raja, mihael.arcan, paul.buitelaar}@insight-centre.org

Abstract

A meme is a form of media that spreads an idea or emotion across the internet. As posting meme has become a new form of communication of the web, due to the multimodal nature of memes, postings of hateful memes or related events like trolling, cyberbullying are increasing day by day. Hate speech, offensive content and aggression content detection have been extensively explored in a single modality such as text or image. However, combining two modalities to detect offensive content is still a developing area. Memes make it even more challenging since they express humour and sarcasm in an implicit way, because of which the meme may not be offensive if we only consider the text or the image. Therefore, it is necessary to combine both modalities to identify whether a given meme is offensive or not. Since there was no publicly available dataset for multimodal offensive meme content detection, we leveraged the memes related to the 2016 U.S. presidential election and created the MultiOFF multimodal meme dataset for offensive content detection dataset. We subsequently developed a classifier for this task using the MultiOFF dataset. We use an early fusion technique to combine the image and text modality and compare it with a text- and an image-only baseline to investigate its effectiveness. Our results show improvements in terms of Precision, Recall, and F-Score. The code and dataset for this paper is published in `https://github.com/bharathichezhiyan/Multimodal-Meme-Classification-Identifying-Offensive-Content-in-Image-and-Text`

Keywords: multimodal data, classification, memes, offensive content, opinion mining

1. Introduction

A meme is "an element of a culture or system of behavior passed from one individual to another by imitation or other non-genetic behaviors"[1]. Memes come in a wide range of types and formats including, but not limited to images, videos, or twitter posts which has an increasing impact on social media communication (French, 2017; Suryawanshi et al., 2020). The most popular form of content corresponds to memes as images containing text in them. Due to the multimodal nature of the meme, it is often difficult to understand the content from a single modality (He et al., 2016). Therefore, it is important to consider both modalities to understand the meaning or intention of the meme. Unfortunately, memes are responsible for spreading hatred in society, because of which there is a requirement to automatically identify memes with offensive content. But due to its multimodal nature, memes which often are the combination of text and image are difficult to regulate by automatic filtering.

Offensive or abusive content on social media can be explicit or implicit (Waseem et al., 2017; Watanabe et al., 2018; Rani et al., 2020) and could be classified as explicitly offensive or abusive if it is unambiguously identified as such. As an example, it might contain racial, homophobic, or other offending slurs. In the case of implicit offensive or abusive content, the actual meaning is often obscured by the use of ambiguous terms, sarcasm, lack of profanity, hateful terms, or other means. As they fall under this criterion, memes can be categorized as implicit offensive content. Hence it is difficult to classify them as offensive for human annotators as well as for machine learning approaches.

To address the issues with identifying offensive meme, we created the MultiOFF dataset by extending an existing memes dataset on the 2016 U.S. Presidential Election. Details about the data annotation process are explained in Section 4.. We address the classification task through an early fusion deep learning technique that combines the text and image modalities of a meme.

Our contributions are as follows:

I We created the MultiOFF dataset for offensive content detection, consisting of 743 memes which are annotated with an offensive or not-offensive label.

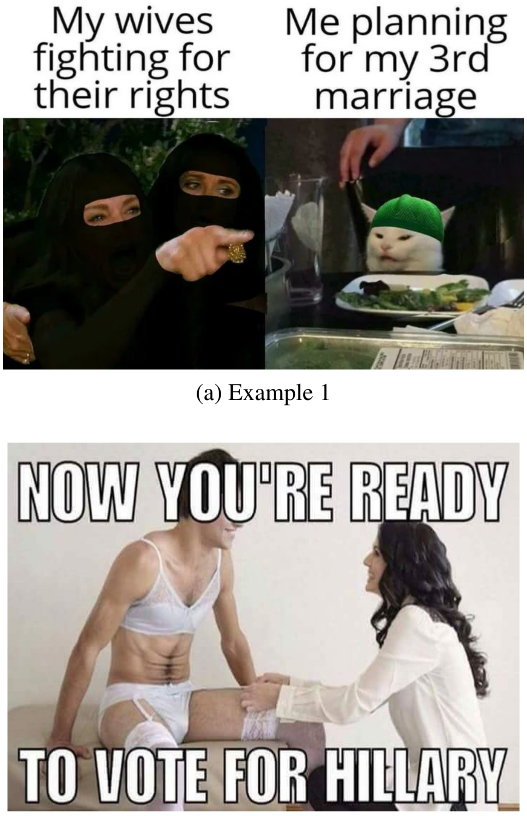
II We used this dataset to implement a multimodal offensive content classifier for memes.

III We addressed issues associated with multimodal classification and data collection for memes.

2. Offensive Content

Offensive content intends to upset or embarrasses people by being rude or insulting (Drakett et al., 2018). Past work on offensive content detection focused on hate speech detection (Schmidt and Wiegand, 2017; Ranjan et al., 2016; Jose et al., 2020), aggression detection (Aroyehun and Gelbukh, 2018), trolling (Mojica de la Vega and Ng, 2018), and cyberbullying (Arroyo-Fernández et al., 2018). In the case of images, offensive content has been studied to detect nudity (Arentz and Olstad, 2004; Kakumanu et al., 2007; Tian et al., 2018), sexually explicit content, objects used to promote violence, and racially inappropriate content (Connie et al., 2018; Gandhi et al., 2019).

[1] https://www.lexico.com/en/definition/meme


My wives fighting for their rights
Me planning for my 3rd marriage
NOW YOU'RE READY
TO VOTE FOR HILLARY

(a) Example 1

(b) Example 2

Figure 1: Examples of offensive memes from MultiOff dataset.

