NIT-Agartala-NLP-Team at SemEval-2020 Task 8: Building Multimodal Classifiers to tackle Internet Humor

Steve Durairaj Swamy Shubham Laddha Basil Abdussalam

Debayan Datta Anupam Jamatia
Department of Computer Science and Engineering
National Institute of Technology
Agartala, Tripura, India
{steve050798, laddha.shubham97, basilwkrh, debayan.datta98, anupamjamatia}@gmail.com

Abstract

The paper describes the systems submitted to SemEval-2020 Task 8: Memotion by the 'NIT-Agartala-NLP-Team'. A dataset of 8879 memes was made available by the task organizers to train and test our models. Our systems include a Logistic Regression baseline, a BiLSTM + Attention-based learner and a transfer learning approach with BERT. For the three sub-tasks A, B and C, we attained ranks 24/33, 11/29 and 15/26, respectively. We highlight our difficulties in harnessing image information as well as some techniques and handcrafted features we employ to overcome these issues. We also discuss various modelling issues and theorize possible solutions and reasons as to why these problems persist.

1 Introduction

Over the years, the internet and social media have become an indispensable part of our lives. Today, an average netizen spends over 45 minutes on some form of social media everyday[1]. Therefore, social media has become a goldmine for data to model and study human opinions and behaviour. Social media analysis conventionally deals with data in the form of text, audio or video — but often only one prominent modality is studied in isolation. This lack of multimodality in research works has led to entire modes of communication being disregarded. One such form of internet communication that we have yet to tap into is memes. While a meme was initially defined as *"an idea, behaviour, or style that spreads from person to person within a culture—often to convey a particular phenomenon, theme, or meaning*[2]*,"* it has transformed into the umbrella term for a suite of referential humor that plagues the internet. Memes can more aptly be defined as *"a form of referential humor that incites humor by leveraging images, text and sometimes audio."* The image is more often than not coupled with a real-life event or media. The images are then re-purposed to incite humor through the subversion of expectations and via reference. Through the years, memes have been leveraged differently by various communities — apart from innocent, humor inducing purposes. Corporations and companies are increasingly interested in harnessing memes to sway their younger customers. More adversely, certain memes such as *Pepe the Frog*[3] were adopted by alt-right groups and used as a calling card of sorts. Memes are also used as performative acts, which involve a conscious decision to support or oppose a movement (as seen in the case of the recent Hong Kong protests [4]). The abundance of memes on social media platforms such as Facebook, Instagram, and Twitter (Zannettou et al., 2018b) further suggests that — these digital constructs have become a flagship of internet culture, so and so that, to understand memes would mean to understand the views of a community. The SemEval Task 8: Memotion (Sharma et al., 2020) shared task is one such grassroots endeavour to understand memes better and incorporate multimodality in social media analysis. Apart from the rich information that could be derived from understanding social media, solving the various challenges of the

[1] www.bit.ly/2NZQ5L5
[2] www.lexico.com/en/definition/meme
[3] https://en.wikipedia.org/wiki/Pepe_the_Frog
[4] www.pri.org/stories/2019-07-16/memes-hong-kong-protests

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1179–1189
Barcelona, Spain (Online), December 12, 2020.

task also pose unique research value — such as the training of models that could understand reference derived from background knowledge and complex human expressions such as humor and sarcasm.

In this paper, we share the insights we gain from our time working with the task and dataset. In Section 2, we discuss some previous work and related approaches to tackling the problem. Then, in Section 3, we introduce the task and the dataset. Subsequently, in Section 4, we outline the features employed and the model architectures. Section 5 reports our results and model performance. In Section 6 and 7, we analyze our results and discuss the possible source and solutions to our errors and issues. We summarize our key results and conjectures in Section 8.

