JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IDENTIFICATON OF OFFENSIVE MEME CONTENT USING CNN

Mrs.A.Barveen[1], S.Abirami[2], L.Kris Dayana[3], R.Meena[4]

M.E., P.h.D., Assistant Proffessor[1], BE[2], BE[3], BE[4]
Department Of Computer Science and Engineering
M.I.E.T.Engineering College,Trichy,India.

Abstract- People can express themselves more easily through the creation or exchange of information, ideas, or other types of content on social media, an interactive platform for communication. A image can best describe a meme. Memes make it even harder because they imply humour and sarcasm, so if we only look at the image, the meme might not seem offensive. Memes can be used to make personal attacks, homophobic, racist, or minority attacks, among other things. Memes should be the main emphasis as the use of these media platforms grows, and automatic multimodal content detection should be necessary. However, a (CNN) deep learning strategy can help to partly solve this issue.

Index Terms— ***Meme dataset, Offensive content, Social Media, Convolutional Neural Network***

I. Introduction

Using imitation or other non-genetic behaviors, a meme is "an element of a culture or system of behavior passed from one individual to another. "Images, videos and Twitter posts are just a few examples of the many different kinds and formats of memes, which are having an increasingly big impact on social media communication. (French, 2017; Suryawanshi et al., 2020). Memes as pictures are the most widely used content format. It is frequently challenging to comprehend the content from a single modality because memes are multimodal. (He et al., 2016). Therefore, it's crucial to take into account both modes in order to comprehend the meme's meaning or intended meaning. Sadly, memes are to blame for the spread of hatred in society, which is why it's necessary to instantly detect memes with offensive content. However, because of its frequently occurring memes, the picture is challenging for automatic filtering to control. Or, to put it another way, memes are multimodal in nature, making it unpredictable whether the offensive content will be linked to the background picture. They cannot be labelled as objectionable based solely on the findings of one modality. When we look at the picture alone, we might not be able to see anything dangerous, but when we look at both modalities, the context changes. In order to more accurately categories memes as offensive or not, this project seeks to build a multimodal system that can process images on the meme in parallel. The classification of the meme as offensive or not is then determined using these image characteristics [21] and the text features. We used an Instagram-based social media website to categories and create image memes. The actual meaning of implicitly offensive or abusive content is frequently hidden by the use of ambiguous language, sarcasm, an absence of profanity, or other techniques. Memes are able to be categories as implicitly offensive material because they meet this criterion. Machine learning is an idea in AI where a machine can train itself to perform a job. Image classification can be done using Convolution Neural Networks (CNN).

1.1Offensive Content

By being rude or insulting, offensive material seeks to upset or humiliate others. (Drakett et al., 2018). Previous research on objectionable content detection concentrated on trolling (Mojica de la Vega and Ng, 2018), aggression detection (Aroyehun and Gelbukh, 2017), hate speech detection (Schmidt and Wiegand, 2017; Ranjan et al., 2016; Jose et al., 2020), and cyberbullying. (Arroyo-Fernandez et al., 2018). Arentz and Olstad (2004), Kakumanu et al. (2007), Tian et al. (2018), Gandhi et al. 2019), Connie et al. (2018), and Gandhi et al. (2019) conducted studies to identify offensive content in images, including nudity, sexually explicit material, items that encourage violence, and racially insensitive material.

(a) Offensive image 1 (b) Offensive image 2

Figure 1. Offensive Images

Examples of offensive images from the dataset are shown in Figure 1.

Due to the range of terminologies and descriptions used in the literature for such content, Zampieri et al. (2019) categorized offensive text as targeted, untargeted, and if targeted, to a group or a person in the Sem Eval 2019 task. As a result, we describe an offensive meme as a medium that spreads an idea or emotion with the aim of undermining the social identity of the target person, group, or community. A meme can be interpreted as implicitly offensive since it blends a neutral text with a provocative image, or the other way around.. The actual meaning of a derogatory picture is frequently obscured by the use of unrelated text, or vice versa. We gave numerous examples of both offensive and non
offensive memes because the obscure nature of the meme caused disagreements among the annotators. Examples of these are provided in Appendix A. The meme is criticizing a minority in the first illustration from Figure 1 b) and attempting to portray religion in a negative light. This is apparent from the visual cues in the picture, specifically the clothing worn by the characters. The first figure 1 is attacking Hillary (Democratic candidate in 2016 U.S. presidential election) supporters by shaming them.