Due to the multitude of terms and definitions used in literature for offensive content, the SemEval 2019 task categorized offensive text as targeted, untargeted offensive text, if targeted then targeted to a group or an individual Zampieri et al. (2019). Inspired by this, we define an offensive meme as a medium that spreads an idea or emotion which intends to damage the social identity of the target person, community, or lower their prestige.

A meme can be considered as implicitly abusive since it uses a non-offensive sentence in combination with a provoking image or the other way around. The use of an unrelated text often obscures the actual meaning of a derogatory image or the other way around. The obscure nature of the meme resulted in the differences in opinion amongst the annotators, hence we provided multiple examples of offensive memes and non-offensive memes. The examples are shown in Appendix A. In the first example from Figure 1, the meme is attacking a minority as it tries to paint religion in a bad manner. This is noticeable from the visual cues from the image, i.e., attire of the characters in the image. The second example 1 is attacking Hillary (Democratic candidate in 2016 U.S. presidential election) supporters by shaming them. This meme follows similar behavior as the first example as the idea behind the meme is unknown due to obscure text. Nevertheless, the image associated with the text clears this doubt and conveys the idea. To build an automatic offensive detection system, we therefore have to have a good understanding of the textual and visual features of the meme.

3. Related work

The related section covers the work done in identifying offensive content in text and image. It also describes the research done in the area of meme analysis as well as multimodality.

3.1. Offensive Content in Text

Warner and Hirschberg (2012) model offensive language by developing a Support Vector Machine (SVM) classifier, which takes in features manually derived from the text and classifies if the given text is abusive or not. Djuric et al. (2015) have used n-gram features to classify if the speech is abusive or not. There are many text-based datasets available for aggression identification (Watanabe et al., 2018), hate speech identification (Davidson et al., 2017) and Offensive language detection (Wiegand et al., 2018; Zampieri et al., 2019). Amongst the work mentioned, Watanabe et al. (2018) relies on unigrams and pattern of the text for detecting hate speech. These patterns are carefully crafted manually and then provided to machine learning models for further classification. Wiegand et al. (2018; Zampieri et al. (2019) deals with the classification of hateful tweets in the German language and addresses some of the issues in identifying offensive content. All this research puts more weight on features of single modality i.e. text and manual feature extraction. We work on memes which have more than one modality, i.e. image and text and feature extraction is automatically done with deep learning techniques.

3.2. Offensive Content in Image

Identifying offensive content in an image based on skin detection techniques have been proposed for nudity detection (Arentz and Olstad, 2004; Kakumanu et al., 2007; Tian et al., 2018). Several works proposed convolutional neural networks (CNNs) to identify appropriate or in-appropriate images for children (Connie et al., 2018). The research done by Gandhi et al. (2019) deals with offensive images and non-compliant logos. They developed an offensive and non-compliant image detection algorithm that identifies the offensive content in the image. They have categorized images as offensive if it has nudity, sexually explicit content, objects used to promote violence or racially inappropriate content. The dataset that has been used by authors is being created by finding similar images by comparing the embeddings of the images. The classifier takes advantage of a pre-trained object detector to identify the type of an object in the image. This research heavily relies on object detection. In our research, we are relying on automatically derived features through a pre-trained CNN, which is capable of classifying memes with relatively fewer resources. Hu et al. (2007) proposed a novel framework for classifying pornographic web pages by using both image and text. The authors used a decision tree to divide Web pages into continuous text, the discrete text, and the image. According to content representations, the algorithm fuses the result from the image classifier and the text classifier to detect inappropriate content. They showed that their fusion algorithm outperforms those by individual classifiers. While this work is identifying pornographic content on the web page, it relies on skin detection. Unlike our research, the content that they are trying to identify is less obscure and rather explicit.

3.3. Offensive Content in Memes

He et al. (2016) proposed a meme extraction algorithm that automatically extracts textual features from data posted during events such as the anti-vaccination movement[2]. The process of extraction is done by identifying independent phrases and by clustering the mutation variant of each phrase associated with the meme. This work studies the convergence and peak times of memes. Drakett et al. (2018), in their research, address online harassment of marginalized groups by abusing memes, using thematic analysis of 240 sample memes. This research studies memes from a psycho-linguistic perspective.

3.4. Multimodal Datasets

TUMBLR dataset by (Hu and Flaxman, 2018) is a multimodal sentiment analysis dataset collected from Tumblr (a microblogging site). This dataset has been loosely labelled on the tags attached to the posts available on Tumblr. Their dataset relies on the tag attached to the social media posts as a label while the MultiOFF dataset used in by us is annotated manually. They emphasize more on emotion analysis, unlike our research which gives importance to the detection of offensive content. Duong et al. (2017) proposes different types of architectural designs that can be used to classify multimodal content. While their research delves into emotion classification based on multimodal data, it does not match with the objective of this research, i.e. binary classification of memes into offensive and non-offensive. Smitha et al. (2018) suggests manual extraction of features from the given meme which can be used to classify them in positive, negative and neutral classes. On one hand, sentences related which belong emotions such as sadness, anger, disgust would be classified as negative. On the other hand, the sentences which hint happiness and surprise would be categorized in positive classes and the rest of the memes are treated as neutral. Their dataset is not publicly available. While our work is the first to create a dataset for the memes to detect offensive content using voluntary annotators.