2 Related Work

Research on Memes, in particular, are few and far in between. Some works have experimented with memes as sentiment predictors: In their work on Facebook sentiment analysis, French (2017) reported a positive correlation between the category of the meme used and the affection of the discussion from its texts. French (2017) go on to confirm that memes were far more successful in conveying the sentiment of the debate over textual data. Other works attempt to automate the meme generation process (Peirson and Tolunay, 2018; Oliveira et al., 2016) but fail at replicating the humor expressed by particular memes. While Peirson and Tolunay (2018) attempts to re-purpose image captioning algorithms to capture the humor, Oliveira et al. (2016) attempt to use macros within news headlines to do the same. However, both works fail to generate coherent humor that can persuade annotators consistently. Another similar work by Wang and Wen (2015) apply multimodal techniques to generate captions for popular meme formats. Finally, the propagation and effect of memes on social media are also studied in works such as the one by Zannettou et al. (2018a) which provides an assessment of the popularity and use of certain memes, in the context of each community. In another work, Ferrara et al. (2013) aim to detect and analyze meme usage in social media streams, particularly by using an unsupervised clustering framework. Work in the space also make use of meta-information (Wang and Wen, 2015; Ferrara et al., 2013) regarding the meme to provide context to their frameworks. Apart from memes, other work on multimodal social media analysis and humor analysis could provide valuable insights. Shin et al. (2018) show how multimodal information can be used to enhance the analysis of unstructured data on social media. There are numerous works that attempt to predict the sentiment of images (Kanishcheva and Angelova, 2015; Xu et al., 2014). However, these works focus on a single modality — images. Works on Multimodal sentiment analysis like the one by Hu and Flaxman (2018) use both image and text embeddings in their models to show slight improvements over the text-only models. A comprehensive overview of multimodal sentiment analysis can be found in work, such as that Kaur and Kautish (2019) and Soleymani et al. (2017). In the realm of humor analysis — work such as the recent paper by Weller and Seppi (2019) have shown that deep learning models can outperform humans in classifying humor mainly due to a disparity in the sense of humor of the annotators and the testers. In an interesting work by Chandrasekaran et al. (2018) an attempt is made to capture wit by using synonyms of words in image descriptions to incite puns through the subversion of expectations. While the work reports impressive results and beats out humans in a controlled vocabulary setting, humans quickly regain dominance when the vocabulary is unconstrained.

3 Data

The task organizers have made available a dataset (Sharma et al., 2020) of 8879 annotated memes scrapped from various sources across the internet. Each meme was annotated by two annotators to ensure annotation quality. The text was extracted from the image using the Google OCR system and manually corrected by crowdsourced workers — to ensure that model accuracy doesn't depend on the quality of the OCR techniques used. We briefly describe each subtask of the Memotion Shared Task below:

- **Task A — Sentiment Classification:** Given a meme, the task is to classify it as a positive, negative or neutral meme.
- **Task B — Multilabel Characteristic Classification:** Given a meme, the system has to identify the existence of the following characteristics — humor, sarcasm, offense and motivation. Being a

multilabel classification task a meme can exhibit any combination of the above characteristics or none at all.

- **Task C — Scales of Semantic Classes:** The third task is to quantify the extent to which a particular effect is being expressed, i.e. if the meme is humorous whether it is funny, very funny or hilarious and so on.

The dataset shows a significant imbalance, particularly in the representation of the labels `negative`, `hilarious`, `very_twisted` and `hateful_offensive` of the sentiment, humor, sarcasm and offensive categories, respectively. We represent the distribution of labels in Table 1.

Categories	Tags	No. of Samples	Percentage(%)
Sentiment Analysis	`negative`	631	09.01
	`neutral`	2,205	31.50
	`positive`	4165	59.49
Humor	`not_funny`	1,651	23.58
	`funny`	2,457	35.10
	`very_funny`	2,241	32.01
	`hilarious`	652	09.31
Sarcasm	`general`	3,512	50.16
	`not_sarcastic`	1,546	22.09
	`twisted_meaning`	1,549	22.12
	`very_twisted`	394	05.63
Offensive	`not_offensive`	2,715	38.78
	`slight_offensive`	2,596	37.08
	`very_offensive`	1,469	20.98
	`hateful_offensive`	221	03.16
Motivational	`not_motivational`	4,530	64.70
	`motivational`	2,471	35.30

Table 1: Sample distribution of the corpus

4 Preprocessing and System Overview

Before we outline our features and models, we digress to briefly explain the steps we undertook to reduce noise in the text. We perform basic lower casing and the removal of unnecessary punctuation as the initial step. The second step was to remove specific noise inducing aspects such as the removal of URLs and User mentions (using regular expressions).

