

BLUE at Memotion 2.0 2022: You have my Image, my Text and my Transformer

Ana-Maria Bucur^{1,2}, Adrian Cosma³ and Ioan-Bogdan Iordache⁴

¹Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania

²Universitat Politècnica de València, Spain

⁴Faculty of Automatics and Control, University Politehnica of Bucharest, Romania

³University of Bucharest, Romania

Abstract

Memes are prevalent on the internet and continue to grow and evolve alongside our culture. An automatic understanding of memes propagating on the internet can shed light on the general sentiment and cultural attitudes of people. In this work, we present team BLUE's solution for the second edition of the MEMOTION shared task. We showcase two approaches for meme classification (i.e. sentiment, humour, offensive, sarcasm and motivation levels) using a text-only method using BERT, and a Multi-Modal-Multi-Task transformer network that operates on both the meme image and its caption to output the final scores. In both approaches, we leverage state-of-the-art pretrained models for text (BERT, Sentence Transformer) and image processing (EfficientNetV4, CLIP). Through our efforts, we obtain first place in task A, second place in task B and third place in task C. In addition, our team obtained the highest average score for all three tasks.

Keywords

memotion, memes, multi-modal network, multi-task learning, transformers, ordinal regression, fine-tuning

1. Introduction

The concept of a meme was first introduced by Richard Dawkins [1], as a fundamental unit of propagation of ideas and cultural information, similar to genes for transmitting genetic information across time. Dawkins proposed a memetic theory, in which memes have very similar patterns evolution by natural selection as genes, in which memes evolve and replicate across time, through mutation and cross-over with other memes. While this theory was criticized [2] from the onset, it remained a valuable tool for viral marketing, social analytics and understanding of cultural evolution across history.


A widespread usage of the term "meme", aside from the more academic definitions from memetic theories, is that of internet memes in the sense of catch-phrases, images, gifs and videos. Internet memes "evolved" from simple images templates into modern, ironic and absurdist images, similar to tendencies in postmodern art (i.e. "deep fried" memes, dank memes).

De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada

✉ ana-maria.bucur@drd.unibuc.ro (A. Bucur); cosma.i.adrian@gmail.com (A. Cosma); iordache.bogdan1998@gmail.com (I. Iordache)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Websites such as 9gag, Tumblr, and Reddit were at the forefront of internet meme propagation and mainstream spread.

Internet memes continue to grow and evolve alongside our culture. An automatic understanding of memes propagating on the internet can shed light on people’s general sentiments and cultural attitudes.

In this work, we present team BLUE’s solution to the 2022 edition of the MEMOTION 2.0 shared task [3]. We focused our efforts on two main approaches: i) text-based fine-tuning using BERT and ii) a Multi-Modal-Multi-Task transformer that uses features from both images and text. Furthermore, we provide ablation studies on different modalities and training parameters, and show that there is not one combination of modalities suitable for all tasks.

In the following sections, we make an overview of related methods from the previous shared task, briefly describe the task, describe the available training, validation and testing data for the current edition of the shared task. Finally, we describe our two approaches and report our results on both validation and test sets.

2. Related Work

With the increase in popularity of social media websites (e.g. Facebook, Reddit, Twitter), NLP researchers started using the textual data collected from these platforms for detecting emotions [4, 5, 6], offensive content [7, 8, 9], hate speech [10, 11], humour [12, 13, 14], sarcasm [15, 16, 17], pejorative language [18], inspirational content [19], optimism [20, 21] and the manifestations of mental health problems such as depression [22, 23], suicide ideation [24, 25] and anxiety [26]. Researchers explored the online content from social media even further and began focusing on the multi-modal data [27, 28], including internet memes. Efforts to automatically detect the offensive [29] or harmful memes [30] are being made to help the content moderators in charge of removing the posts containing hate speech.

Several competitions took advantage of the high availability of internet memes and used the multi-modal data for various tasks: DANKMEMES from Evalita 2020 [31], MEMOTION from SemEval 2020 [32], The Hateful Meme Challenge [33], Fine Grained Hateful Memes Detection Shared Task from The 5th Workshop on Online Abuse and Harms [34], Detection of Persuasion Techniques in Texts and Images from SemEval 2021 [35], Multimodal Fact-Checking Task from the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (De-Factify) ¹ and Multimedia Automatic Misogyny Identification (MAMI) from SemEval 2022 ².