3.5. Summary

Most of the studies mentioned above focus on meme classification on a single modality. The ones that have been dealing with multimodal content rely on machine learning approaches that require handcrafted features derived from the data to classify the observations. Internet memes are in the form of images with text, this adds visual elements to the message. As multimodal approaches that are capable of classification rely on manual feature extraction, the system with automatic feature extraction can be used to provide a generic and robust solution to these difficulties. Deep neural network has the capability of deriving such features with minimal manual intervention, however, an annotated dataset for memes was not publicly available. Recently, a shared task on emotions in memes (Memotion Analysis) was published in Semeval 2020 (Sharma et al., 2020) while we were creating our dataset. The details of the data collection are not explained in the shared task. However, we are the first one to collect a multimodal offensive meme dataset using voluntary annotators.

4. MultiOFF Dataset

An event such as the 2016 U.S. Presidential Election can be used as a reference to identify offensive content on social media. The initial dataset has been accessed from Kaggle.[3] This dataset has image URLs and the text embedded in the images. The memes have been collected from social media sites, such as Reddit, Facebook, Twitter and Instagram.

4.1. Data Pre-processing

The dataset from Kaggle has many images and may unrelated features such as a timestamp (date published), link (post URL), author, network, likes or upvotes. Those that did not serve the objective of the research were removed, i.e., only the URL link and text (caption) were used from the existing dataset. The captions contained a lot of unwanted symbols such as `//n` or `@`. As this was hindering the readability of the text, all such symbols were removed from the text during the initial data pre-processing step. Furthermore, the observations in the form of long text posts were removed from the dataset and only the one with less than or equal to 20 sentences of text were kept. Each of the image URLs has been verified for its availability and the image has been obtained locally for training the classifiers for offensive content.

4.2. Data Collection and Annotation

We constructed the MultiOFF dataset by manually annotating the data into either the offensive or non-offensive category. The annotators, which used Google Forms (Chakravarthi et al., 2019; Chakravarthi et al., 2020b; Chakravarthi et al., 2020a), were given instructions to label if a given meme is offensive or non-offensive based on the image and text associated with it. The guidelines about the annotation task are as follows:

I The reviewer must review the meme as shown in Figure 6a in two categories either offensive or Non-offensive.

II Memes can be deemed offensive if it intends the following:

  (a) Personal Attack (Figure 6b)
  (b) Homophobic abuse (Figure 6c)
  (c) Racial abuse (Figure 5a)
  (d) Attack on Minority (Figure 5b)
  (e) Or Non-offensive otherwise (Figure 5c)

III Most of the memes come with an image and caption.

IV The reviewer must understand that images here are acting as context and play an important role in conveying the intention behind it. So indeed, images or text alone sometimes may not be meaningful.

[2] https://www.msdmanuals.com/professional/pediatrics/childhood-vaccination/anti-vaccination-movement

[3] https://www.kaggle.com/SIZZLE/2016electionmemes

V In case of doubt that if the meme is sarcastic, the benefit of the doubt should be given and it should be labelled as offensive.

VI While annotating the data, annotators should consider the population exposed to the content in the meme overall.

Once pre-processing and annotation guidelines were made, only six male annotators volunteered for the task. To avoid gender bias, efforts were made to balance the gender ration of the annotation task. Finally, eight annotators (six male; two female) agreed to participate in the annotation campaign.
The annotation process has been done in two steps. In the first step, a set of 50 memes has been given to each of the eight annotators. As there was no ground truth defined, the majority of the vote has been considered as the gold standard and the Fleiss' kappa (Fleiss and Cohen, 1973) has been calculated for this majority vote. Initially, the maximum and minimum value of kappa lied in the interval between 0.2 and 0.3, which showed a ”fair agreement” between the annotators. After the initial run, we asked the annotators for their feedback on the task. The issues that annotators faced while labelling the data were as follows:

I Annotators had a different interpretation of sarcastic memes. The majority of sarcastic memes had a conflict of opinion between the annotators. Example number two from Figure 1 is one such meme.

II As the annotators were unfamiliar with US politics, they were labelling the memes as offensive simply if their sentiments were hurt.

In an attempt to resolve these issues and concerns raised by the annotators, we updated the annotation guidelines and added **V** and **VI** in the given annotation guideline.
After improving the annotation guidelines, a set of 50 new memes were identified and distributed to each annotator. Similar to the first set of annotations, kappa was calculated, resulting in a ”moderate agreement” between the annotators (0.4 and 0.5).
After achieving moderate agreement, we sent all the memes to the annotators. In this phase, each meme was annotated by only one annotator. The response provided by the annotators has been taken as the final ground truth. According to psychology (Gilbert, 2006), gold standards for measuring sentiments can be a reported reaction of the audience on the content and this response can be taken as ground truth. Data annotation in itself is a challenging and emotionally draining task for the annotators as the memes in the dataset do hurt the sentiment and opinions of the annotators. Defining annotation guidelines, analyzing the annotation and overcoming the disagreement is an achievement in itself.

4.3. Dataset Statistics

After the initial data pre-processing 4.1. and data collection 4.2., the newly created dataset has 743 annotated memes. Table 1 shows a summary of the dataset used for training, validating and evaluating our work.

Data	avg#w	avg#s	off	n-off	Total
Train	41	2	187	258	445
Test	47	2	59	90	149
Val	45	2	59	90	149

Table 1: Summary statistics for the meme dataset based on the 2016 U.S. Presidential Election (avg#w: average number of words, avg#s: average number of sentences, off: offensive, and n-off: non-offensive).

Since the number of non-offensive memes is higher than that of offensive ones, we balanced this by using different class weights while training our classifier.

5. Methodology

In this section, we give insights on baselines and multimodal approaches for meme classification on our MultiOFF dataset. The subsection regarding data transformation gives insights on text and image vectorization. Baselines for text and image elaborates on the baseline models used on each modality. Finally, the multimodal approach summarises the multimodal experiments performed on the MultiOFF dataset.