We did not perform any global image preprocessing steps; however, we did perform basic preprocessing steps such as resizing and grayscale conversion for specific feature extraction techniques.

4.1 Features

Depending on the model we use, we employ three different text vectorization techniques — word (1,2)-gram `TFIDF` features for the Logistic Regression model; A GLoVe (Pennington et al., 2014) embedding pretrained on 27 billion tweets for the BiLSTM + Attention based model; Contextual `BERT` embeddings (Devlin et al., 2019) for our transfer learning approach.

Beyond text vectorization we explore a few *hand crafted features* to improve model performance. We use features previously employed by Bertero and Fung (2016) in their work to detect humor in sitcoms and Mahajan and Zaveri (2017)'s submission to the SemEval 2017 Task 6. The stylistic features are as follows: number of words, number of parts-of-speech (POS) tags such as nouns, adjectives and verbs and their ratio to the number of words. The POS tags mentioned are obtained using CMU POS Tagger (Owoputi et al., 2013).

Ambiguity features are useful in representing the multiple meanings that can be delivered simultaneously as found in pun related humor (Yang et al., 2015; Miller and Gurevych, 2015). For this purpose we use a concept called *Synset* (short for Synonym set). A Synset is defined as a set of one or more synonyms

that can be interchangeably used in the same context to express the same meaning which was originally embedded. We derive these synonyms using the NLTK Corpus (Loper and Bird, 2002). For example the Synset for the word '*new*' would be { new, fresh, raw, newfangled, modern, newly }. The ambiguity features used are as follows: Mean Synset Length (it is the mean of the length of synset of each word of the text) , Maximum Synset Length (it is the maximum length of synset that a single word in the text can have), Synset Length Gap (it the difference between the Maximum Synset Length and Mean Synset Length).

Initially, we experimented with the use of pretrained feature extractors to extract image information. Our approach, using various pretrained models trained on the ILSVRC (Deng et al., 2009) such as the popular Inception_V2_ResNet (Szegedy et al., 2017), exhibited underwhelming results and sometimes proved to be detrimental to model performance. In this regard, we elected to employ more hand crafted features rather than pretrained feature extractors. For each image: hue, saturation and value are calculated by converting the RGB images into HSV channels using the scikit-image toolkit (van der Walt et al., 2014). We average the hue, saturation and value over all the pixels in the image to obtain the hue (H_{image}), saturation (S_{image}) and luminance (V_{image}) of the image. We also include RMS contrast features for each image — which can simply be defined as the standard deviation of pixel intensities (i.e. the brightness). We draw inspiration from previous work by Zhang et al. (2015b) and explore more features that can be quantitatively derived from the HSV model such as Colourfulness — a metric defined Hasler and Suesstrunk (2003) that exhibits high correlation to human perception of colourfulness — and metrics by Valdez and Mehrabian (1994) features – that relate brightness and saturation to the following emotions: Pleasure, Arousal and Dominance. Following this, we also entertain the idea of employing facial expressions as an extra image feature. To do this we re-purpose an opensource emotion detection application [5]. The approach used a convolution neural network at its core, to detect the following emotions — Angry, Disgusted, Fearful, Happy, Neutral, Sad and Surprised. The model is pretrained on the FER2013 Kaggle dataset [6] and used as feature extractor in our pipeline.