In the first iteration of the MEMOTION task at SemEval 2020 [32], the participating teams surpassed the baseline models by only a small percentage. Two participating teams used only the textual data extracted from the memes for all their experiments. The rest of the teams experimented with systems using visual-only or textual-only information or the fusion of both features. The majority of the participating teams relied on approaches based on pretrained models such as ResNet [36], VGG-16 [37] and Inception-ResNet [38] to extract the visual features. For textual information the teams used approaches based on Recurrent Neural Network architectures or pretrained transformer models such as BERT [39]. For task A, the system with

¹<https://aiisc.ai/defactify/>

²<https://github.com/MIND-Lab/MAMI>

the best performance used only the textual data from the internet memes for sentiment detection. For tasks B and C, the best performing systems used both image and text data for identifying the emotion of the memes.

In our participation in the MEMOTION 2.0 shared task, we used two approaches: a text-only approach using BERT and a Multi-Modal-Multi-Output transformer using both visual and textual features. For the second approach, we also used features extracted from CLIP, this model being suitable for multi-modal inputs.

3. Task Description

The second iteration of the MEMOTION shared task, previously conducted at SemEval 2020 [32], is comprised of three tasks for detecting the sentiment and the emotions of memes as described below:

- **Task A: Sentiment Analysis:** Identify if a meme is positive, negative or neutral.
- **Task B: Emotion Classification:** Identify the emotion expressed by a meme: humour, sarcasm, offensive and motivation. A meme can convey more than one emotion.
- **Task C: Scales/Intensity of Emotion Classes:** Quantify to which extent a particular emotion is being expressed in a meme. The intensities are on a scale from 0 to 3 for humour, sarcasm and offensiveness (e.g. 0 - not funny, 1 - funny, 2 - very funny, 3 - hilarious) and only 0 and 1 for motivation (0 - not motivational, 1 - motivational).

The tasks are challenging, as identifying the sentiment and emotion in a meme is more complex than performing the same task only on textual data. For memes, comprised of image and text information, a multi-modal approach for understanding both visual and textual cues is needed. The dataset from the shared task contains memes with overlapping emotions, increasing the difficulty of the tasks. Most funny memes are also sarcastic, and some motivational memes are also offensive [32].

The teams' performance is evaluated by the weighted F1 score for task A. For tasks B and C, the weighted F1 score is computed for each subtask (humour, sarcasm, offensive, motivation), and the average F1 score of these subtasks is used to rank the systems.

4. Data

The dataset used in the MEMOTION 2.0 shared task [40] is comprised of internet memes collected from the public domain. These memes were annotated by Amazon Mechanical Turk workers for sentiment and emotion. The dataset contains the image of the memes and the corresponding OCR extracted text. The annotators were also asked to provide the corrected text if the OCR extracted text was inaccurate.

The training set is comprised of 7K memes, and the validation and test splits contain 1.5K memes each. The distribution of labels for the three splits is presented in Table 1, the dataset is heavily imbalanced. For the sentiment labels, there are more positive memes than negative ones in the training and validation splits, while in the test split, there is only a very small number of memes with positive sentiment.

Label	Train data				Validation data				Test data			
	Negative	Neutral	Positive		Negative	Neutral	Positive		Negative	Neutral	Positive	
Sentiment	973	4510	1517		200	975	325		451	971	78	
Label	0	1	2	3	0	1	2	3	0	1	2	3
Humour	918	3666	1865	551	229	745	419	107	62	892	398	148
Sarcasm	3871	1759	1069	301	804	388	246	62	185	248	892	175
Offensiveness	5182	1107	529	182	1110	238	107	45	943	457	87	13
Motivation	6714	286	-	-	1430	70	-	-	1480	20	-	-

Table 1

The distribution of labels for sentiment and emotion in the MEMOTION 2.0 tasks. The dataset is heavily imbalanced.

Given the multi-modal content found in internet memes, the textual or visual information alone may not be sufficient for identifying the sentiment and the emotions in a meme. Some examples from the dataset are presented in Figure 1, in which visual content is necessary for a complete understanding of the meme context. As such, we propose, alongside a text-only based method, a fusion approach combining both visual and textual information through a Multi-Modal-Multi-Task transformer.



Figure 1: Example of memes from the dataset used in the MEMOTION 2.0 shared task. As seen in these examples, the memes cannot be understood without considering both the visual and textual cues.