5.1. Data Transformation

The text in each observation contained stopwords, non-alphanumeric symbols, words with both upper and lower cases were removed and the rest of the text has been lowercased. As a next step, the processed text has been transformed into vector sequences. Text transformation is different for each baseline. For text baseline models (Logistic Regression, Naive Bayes, Deep Neural Network), the text has been transformed into vectors according to the index and count of the word in the local vocabulary. The rest of the classifiers are using GloVe (Pennington et al., 2014) as word embeddings. Images that were locally obtained during the initial data pre-processing were converted into trainable vectors using automatic feature extraction in Convolutional Neural Network (CNN) trained on the ImageNet dataset (Deng et al., 2009).

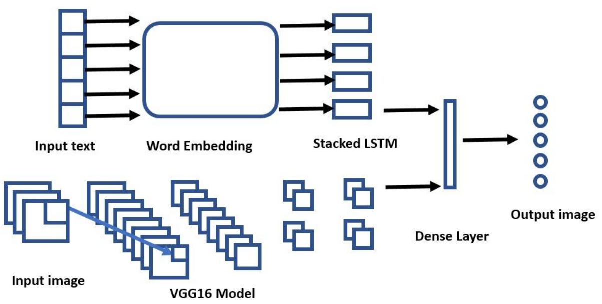


Figure 2: Early fusion model for combining visual and textual data associated with the meme.

5.2. Baseline Models for Textual Data

Logistic regression (LR) and *Naive Bayes* (NB), have been used to classify memes based on the provided textual information for a single modality experiment. The standard bag-of-word approach has been followed. Apart from these machine learning algorithms, a neural network with four layers, a stacked Long Short Term Memory (LSTM) network (Gers, 1999), a Bidirectional LSTM and a CNN have been compared for meme classification based on text.

Logistic regression (LR) used for classification is helpful if the targeted classes in the data are linearly separable (Hosmer Jr et al., 2013). This bag of word approach has been used for creating a text vector x_i. LR works with the basic assumption that the class of the observation and features are in a linear relationship with each other. The probability of the class p is being predicted for the text data which has been classified either *offensive* (p) or *non-offensive* $(1-p)$.

$$\ell = \log_b \frac{p}{1-p} = \beta_i x_i$$

Where, ℓ is the log-odds, b is the base of the logarithm, and β_i are parameters of the model. If this probability is beyond the threshold then the observation has been set as *offensive*, *non-offensive* otherwise.

Naive Bayes (NB) builds the hypothesis with the assumption that each feature is independent of each other features (McCallum et al., 1998). Eventually, NB calculates the probability of the classes given the text vector. In NB, probabilities of class *offensive* and *non-offensive* class given the text vector have been calculated. Training examples have been labeled as per the conditional probability of the class.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Where, C_k is class label, x is a feature vector, $p(C_k)$ is prior probability, $p(x|C_k)$ is likelihood, $p(x)$ is probability of feature or evidence.

A Deep Neural Network (DNN) has been used as the third baseline. A neural network with four layers has been designed to classify the meme based on text. The embedding layer that has been used in this baseline is made from the training vocabulary. The neural network has been trained solely on the training data from scratch and no transfer learning approaches have been used in this baseline. A text vector representation is a count of the word sequence from the vocabulary which is using a local word embedding to represent each word. The embedding layer takes in the input of 100 dimensions and provides embeddings of 50 dimensions. A flatten layer precedes a fully connected layer to ensure that all the embeddings get flatten before sent to fully connected layers. The output of this neural network is "sigmoid" to calculate the probability of the class. Binary cross-entropy loss function and gradient descent are used to tune all the hyperparameters associated with the hidden layer.

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

Bidirectional LSTM (BiLSTM) uses GloVe for word embeddings. Unlike LSTM, BiLSTM saves the past as well as future data sequences to preserve the context of the targeted word. In this architecture, only one BiLSTM has been used. The output of this layer has been connected to the classification layer with a sigmoid activation function which gives out the probability of the offensive class.

CNN approaches are suitable for text as it can be represented in a learnable vector form. As text can be represented in such form, CNN can be relied upon to classify the text data as well. In this baseline, two basic building blocks of CNN are used. The convolutional layer and maxpooling layer with the output of previously connected to the input of the later has been used. Three such convolutional blocks have been used before the classification layer. The flattening layer before the classification layer converts the vector in one dimension for the fully connected dense layer. Finally, the output of this layer has been cascaded to the final layer with sigmoid as a primary choice for activation function.

5.3. Baseline Model for Images

A CNN architecture developed by the Visual Geometry Group (VGG) at the University of Oxford has been used to classify the targeted image data (Simonyan and Zisserman, 2014). This specific architecture has 16 layers and is known as VGG16. The model is pre-trained on the ImageNet dataset and has been used as the baseline in our experiments. Images were loaded into an array and changed into a fixed shape as per VGG16 specifications. All the values in the matrix were in the range between 0 and 255. VGG architecture has two convolution layers both with Relu as an activation function. The output of the activation function has been fed to the max-pooling layer which later has been followed by a fully connected layer which also uses "Relu" (Wang, 2017) as an activation function. Instead of a fully connected layer, a Global Average Pooling layer has been used which later is connected to a Dense layer with the *Sigmoid* activation function to predict class probability.

In **Network Surgery**, all the 16 layers in VGG16 have been frozen by converting all the parameters in the layers as untrainable. This has been done to prevent the pre-trained network again on new data. The top layer in the model i.e. 1000 classes of ImageNet is not required and hence removed.