At this juncture, it might be apt to recognise the small dataset size and label imbalance exhibited by the Memotion Dataset (Sharma et al., 2020). To tackle the imbalance and data scarcity we consider many techniques such as oversampling, weighted training and resampling techniques (such as SMOTE (Chawla et al., 2002)) in our pipelines. We also consider a text augmentation technique found in work from Zhang et al. (2015a). We re-purpose open source code found on github [7]. The core of the idea is to extend the given dataset by simple word replacement wherein we replace certain words in the sentence with corresponding synonyms or phrase replacements pulled from an auxiliary database/dictionary. The image features are then replicated for the newly created synthetic sample. Our database of choice for such replacements was the *paraphrase bank*[8] database. We apply this technique on the training split of the data and duplicate the corresponding image and dense features.

4.2 Model Descriptions

For a baseline approach to the given problem, we elected an L2 regularized Logistic Regression as an ideal starting point. The input representation for text was `TF-IDF` uni-grams and bi-grams. We also use the handcrafted text and image dense features as well the emotion vectors with this model. All the modelling was done using the sci-kit learn toolkit (Pedregosa et al., 2011). This model uses the previously explained data augmentation technique as well as balanced class weights during training.

Our second model is an attention based deep learning model. Attention is a technique first introduced in the machine translation research space by Bahdanau et al. (2014). The idea was further extended to text classification by work such as the ones by Yang et al. (2016) and Raffel and Ellis (2015) — which are what we implement here. All modelling tasks for this model was carried out using keras with a tensorflow (Abadi et al., 2015) backend. The model uses the 'Adam' learning rate optimizer and categorical cross entropy loss function during the training phase. The model architecture is more intuitively described in

[5] `www.github.com/atulapra/Emotion-detection`
[6] `www.kaggle.com/deadskull7/fer2013`
[7] `www.github.com/Opla/SmallData-Augmentation-MachineLearning`
[8] `www.paraphrase.org`