Due to its cryptic text and comparable behavior to the first example, this meme also lacks clarity regarding its underlying concept. However, the picture that goes with the text dispels this uncertainty and communicates the concept. The visual characteristics of the meme must therefore be well understood in order to develop an automated offensive detection system.

II. Related Work

The work done to identify objectionable content in images is covered in the related part. It also discusses the multimodality and meme analysis study that has been done. The CNN based approach has better performance in terms of recall when image features are considered

For the purpose of locating objectionable information in a picture, nudity recognition based on CNN approaches has been proposed (Arentz and Olstad, 2004; Kakumanu et al., 2007; Tian et al., 2018). Convolutional neural networks (CNNs) have been proposed as a method to categorize images for children as appropriate or inappropriate in a number of publications. 2018 (Connie et al.). A research on offensive visuals and non-compliant logos was undertaken by Gandhi et al. in 2019. They developed an algorithm to recognize offensive content in non compliant and objectionable photos. A photograph is deemed offensive if it contains nudity, sexually explicit content, weapons or other symbols of violence, or if it is racially insulting.

By comparing the embeddings of the images, the authors were able to identify similar images and construct the dataset that they used. To determine the kind of item in the image, the classifier uses a previously trained object detector. Object detection is heavily used in this study. In our study, we depend on automatically generated features through a CNN that has already been trained to classify memes with comparatively fewer resources. A novel framework for categorizing pornographic web sites using images was put forth by Hu et al. in 2007. To categorize Websites into discrete images, the authors used a decision tree. The algorithm fuses the output from the image classifier to identify inappropriate content in accordance with content representations. This work depends on CNN to find pornographic material on the webpage. Unlike our study, their attempt to identify the content is less cryptic and more explicit. He and colleagues (2016) suggested a meme extraction algorithm that uses data posted during occasions like the anti-vaccination movement to automatically extract textual features2

The extraction process is carried out by identifying isolated sentences and combining the mutation variation of each phrase associated with the meme. This research examines the peaks and points of confluence of memes. Drakett et al. (2018) used thematic analysis of 240 example memes to address the issue of online harassment of marginalized groups through the use of memes. This study examines memes from a psycho linguistic angle.

III. Methodology

To identify offensive content in images, we use the meme dataset as shown in Table 1. There are three columns in each sort of data file. The first column contains the name of the picture along with its file extension.

Table. 1. Summary of dataset

Data	Number of offensive data	Number of non offensive data	Total
Training	187	258	445
Validation	58	91	149
Test	58	91	149

The proposed System architecture is shown in Figure 2. Getting Information and Annotating it by manually classifying the data into offensive and non-offensive categories, we created the dataset.

Figure 2: System Architecture

The annotators were asked to use Google Forms to categorize a specific meme as offensive or not offensive based on the accompanying picture. The following are the requirements for the labelling job.

3.1 Data Collection and Annotation

Obtaining and Annotating Information We produced the dataset by manually categorizing the data into contentious and non-offensive groups. The reviewer must classify the shown image either as offensive or inoffensive depending on the image that supported it. The annotators were tasked to identify a specific image as being offensive or not offensive depending on the image that supported it.. Memes that target: (a) personal assault(b) homophobia may be deemed offensive. racial epithets(c)Attack on a minority group(d) non-offensive in any other situation. Most memes consist of an image and a caption. The reviewer needs to be conscious that the context and importance of the images in conveying their meaning are important points to note. Consequently, a picture might rarely have no meaning. The meme should be labeled offensive and granted the benefit of the doubt if there is any doubt as to whether it is being taken seriously. When annotating the data, Annotators should consider the population's total exposure to the meme's content.

Only six male annotators provided their assistance once pre-processing and annotation guidelines were established. The gender distribution of the annotation distribution was balanced in order to avoid gender bias. Eight annotators (six men and two women) eventually consented to take part in the annotation effort. Two steps were taken to finish the annotating process. Every one of the eight annotators received 50 memes in the first round. The vast majority vote has been used as the gold standard although there was no support for truth provided, and the Fleiss' kappa was determined for this majority vote. Initially, "fair agreement" between the annotators was indicated by the spread of kappa's highest and lowest values being between 0.2 and 0.3. The challenges that annotators faced when recording the data included the following: varying annotators had varying interpretations of satirical ideas. Most sarcastic memes were annotated by people who differed from each other. illustration #2 from Figure 1 is one illustration of such a picture. The annotators were merely classifying the images as offensive if they had been offended by them due to their lack of knowledge of US politics.