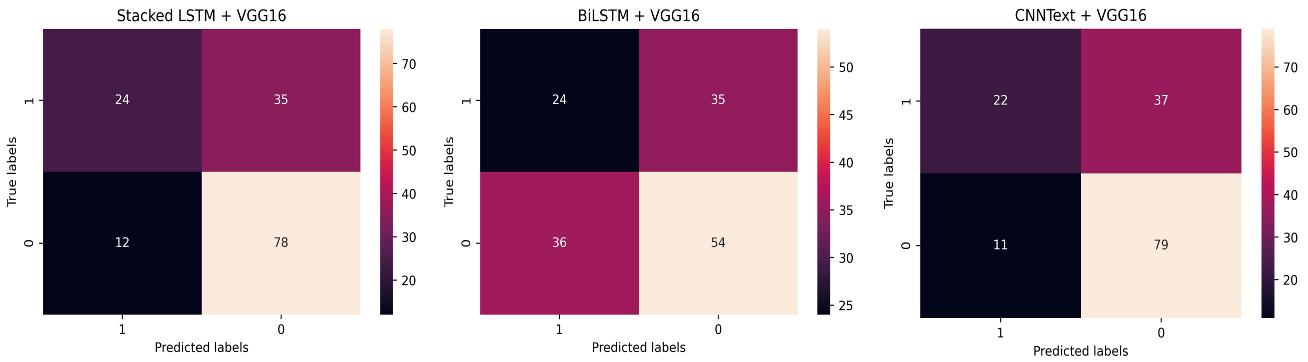
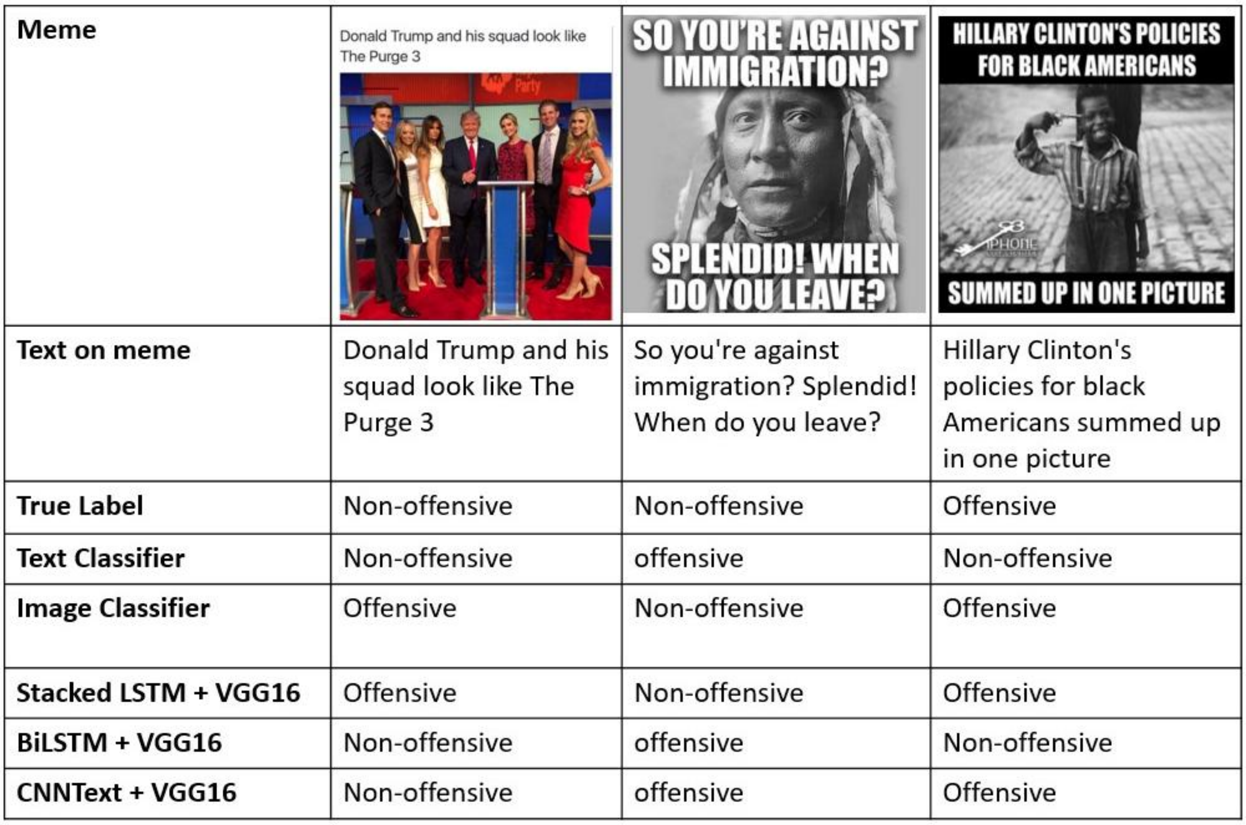


Figure 3: Confusion matrix for Multimodal classifier with Stacked LSTM, BiLSTM and CNN.


Donald Trump and his squad look like The Purge 3
SO YOU'RE AGAINST IMMIGRATION?
SPLENDID! WHEN DO YOU LEAVE?
HILLARY CLINTON'S POLICIES FOR BLACK AMERICANS
SUMMED UP IN ONE PICTURE

Meme			
Text on meme	Donald Trump and his squad look like The Purge 3	So you're against immigration? Splendid! When do you leave?	Hillary Clinton's policies for black Americans summed up in one picture
True Label	Non-offensive	Non-offensive	Offensive
Text Classifier	Non-offensive	offensive	Non-offensive
Image Classifier	Offensive	Non-offensive	Offensive
Stacked LSTM + VGG16	Offensive	Non-offensive	Offensive
BiLSTM + VGG16	Non-offensive	offensive	Non-offensive
CNNText + VGG16	Non-offensive	offensive	Offensive

Figure 4: Predictions from the Stacked LSTM + VGG16 classifier.