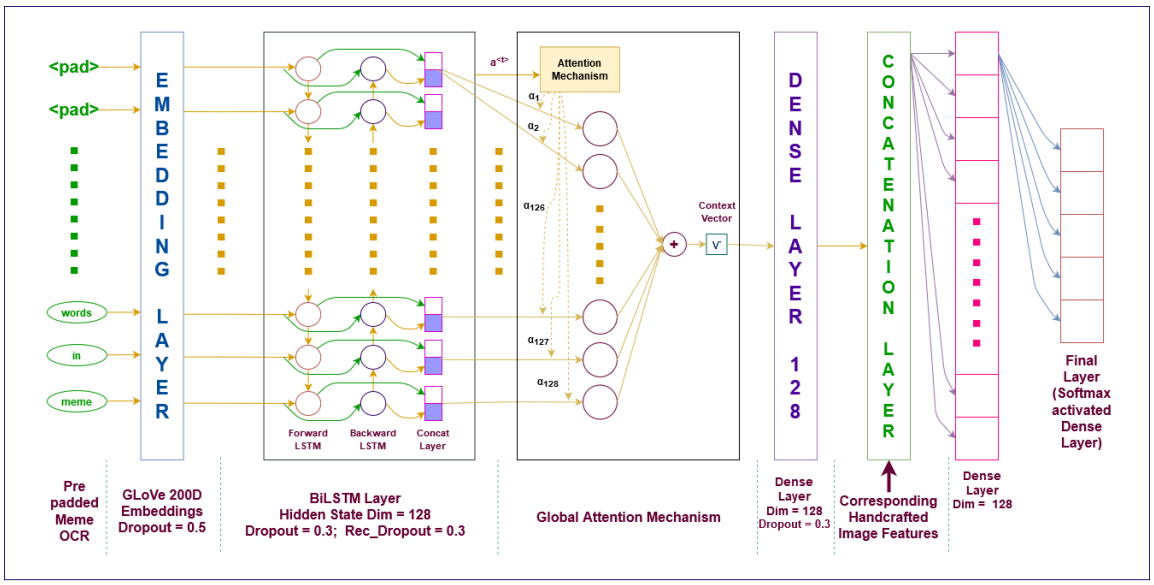

Figure 1: Attention informed model architecture

Figure 1. Due to the small dataset size we associate large dropout values with the embedding layer (0.5), LSTM layers (0.3) and Dense layers (0.3, apart from a L2 regularization kernel). We also implement early stopping, to further alleviate overfitting. This model also makes use of the data augmentation technique previously mentioned and SMOTE (Chawla et al., 2002) for balancing/resampling purposes. The model uses a 200D GLoVe (Pennington et al., 2014) embedding pretrained on 27 billion tweets for text representation and the handcrafted image features and emotion vectors for image representation (which is concatenated to the dense layer output which follows the Attention mechanism). The model was set to train for 100 epochs with a batch size of 2048 and initial learning rate of 0.001 but we found that the model started overfitting at an average of 40 epochs.

We also explore the possibility of leveraging dynamic contextual embeddings through the use of a transfer learning model — `BERT`. The developers of BERT provide a simple classification API for BERT through the `run_classifier` API available on their github page [9]. Our underlying model of choice was the $\texttt{BERT}_{base,uncased}$ — which trains a total of 110 million parameters, contains 12 transformer blocks and 12 self-attention heads with a hidden layer dimension of 768 . This model simply draws the [CLS] token embedding of the second to last layer of the BERT model for classification. During the training process the weights of the model are modified slightly to better cater to the task at hand. We finetune hyper parameters such as the learning rate, batch size and maximum sequence length to improve performance for different categories. In this case the model only takes advantage of the data augmentation and preprocessing techniques mentioned above, no extra image and dense features were provided.

5 Experimental Setup and Results

The dataset was provided to task participants as a pre-annotated training dataset (containing 7001 samples) and an un-annotated test dataset (containing 1878 samples). We first perform a model and ablation analysis using 10-fold cross validation methods on the training dataset and follow up with scores obtained by our systems on the test dataset, as provided by the task organizers. All results are represented using Accuracy and Macro F_1 metrics which were used for ranking the systems by the task organizers.

For the training Dataset, Initially, a validation dataset is also maintained to diagnose variance and bias issues that arise in the training phase and to aid hyperparameter tuning. Our overall train, validation, test split ratio is 80:10:10. However, for our final results, we conflate the validation and training set and represent 10-fold cross validation results. We represent our results for Subtask A and the categories within Subtask C (as subtask B and C only vary by semantic levels) in Table 2. We then perform an ablation

[9] `www.github.com/google-research/bert`

Model	Sentiment Analysis		Humor		Sarcasm		Offensive		Motivational	
	F_1 (%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)
Logistic Regression	31.97	**53.84**	**25.85**	31.67	**23.36**	42.12	**24.96**	**35.74**	49.71	53.75
Attention based Learner	24.43	36.67	22.48	30.00	19.62	25.54	23.18	31.18	46.82	**60.58**
Fine-tuned BERT	**32.47**	49.89	25.20	**33.00**	22.69	**43.21**	23.95	33.58	**50.29**	56.82

Table 2: 10-fold cross validation results (macro-F_1 and accuracy) on subtask A and C, with the training dataset

Features	Sentiment Analysis		Humor		Sarcasm		Offensive		Motivational	
	F_1 (%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)
`TFIDF Word (1,2)-gram + Dense Features`	27.15	**59.15**	23.01	33.84	17.60	**49.89**	22.04	37.57	42.09	**64.29**
+ balanced training	30.88	55.69	25.99	31.82	22.41	44.07	24.40	35.42	49.23	53.29
+ augmentation	28.15	57.95	23.81	33.84	18.70	49.04	22.76	37.44	43.87	63.21
+ image features	27.29	59.08	23.35	**33.94**	22.33	37.84	22.33	**37.84**	42.36	64.18
+ balanced training + augmentation	31.51	54.06	25.38	31.45	23.35	42.54	24.95	35.34	49.01	53.26
+ balanced training + image features	31.25	55.49	**26.02**	31.75	22.72	43.28	24.92	35.82	49.56	53.32
+ augmentation + image features	28.43	58.16	23.58	33.45	18.78	48.72	23.09	37.84	44.14	62.96
+ balanced training + augmentation + image features	**31.97**	53.84	25.85	31.67	**23.36**	42.12	**24.96**	35.74	**49.71**	53.75