5. Method

In this section, we describe two methods for meme classification, as previously mentioned. We participated in all three tasks from the MEMOTION 2.0 shared task and used two approaches: a text-only meme classification using BERT [41] and a Multi-Modal-Multi-Task transformer

network.

5.1. Text-Only Multi-Task Meme Classification

Previous methods have shown that combining image features with textual features can bring more noise to the dataset [42], and for some tasks, the best results were achieved by text-only approaches [43].

We propose a text-only multi-task method in which we extract textual features using a pretrained BERT model [39]. Neural network architectures based on transformers [44], such as BERT, were shown to obtain great performance on many different Natural Language Processing tasks. Moreover, these results can be obtained by training such architectures on a large set of texts and then fine-tuning the model for different downstream tasks.

Processing Pipeline. Our implementation is based on a BERT base model provided by the HuggingFace library [45], consisting of a 12-layer transformer, with a hidden size of 768 and 12 attention heads. The OCR-extracted text of a meme is firstly tokenized using a SentencePiece [46] tokenizer provided with the BERT model, then the list of tokens is passed through the encoder. The features extracted by BERT (the embedding of the [CLS] token) are passed through 5 classification heads, corresponding to each of the emotions defined for a meme. A classification head is implemented as a feed-forward layer with the output size specified by the task.

Training details. During training, we applied dropout with a rate of 0.1 between the feature extractor and the classification heads. We used cross-entropy loss to compute the loss for each task. The training was done in mini-batches of size 16 (each batch containing examples for a single task), for 5 epochs, saving the model with the best performance on the validation set. We fine-tuned the architecture end-to-end using the AdamW optimizer [47], with a learning rate of 0.00002, decreased at each step using a linear scheduler, and no weight decay. For each task, we randomly oversampled instances from the training set, due to heavy dataset imbalance.

5.2. Multi-Modal-Multi-Task Transformer (MMMT)

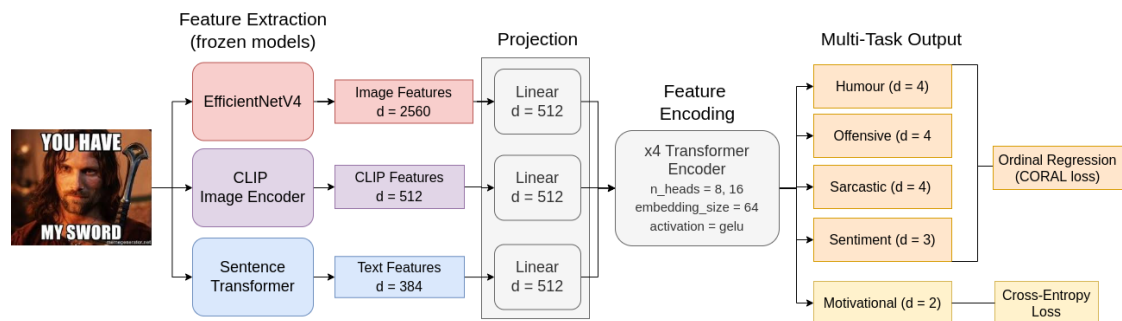


Figure 2: The general architecture of our Multi-Modal-Multi-Task model. We use image features from a pretrained EfficientNetV4 [48], CLIP [49] features and sentence features from a pretrained sentence transformer [50]. The features are passed through a 4-layer transformer encoder which outputs a classification head for each required aspect of the memes.

Internet memes are inherently multi-modal, often having a pop-culture image reference and a caption that accompanies the image, overlaid on top of it. While the meaning can be estimated using only the caption, images offer important additional context for higher-level semantic understanding and differentiating between types of memes (i.e. dank, deep-fried, "classical"), each with its predominant sentiment.

Several previous methods have reported using multi-modal approaches in the computational pipeline [27, 28, 29, 30]. In our pipeline, however, we explore semantic image features in two ways: i) direct image features provided by a pretrained EfficientNetV4 [48] on ImageNet dataset [51], and ii) features from the image encoder of CLIP [49]. CLIP was trained on a large-scale multi-modal dataset of image-text pairs in a contrastive learning fashion for use in zero-shot image classification. Features extracted from CLIP have been shown to respond to multi-modal inputs [52], such as the same concept being represented explicitly as an image, a high level representation of it (i.e. a sketch), or in written form. This makes CLIP suitable for use in our scenario. However, fine-tuning classification results are still lagging behind pretrained models trained only on images, so we also opted for using direct image features from EfficientNetV4 [48].