5.4. Multimodal Approach

To support our research hypothesis, the text and image classifiers are evaluated individually. Additionally, we combined the modalities (text and image), which is known as the ”Early Fusion Approach” (Duong et al., 2017).

As shown in Figure 2 (Hu and Flaxman, 2018), the text and image modalities in their vector form have been fed into the classifier. In this architecture, both modalities are required to classify the offensive content. A new vector has been formed by the concatenation of both modalities which represents a meme as a whole and hence can be used for classification.

The setup for each of the experiment remains the same in the case of training. As the amount of data is insufficient to train a DNN, we take advantage of pre-trained embeddings. On the one hand, pre-trained VGG16 on the ImageNet dataset has been used for images, while GloVe has been used to represent word embeddings.

Stacked LSTM + VGG16: VGG16 has been used to extract image features. It is a CNN model, pre-trained on the ImageNet dataset. The same Stacked LSTM approach used in the text baseline has been used in the multimodal experiment.

BiLSTM + VGG16: In this experiment, Bi-directional LSTM has been used to vectorise the text, which was combined with the image features. This combination gives rich information about the training example and stands a better chance of getting classified in the correct category.

CNNText + VGG16: In this experimental setting, image features have been carried out by a pre-trained VGG16 net-

Type	Classifier	P	R	F
Text	LR	0.58	0.40	0.48
	NB	0.52	0.45	0.49
	DNN	0.47	0.54	0.50
	Stacked LSTM	0.39	0.42	0.40
	BiLSTM	0.42	0.23	0.30
	CNN	0.39	0.84	0.54
Image	VGG16	0.41	0.16	0.24
Multi	Stacked LSTM + VGG16	0.40	0.66	0.50
	BiLSTM + VGG16	0.40	0.44	0.41
	CNNText + VGG16	0.38	0.67	0.48

Table 2: Precision, recall and F1-score for the baseline and multimodal classifiers.

work on the ImageNet dataset and textual features have been extracted by using a CNN model. These features are concatenated and fed as input to a stacked LSTM model. The output of the LSTM model is connected to the dense layer which then is combined with the image features to represent the meme. The CNNText+ VGG16 approach leverages the CNN architecture text as used in the baselines above.

6. Results and Discussion

The set of 743 memes has been randomly split into train, validation and test dataset. Table 1 shows the data statistics. All approaches mentioned in the previous section are applied to the text, extracted from the memes, whereby ***early fusion approaches*** have been used to implement a DNN to combine the two targeted modalities. The Table 2 shows the results of the meme classification experiments. Later on, these baselines, except LR, NB, and DNN, have been extended to build the multimodal classifier that can classify the meme based on textual and visual features of the meme. From Table 2, it is evident that **Logistic regression** performs best in predicting the offensive meme category based on the text. Classification of offensive language with the **CNN on text** provides the highest recall, which highlights its capability of retrieving the offensive meme. On the other hand, the precision of 0.39 shows that many memes are being mislabeled as offensive. **VGG16** generates the lowest recall, which shows that only 0.16 of memes were retrieved from the total pool of offensive memes. According to the same table, DNN on text has a 0.5 F1-score, but it showed an inferior recall value of 0.55 when compared to the recall of Stacked LSTM + VGG16 (0.66). As mentioned earlier, DNN is the only model with local embeddings. Hence it is showing better precision, recall, F1-score than other models. It is showing better results for memes related to this domain but may as well fail in generalising.

It can be seen from the Table 2 that the text classifier based on the Stacked LSTM, BiLSTM and CNN text show improvements in terms of recall when text and image features are considered. The last three entries in Table 2 report the evaluation results for the multimodal classifier. On average, the precision of 0.40 is achieved for all three multimodal approaches. This has been achieved without suffering from a poor recall, as recall for all of them is in a range between 0.44 and 0.67. As a result, a balanced F1-score has been achieved which maintains the inclination of getting more precision without reducing recall. Figure 3 shows an interesting fact about the multimodal classifiers. All the classifiers end up identifying the same number of offensive memes, while the recall of each distinguishes them from each other. An ensemble model could be built by leveraging the strength of multiple classifier to identify the offensive content. Figure 4 shows the predictions of stacked LSTM text classifier, VGG16 image classifier, and their combined multimodal classifiers. In the first example, the true label for the meme is non-offensive, whereby the text classifier predicts it correctly. Differently, the image classifier predicts the same meme as offensive, while the BiLSTM + VGG16 and CNNText + VGG16 classifier correctly labels it as Non-offensive. In the third meme, we can see an offensive content in terms of a child holding a gun to his head. This image in itself can be deemed as offensive but the text associated with it is vaguely Non-offensive if considered alone. The text classifier fails to identify the true label. On the other hand, the image classifier identifies the right label followed by the multimodal classifier.

7. Conclusions and Future Work

In this work, we implementer an approach on offensive content classification in memes based on images and text associated with it. For this purpose we enriched an existing memes dataset with offensive or non-offensive labels with the help of voluntary annotators. This MultiOFF dataset was the used to train and evaluate a multimodal classification system for detecting offensive memes. Results demonstrate the improvement in retaining offensive content (recall) when both text and image modality associated with the meme was considered.