Table 3: Ablation study results (macro-F_1 and accuracy) on subtask A and C, with the training dataset

Model	Subtask A		Subtask B		Subtask C	
	F_1 (%)	Rank	F_1(%)	Rank	F_1(%)	Rank
Final Mixed System	**32.48**	**24**	**49.94**	**11**	**30.74**	**15**
ML Model	31.44	28	49.98	11	30.74	15
Attention Model	31.68	28	49.34	18	28.69	21
BERT Model	32.48	24	49.75	11	29.81	19

Table 4: Test set results (macro-F_1 and potential ranks) on all subtasks. Official submission in bold.

analysis with a 10-fold cross validation split using the Logistic Regression model to better understand the effect of the various techniques and features employed and how they affect model performance. Due to the task specific nature of the experiment, we decide to carry out the experiment for all the categories — sentiment analysis, humor, sarcasm, offense and motivation classification. We represent the results in Table 3.

On the test dataset, the task organizers rank each system based on averaged Macro F_1. For subtasks B and C where there are 4 separate categories within the task an average score over the four categories was provided. Ranks were provided for only the top submission of each team. Initially, we submit three different systems for task evaluation — the Logistic Regression model, the BiLSTM + Attention model, the BERT model. However, we saw that the certain models performed better on certain subtasks. Therefore, for our final evaluation, a mixed system (referred to in Table 4 as Final Mixed System) — BERT for Subtask A, Attention for Subtask B (except motivational category) and Logistic Regression for Subtask C (including motivational category in Subtask B) — this represents our official submission to the task. We also represent the macro F_1 score and potential ranks of the individual models in Table 4.

Our model study revealed that Logistic Regression and the BERT model exhibit the best results, trading places based on what metric was considered. The results of the Attention based learner were underwhelming and were not competitive with the other models.

Our ablation analysis reveals that in general balancing and augmentation techniques provide the biggest performance improvements (on the basis of macro F_1). We also note that both these techniques were implemented to address the data scarcity and imbalance issue. Image features too, provide slight improvements when added (in all combinations). However, when these techniques are considered in

combination, the addition of augmentation over *text + balancing + image features*, lead to a drop in performance for the humor category and also in the sarcasm task, where the *text + image features* model outperforms the *text + augmentation + image features* model. In an additional test, we also observed that for the Attention based learner, image information provided little to no improvement on the performance.

On the test dataset, again BERT and the Logistic Regression model obtain the best results. We also observed that the Attention based learner exhibits lower performance on addition of image features, on all sub tasks.

6 Error Analysis

Due to the under-representation of specific labels, we saw our models were unable to classify samples into these classes effectively. This was especially apparent for the highly under-represented `hateful_offensive` (of the offensive category) and `very_twisted` (of the sarcasm category) where very few samples could be correctly classified. While we have reservations regarding the subjective nature of humor and offense, we generally found offensive memes wrongly classified as non-offensive and funny memes wrongly classified as not funny when the context mostly derived from the image. This points to our inability to capture image information effectively. This also indicates the absence of any background knowledge, which is imperative when understanding references that most memes tend to invoke. On analyzing the errors and ground truth of the motivation category, we find many of the annotations puzzling. We are, therefore, more curious as to what the annotation guidelines for this category are and refrain from making any conjectures regarding the classification errors. Another issue we would like to explore is the underwhelming results of the BiLSTM + Attention model. The main issues faced during the training phase are the small overall dataset size and under-representation of specific labels. This led to the inability to train larger and deeper models without facing high variance issues. Another issue that arose due to the small dataset size is the inability to train an image feature extractor from scratch (to alleviate data dissimilarity between the dataset and pretrained models) and better harness image information.