We updated the annotation standards and added numbers V and VI to lists that were previously received in an effort to solve the difficulties brought up by the annotators. Each annotator received 50 brand-new memes when the annotation criteria were updated. When Kappa was determined, it revealed average agreement between the annotators, similar to the previous set of annotations. (0.4 and 0.5). We forwarded every joke to the annotators after we got a good lot of consensus, A reported audience response to the content can serve as the gold standard for gauging sentiments, according to psychology (Gilbert, 2006), and this response may be interpreted as the actual joke. Since the dataset's memes truly undermine the sentiment analysis results, data annotation is a difficult and emotionally taxing task for the annotators.

3.2 Data Pre-processing

In addition to a large number of images, the Kaggle dataset also includes potentially unrelated data such as a timestamp (date of publishing), link (post URL), creator, network, likes, or upvotes. The present dataset was used exclusively, and any URL links (captions) that did not advance the study's objectives were removed. Unwelcome symbols like //n and @ were all over the remarks. During the initial data pre processing step, all of these symbols were removed from the text because they made it challenging to read. Additionally, the availability of each picture URL has been verified, and the image has been downloaded locally in order to train the classifiers for objectionable content.

3.3 Statistics for a Dataset

There are 743 annotated memes in the freshly created dataset after early data collection (3.1) and pre-processing (3.2). The statistics that were used for the instruction, validation, and evaluation of our work are summarized in Table 1.

Table 1 lists the average word count, average sentence count, offensive and neutral terms, as well as summary figures for the meme dataset based on the 2016 U.S. Presidential Election. In order to take into consideration the fact that there are more offensive memes than neutral ones, we trained our classifier with a variety of class weights.

We discuss the criteria for categorizing parodies in our dataset in this part. The baseline models used for each modality are covered in more depth in the baselines for images. The attempts in the dataset are summarized in the end.

3.4. Baseline Model for Images

A CNN architecture was created by the group known as the Visual Geometry (VGG) at the educational institution of Oxford and used to classify the selected image data. 2014's Simoyan and Zisserman. The model served as the basis for our testing because it was already trained employing the ImageNet data set. The matrix's values were all between 0 and 255. Each of the two convolution layers in the VGG architecture has the activation function Relu. The max-pooling layer received the result of the activation function, and after that came a totally linked layer that also employs "Relu" (Wang, 2017) as a function for activation. To forecast class probability, a Globally Pooling Average layer that is linked to a Dense surface with the activation function of the Sigmoid has been employed rather than a fully connected layer. To avoid applying the already trained network to fresh data, this was done. The highest layer of the model, which had 1000 ImageNet classes, was eliminated since it was unnecessary.

He and coworkers (2016) proposed a meme extraction algorithm that automatically extracts textual features from data uploaded during events like the anti-vaccination movement.2

3.5. Model Architecture

Collect a dataset of labelled images: To train a deep CNN to recognize memes, you will need a large dataset of labelled images. You could use a tool like Google Images or Instagram's API to scrape images and manually label them as "meme" or "not meme". Split the dataset into training and validation sets: Once you have your dataset, split it into training and validation sets. The training set will be used to train the deep CNN, while the validation set will be used to test its performance and fine-tune the model's hyperparameters.

Build the deep CNN architecture: There are many deep CNN architectures you could use to recognize memes, such as ResNet, Inception, or VGG. You could either build the architecture from scratch or fine-tune a pre-trained model on a similar task. Train the deep CNN: With the deep CNN architecture defined, you can train it on the labeled dataset. During training, the model will learn to recognize patterns and features that distinguish memes from non-memes. Evaluate the model: Once the model is trained, evaluate its performance on the validation set. This will give you an idea of how well it's able to recognize memes and where it may be making mistakes. Fine-tune the model.

Based on the model's performance on the validation set, you may need to fine-tune its hyperparameters or adjust the model architecture to improve its accuracy. Deploy the model: Once you are satisfied with the model's performance, you can deploy it to recognize memes in real world images on Instagram.

This could involve building an API or integrating the model into an existing Instagram scraper. Keep in mind that detecting memes in Instagram using a CNN is a complex task that requires a significant amount of data and computational resources. Additionally, there may be legal and ethical considerations when collecting and using Instagram data.