Although results in Table 2 show that the ability to retain most of the offensive content will be increased by a multimodal classifier, it is still debatable if the accuracy of such a multimodal approach is reliable. As a remedy, manual evaluation by an administrator should be imcluded before blocking offensive content. The result shown by the text classifier shows accuracy close to the multimodal classifier and sometimes better. While the image classifier has a lesser chance of identifying and retaining offensive memes on its own, the multimodal classifier shows improvements in retaining offensive memes. This suggests that there are more chances of improving accuracy by increasing the weight of textual features while combining it with visual elements of the meme. The future direction of this research focuses on the usage of tags associated with social media posts which are treated as the label of the post while collecting the data. This will help us to gather more training data. For this work, we used the 2016 Presidential Election Memes dataset, but to avoid the biases caused due to use of the specific domain, a variety of memes can be included from different domains. The approach of combining modalities can be extended for other multimedia content such as audio and video. Concatenating the image and text embeddings for representing memes could be improved upon by fusing embeddings. As it is hard to explain the abstract features that are responsible for identifying offensive content, the inclusion of more training data will help us to understand it. For automatic evaluation of a meme, we need text as the different modality. This text is often embedded on the meme. Hence to capture the embedded text, we can use OCR techniques.

Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund, as well as by the H2020 project Prêt-à-LLOD under Grant Agreement number 825182.

References

Arentz, W. A. and Olstad, B. (2004). Classifying offensive sites based on image content. *Comput. Vis. Image Underst.*, 94(1-3):295–310, April.

Aroyehun, S. T. and Gelbukh, A. (2018). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Arroyo-Fernández, I., Forest, D., Torres-Moreno, J.-M., Carrasco-Ruiz, M., Legeleux, T., and Joannette, K. (2018). Cyberbullying detection task: the ebsi-lia-unam system (elu) at coling'18 trac-1. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 140–149, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Chakravarthi, B. R., Priyadharshini, R., Stearns, B., Jayapal, A., S, S., Arcan, M., Zarrouk, M., and McCrae, J. P. (2019). Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland, August. European Association for Machine Translation.

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020a). A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., and McCrae, J. P. (2020b). Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Connie, T., Al-Shabi, M., and Goh, M. (2018). Smart content recognition from images using a mixture of convolutional neural networks. In Kuinam J. Kim, et al., editors, *IT Convergence and Security 2017*, pages 11–18, Singapore. Springer Singapore.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, Miami, Florida, USA. Ieee.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, New York, NY, USA. ACM.

Drakett, J., Rickett, B., Day, K., and Milnes, K. (2018). Old jokes, new media–online sexism and constructions of gender in internet memes. *Feminism & Psychology*, 28(1):109–127.

Duong, C. T., Lebret, R., and Aberer, K. (2017). Multimodal classification for analysing social media. *ArXiv*, abs/1708.02099.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619, 10.

French, J. H. (2017). Image-based memes as sentiment predictors. In *2017 International Conference on Information Society (i-Society)*, pages 80–85, Dublin, Ireland. IEEE.

Gandhi, S., Kokkula, S., Chaudhuri, A., Magnani, A., Stanley, T., Ahmadi, B., Kandaswamy, V., Ovenc, O., and Mannor, S. (2019). Image matters: Detecting offensive and non-compliant content/logo in product images. *arXiv preprint arXiv:1905.02234*.

Gers, F. (1999). Learning to forget: continual prediction with lstm. *9th International Conference on Artificial Neural Networks: ICANN '99 (Edinburgh, UK)*.

Gilbert, D. T. (2006). *Stumbling on Happiness*. Alfred A. Knopf, New York, New York, United States.

He, S., Zheng, X., Wang, J., Chang, Z., Luo, Y., and Zeng, D. (2016). Meme extraction and tracing in crisis events. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 61–66, Tucson, AZ, USA. IEEE.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hu, A. and Flaxman, S. (2018). Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 350–358, London, UK. ACM.

Hu, W., Wu, O., Chen, Z., Fu, Z., and Maybank, S. (2007). Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1019–1034, June.

Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)*.

Kakumanu, P., Makrogiannis, S., and Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern recognition*, 40(3):1106–1122.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Mojica de la Vega, L. G. and Ng, V. (2018). Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., and McCrae, J. P. (2020). A comparative study of different state-of-the-art hate speech detection methods for Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, May. European Language Resources Association (ELRA).

Ranjan, P., Raja, B., Priyadharshini, R., and Balabantaray, R. C. (2016). A comparative study on code-mixed data of Indian social media vs formal text. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 608–611, Dec.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Sharma, C., Paka, Scott, W., Bhageria, D., Das, A., Poria, S., Chakraborty, T., and Gambäck, B. (2020). Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smitha, E., Sendhilkumar, S., and Mahalaksmi, G. (2018). Meme classification using textual and visual features. In *Computational Vision and Bio Inspired Computing*, pages 1015–1031. Springer.

Suryawanshi, S., Chakravarthi, B. R., Verma, P., Arcan, M., McCrae, J. P., and Buitelaar, P. (2020). A dataset for troll classification of Tamil memes. In *Proceedings of the 5th Workshop on Indian Language Data Resource and Evaluation (WILDRE-5)*, Marseille, France, May. European Language Resources Association (ELRA).

Tian, C., Zhang, X., Wei, W., and Gao, X. (2018). Color pornographic image detection based on color-saliency preserved mixture deformable part model. *Multimedia Tools and Applications*, 77(6):6629–6645, Mar.