7 Discussion

In our time with the task of meme emotion analysis, we have come to believe that there are many inherent challenges in modelling the task apart from the machine learning algorithms being leveraged. We want to bring attention to the fact that memes are ever-changing — new memes enter and old memes leave the meme ecosystem very frequently and are quite often informed by current affairs. This could imply that training a model on historical meme samples could provide little insight into the meme ecosystem of today. The trends learnt by training with historical samples might not translate well to the trends of today. In the data collection and annotation phase, we make the following observations:

Meme Heterogeneity: Memes come in different forms and styles — or so-called templates. These templates can either represent a certain punchline or message and may not share the same referential image. For example, the set of memes {"The Scroll of Truth"[10], "Nancy Pelosi Ripping Paper"[11],"Most People rejected his message" [12]} represent the same idea of 'ignoring a fact or supposed truth' but use different referential images to enhance the humor. Memes belonging to the same template can be assumed to convey a similar message. There exists a countless number of such templates that are used online. Therefore, we think datasets should not attempt to draw memes from all templates but certain ubiquitous ones — as in the former case, the dataset does not contain enough samples to adequately represent the meme template's idea. As a product of this, models face the insurmountable task of learning trends from a large number of templates, each with a minimal set of samples. It may be apt to reverse this idea and create datasets with few selected templates but a large number of samples corresponding to each template.

Annotator Bias: As humans, many of us share different tastes and consequently have different senses of humor. What one may find hilarious might not affect another. Therefore, it may be the case that in tasks such as these, we are modelling humor specific to the annotators — thereby working with an annotator

[10] `https://knowyourmeme.com/memes/the-scroll-of-truth`
[11] `https://knowyourmeme.com/memes/nancy-pelosi-ripping-paper`
[12] `https://knowyourmeme.com/memes/most-people-rejected-his-message`

Annotators	Sentiment Analysis		humor		Sarcasm		Offensive		Motivational	
	F_1 (%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)	F_1(%)	Acc.(%)
Annotator 1	24.28	26.00	20.83	29.00	19.08	30.00	36.67	53.00	53.38	68.00
Annotator 2	27.92	32.00	18.64	22.00	25.40	35.00	31.82	41.00	46.15	49.00
Annotator 3	33.21	38.00	21.69	30.00	24.34	32.00	24.37	34.00	50.47	66.00
Annotator 4	32.54	36.00	28.32	32.00	34.63	39.00	27.24	44.00	47.92	72.00
Annotator Average	29.49	33.00	22.37	28.25	25.86	34.00	30.02	43.00	49.48	63.75
Model Performance	43.79	59.00	21.61	27.00	22.48	47.00	26.37	38.00	51.00	57.00

Table 5: Annotator bias experiment

bias. In a recent work by Weller and Seppi (2019), it was seen that a sample of random humans could not outperform a model trained on the same jokes. This indicates that the model has learned annotator specific biases and may not generalise well to the populace. We are also curious as to what could be the human error on a task such as this.

To address these concerns, we perform an experiment that randomly samples 100 memes from the training dataset (previously annotated) and gets four independent annotators to tag them on all five categories — sentiment, humor, sarcasm, offensive and motivational. The annotation process was mainly carried out in-house by the authors of this manuscript. We use basic guidelines (in the form of dictionary definitions and examples from the dataset) on tagging the categories and labels. We then calculate macro F_1 scores and accuracy for each annotator and compare their performance with the Logistic Regression model's predictions on the same 100 memes. We also calculate inter-annotator agreement metrics to understand better if all the annotators are on the same page. We represent the results of this in Table 5. On calculating Randolph (2005) free-marginal multi-rater Kappa metrics, we found poor agreement on annotations over all the categories — Sentiment (0.22), humor (0.10), Sarcasm (0.09), Offense (0.36), Motivational (0.38). The poor agreement scores are an indicator of how concepts such as humor and offense can be subjective. We also find that the human annotators barely outperform our Logistic Regression model, which are indicators of low human performance in such tasks. Few Annotators also report different memes as random or nonsensical. On manual checking, we found these memes to be related to pieces of media (e.g., Star Wars, Lord of the Rings) that the annotator was not familiar with.