Preprocess the images: Before feeding the images into the CNN, you will need to preprocess them. This could include resizing the images to a standard size, normalizing the pixel values, and data augmentation techniques such as flipping, rotating, or shifting the images. The outputs of the models were then aggregated using the fine tune model to generate the final output. The Results achieved is shown in Table 2 and Figure 3.

Table 2. Results achieved

Meme			
True Label	Non-Offensive	Non-Offensive	Offensive
Image Classifier	Offensive	Non-Offensive	Offensive

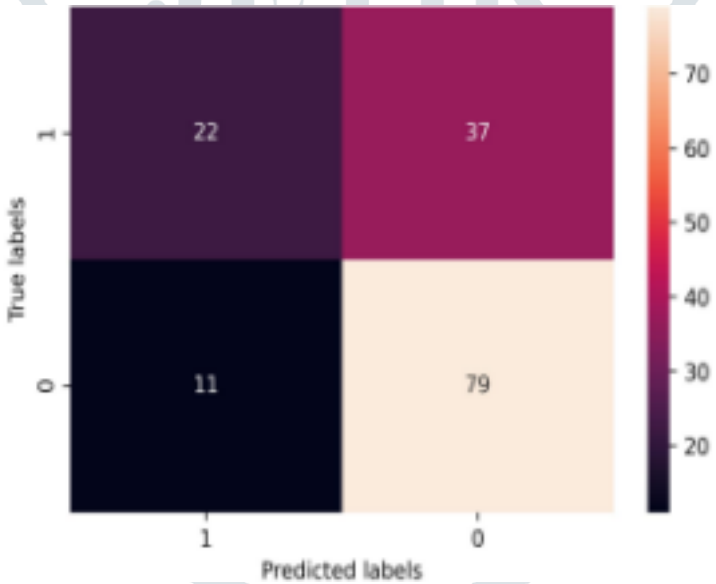

Figure 3: CNN

IV. Conclusions and Future Work

In this study, we put into practice a method for classifying offensive memes based on the images that go along with them. Results show that taking into account the meme's associated picture modalities improved recall of offensive material. Text Classifier performs significantly better at keeping offensive memes than the image classifier, which has a lower likelihood of doing so on its own. This meme indicates that increasing the weight of aspects of the system has a greater chance of improving accuracy. This suggests that when visual components are combined with textual meme traits, there is greater potential to boost accuracy by raising the weight of the visual components. The future path of this study is the usage of tags associated with social media postings, which are utilized as the post's title when data is gathered. As a consequence, we will be capable to gather additional training data. A variety of memes from various domains can be incorporated to avoid the biases created by the usage of the specific domain, even though we utilized the 2016 presidential vote Memes datasets for this study. More training data will aid in our understanding of the abstract features that determine offensive content because they are difficult to describe.

V. References

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

[2] Aroyehun, S. T. and Gelbukh, A. (2018). By using deep neural networks, data augmentation, and pseudo labelling for "aggression detection in social media." Pages 90–97 are included in the proceedings of the first workshop on trolling, aggression, and cyberbullying (TRAC-2018), which was held in August in Santa Fe, New Mexico, USA. Association for Computational Linguistic

[3] Connie, T., Al-Shabi, M., and Goh, M. (2018)."Using a combination of convolutional neural networks, smart content recognition from images."In Kuinam J. Kim, et al., editors, IT Convergence and Security 2017, pages 11–18, Singapore. Springer Singapore.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei (2009).Imagenet: “A large-scale hierarchical image database”.Miami, Florida, USA: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255. Ieee

[5] 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

[6] Mojica de la Vega, L. G. and Ng, V. (2018). “Modeling trolling in social media conversations”.In the papers presented at the eleventh international conference on language resources and evaluation (LREC 2018), which took place in May in Miyazaki, Japan. European Language Resources Association (ELRA).

[7] Simonyan, K. and Zisserman, A. (2014). "Using incredibly deep convolutional networks for large-scale picture recognition". arXiv preprint arXiv:1409.1556.

[8]Zampieri, Malmasi, Nakov, Rosenthal, Farra, and Kumar. (2019). "Identifying and categorising offensive language in social media" is SemEval-2019 job 6 (OffensEval). Proceedings of the Thirteenth International Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA, June, pages 75 to 86.Association of Computational Linguistics.