Wang, Z. (2017). Temporal-related convolutional-restricted-boltzmann-machine capable of learning relational order via reinforcement learning procedure. *International Journal of Machine Learning and Computing*, 7:1–8, 02.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

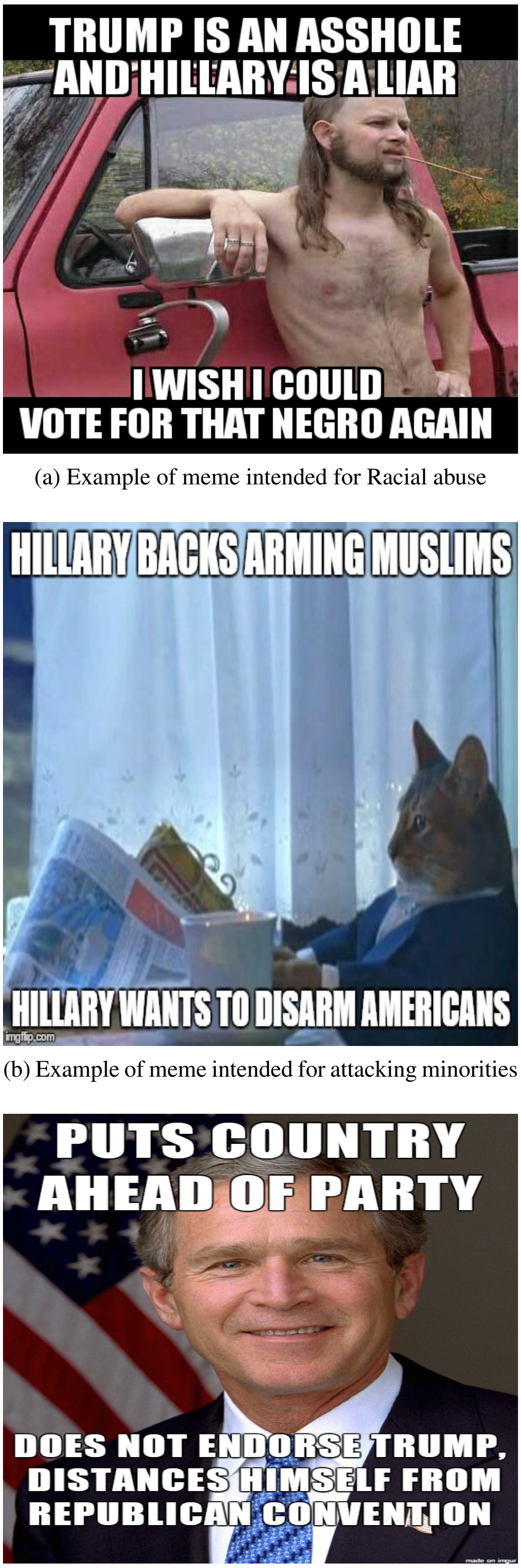
Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria.

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). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

A Appendix: Examples from Annotation Guidelines

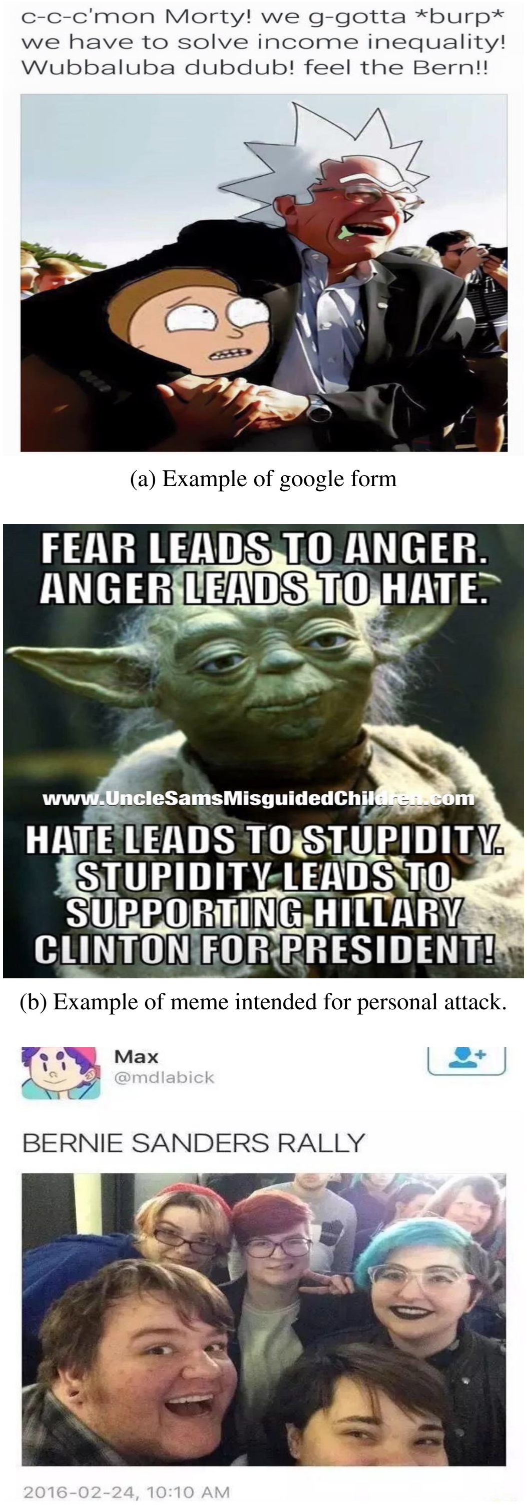


(a) Example of meme intended for Racial abuse

(b) Example of meme intended for attacking minorities

(c) Example of non-offensive meme

Figure 5: Example images



(a) Example of google form

(b) Example of meme intended for personal attack.

(c) Example of meme intended for Homophobic abuse

Figure 6: Example images