Processing Pipeline. Our pipeline is described in Figure 2. The first step in the computation is feature extraction using pretrained models. These models are frozen, and are not fine-tuned during training. We employ EfficientNetV4 [48] and CLIP image encoder [49] for image features, and a Sentence Transformer [50] for processing the meme caption. Each of these models outputs a vector of different dimensionality, so we employ a different linear projection layer to change the dimensionality to a common 512-element vector. These 3 vectors are considered a set of features and are processed using a popular transformer architecture. The order of the different features does not matter in the final computation, and for that reason, we do not employ positional embeddings. The final features are averaged and are then followed by an output layer corresponding to each aspect of memes (humour, sarcasm, offensive, sentiment and motivation). For humour, sarcastic, offensive and sentiment heads, we employed CORAL [53] loss for ordinal regression, to consistently estimate the degree of humour, sarcasm, etc.

Training details. The transformer network has 4 layers, with 8, 8, 16 and 16 attention heads, respectively. The internal embedding size is 64, and we used GELU activation [54]. We trained the network for 100 epochs, or until it overfits the validation set, with a batch size of 256 and Adam optimizer [47]. We employed a cyclical, triangular learning rate schedule [55], with a step size of 5 epochs for a 10x increase in learning rate, and an initial learning rate of 0.0001. Since the training data is severely imbalanced, we oversampled minority classes.

6. Experiments & Results

6.1. Text-Only Experiments

In order to measure the benefits of multi-task learning for classifying emotion intensities, we performed an ablation study by comparing the weighted F1 scores computed for each emotion intensity predictions made by models trained in two settings. Firstly, we trained independent models for each of the emotion subtasks as defined by tasks A and C. Secondly, we trained a single model in the multi-task setting, by fine-tuning the BERT encoder on all of the emotion

Setting	Sentiment	Humour	Sarcasm	Offensive	Motivation
Single-Task	0.5184	0.3682	0.3641	0.6145	0.9285
Multi-Task	0.5054	0.3752	0.3973	0.6148	0.9286

Table 2

Comparing the weighted F1 scores computed on the validation dataset for tasks A and C, using the text-only approach and training the model either in the multi-task setting, or independently for each of the emotions

subtasks at once.

From Table 2 we can see that, with the exception of sentiment classification, the predictions of all emotion intensities have benefited from being learnt jointly by the model. Following these observations, our submission used the single-task model to predict the sentiment labels for the test dataset (task A), and the multi-task model to predict all other fine-grained emotion labels (task C).

The difference between task B and task C is that for humour, sarcasm and offensive emotions, task B is a binary classification task instead of a multi-class one. The previously defined models trained for task C can be used to make predictions for task B, by mapping the non-zero intensity predictions to the positive class in the binary setting. Because of this, we wanted to see if we could gain any performance improvement by training the classification heads specifically for the classification defined by task B.

Table 3 displays the performance obtained for each setting. When training in the multi-task paradigm, we use the same model definition and training strategy and we only change the classification heads’ output sizes to 2 (for humour, sarcasm and offensive emotions). We observe that the best performing method is training single-task models for each of the three subtasks. Thus, we used these models’ predictions on the test dataset for our task B submission.

Target Task	Setting	Humour	Sarcasm	Offensive
Task B	single-task	0.7817	0.52	0.6529
Task B	multi-task	0.6551	0.5081	0.5404
Task C	single-task	0.7688	0.4746	0.6316
Task C	multi-task	0.7676	0.49	0.6368

Table 3

Performance of text-only models on predicting binary labels for humour, sarcasm and offensive emotions (task B). We compare the performance of models trained for classifying the various levels of emotion intensity (task C) and used to predict the binary labels by assigning the positive class to all non-zero intensities, with the performance of models trained directly for the binary classification subtask. We also compare training on either the single-task or the multi-task setting.

6.2. Multi-Modal Experiments

Table 4 showcases an ablation study performed on the different modalities. The architecture remains the same in all situations. While the results are close, it is clear that the addition of CLIP image features in the computation positively improved overall results. For the final submission, we used all the available modalities.