8 Conclusion

Our work on this dataset highlights the difficulty in effectively harnessing image information to understand reference and media that are part and parcel of a meme. Consistent with previous work, we are only able to obtain small performance improvements by employing image-based features. We also highlight modelling issues that arise due to the heterogeneous nature of data (a small number of samples, a large number of templates) and the high variance issues. We also bring to question — annotation and data collection practices that might not immediately translate to the task of meme analysis. A general suggestion would be to use a large number of annotators or to harness popularity metrics such as likes and upvotes used on websites like Reddit and Instagram. Popularity metrics could provide a more generalized view of the humor or other characteristics of the meme. We think that memes can be imperative to understanding user sentiment in the present internet ecosystem and beyond. We look forward to the new interest that enters this space due to work done here and in the memotion shared task.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals,

Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Dario Bertero and Pascale Fung. 2016. Predicting humor response in dialogues from tv sitcoms. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5780–5784.

Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. 2018. Punny captions: Witty wordplay in image descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 770–775, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16(1):321–357, June.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. 2013. Clustering memes in social media. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 548–555. IEEE.

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

David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. International Society for Optics and Photonics.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 350–358, New York, NY, USA. ACM.

Olga Kanishcheva and Galia Angelova. 2015. About emotion identification in visual sentiment analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 258–265, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, Bulgaria.

Ramandeep Kaur and Sandeep Kautish. 2019. Multimodal sentiment analysis: A survey and comparison. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, 10(2):38–58.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rutal Mahajan and Mukesh Zaveri. 2017. SVNIT @ SemEval 2017 task-6: Learning a sense of humor using supervised approach. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 411–415, Vancouver, Canada, August. Association for Computational Linguistics.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of English puns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–729, Beijing, China, July. Association for Computational Linguistics.

Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. 2016. One does not simply produce funny memes!–explorations on the automatic generation of internet humor. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016). Paris, France*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830, November.

Abel L. Peirson and E. Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks. *CoRR*, abs/1806.04510.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 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, October. Association for Computational Linguistics.

Colin Raffel and Daniel P. W. Ellis. 2015. Feed-forward networks with attention can solve some long-term memory problems. *CoRR*, abs/1512.08756.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

Donghyuk Shin, Shu He, Gene Moo Lee, Andrew B Whinston, Suleyman Cetintas, and Kuang-Chih Lee. 2018. Enhancing social media analysis with visual data analytics: A deep learning approach. *Available at SSRN 2830377*.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4278–4284. AAAI Press.

Patrick Valdez and Amirhossein Mehrabian. 1994. Effects of color on emotions. *Journal of experimental psychology. General*, 123 4:394–409.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. 2014. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6.

William Yang Wang and Miaomiao Wen. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 355–365, Denver, Colorado, May–June. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3612–3616, Hong Kong, China, November. Association for Computational Linguistics.

Can Xu, Suleyman Cetintas, Kuang chih Lee, and Li-Jia Li. 2014. Visual sentiment prediction with deep convolutional neural networks. *ArXiv*, abs/1411.5731.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal, September. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June. Association for Computational Linguistics.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018a. On the origins of memes by means of fringe web communities. *CoRR*, abs/1805.12512.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018b. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 188–202, New York, NY, USA. ACM.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Yaowen Zhang, Lin Shang, and Xiuyi Jia. 2015b. Sentiment analysis on microblogging by integrating text and image features. In Tru Cao, Ee-Peng Lim, Zhi-Hua Zhou, Tu-Bao Ho, David Cheung, and Hiroshi Motoda, editors, *Advances in Knowledge Discovery and Data Mining*, pages 52–63, Cham. Springer International Publishing.