Features Used	Task A	Task B	Task C	Mean
Only Text	0.5127	0.6494	0.5001	0.5541
Only Image	0.5139	0.6404	0.5117	0.5553
Only CLIP	0.5113	0.6559	0.4835	0.5502
Image + Text	0.5118	0.6452	0.5041	0.5537
CLIP + Image	0.5077	0.6398	0.5053	0.5510
CLIP + Text	0.5118	0.6551	0.5032	0.5567
<i>Image + CLIP + Text (submission)</i>	0.5178	<i>0.6394</i>	<i>0.5029</i>	<i>0.5534</i>

Table 4

Ablation study of our MMT Transformer. The training conditions are the same except the different input features present. The differences between modalities are very small.

6.3. Shared Task Results

Emotion	Task	Only Text	MMMT
Sentiment	Task A	0.5072	0.5318
Humour	Task B	0.9239	0.8111
Humour	Task C	0.4131	0.4036
Sarcasm	Task B	0.6386	0.8191
Sarcasm	Task C	0.1604	0.3083
Offensive	Task B	0.5581	0.485
Offensive	Task C	0.5045	0.485
Motivation	Tasks B & C	0.9764	0.98

Table 5

Weighted F1 scores computed for each emotion defined in the test dataset.

We report in Table 5 the scores obtained by both of our approaches, separately for each emotion subtask from the test dataset. We observe that none of the models outperforms the

other on all subtasks. The text-only approach seems to be better suited for identifying humour and offensiveness, while the multi-modal model performs better on all of the other emotions. Looking at Table 6, even when comparing the models by their results on the three main tasks, the multi-modal approach does better only on tasks A and C.

We also provide the scores achieved by all participating teams for each task in Table 7. Our team managed to place first for task A, second for task B and third for task C. Moreover, our team obtained the highest average score across the three tasks.

Model	Task A	Task B	Task C	Mean
Only Text	0.5072	0.7743	0.5136	0.5984
MMMT	0.5318	0.7738	0.5443	0.6166

Table 6
Weighted F1 scores computed for each task for the test dataset.

Team Name	Task A	Task B	Task C	Mean
<i>BLUE (our team)</i>	0.5318	0.8059	0.5443	0.6273
BROWALLIA [56]	0.5255	0.767	0.5453	0.6126
Amazon PARS [57]	0.5025	0.7609	0.5564	0.6066
HCILab [58]	0.4995	0.7414	0.5301	0.5903
weipengfei	0.4887	0.6915	0.5033	0.5612
BASELINE	0.434	0.7358	0.5105	0.5601
Yet [59]	0.5088	0.6106	0.51	0.5431
Greeny	0.5037	0.6106	0.484	0.5328
Little Flower [60]	0.5081	0.8229	N/A	N/A*

Table 7
Scores obtained by all of the participating teams, for each task. We also report the average score achieved by the teams over all of the three tasks. Our team obtained the highest average score across the three tasks. * the team participated only on tasks A and B.

7. Conclusions & Future Work

This work presented team BLUE’s approach for the 2022 edition of the MEMOTION 2.0 workshop. We described two solutions for meme classification: i) text-only approach through fine-tuning a BERT model and ii) a Multi-Modal-Multi-Task transformer network that operates on both images and text. Different from most previous methods, we employed CORAL [53] for performing ordinal regression for ordinal outputs (e.g. humour intensity).

By making use of powerful, state-of-the-art, pretrained models for text and images, we obtained the first place on task A with a weighted F1 score of 0.5318, second place on task B with a score of 0.8059 and third place on task C with a score of 0.5453. In addition, we obtain the highest average score for all three tasks.

For future work, we aim to address the issue of severely imbalanced training data and small amount of images by designing a pipeline for self-supervised pretraining on internet-meme images. By only fine-tuning on a small and densely annotated images, the model is more robust to overfitting and predicting the majority class in training.

Moreover, the text in the MEMOTION 2.0 dataset is cleaned by human annotators. However, for a large-scale meme dataset used for pretraining, one can employ lexical normalization models [61, 62] to automatically correct faulty OCR and transform the text to its canonical form, which was a significant problem in computational pipelines from the first edition of this shared task.

References

- [1] R. Dawkins, *The Selfish Gene*, Oxford University Press, Oxford, UK, 1976.
- [2] M. Midgley, *The Solitary Self: Darwin and the Selfish Gene*, Routledge, 2014.
- [3] P. Patwa, S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Findings of memotion 2: Sentiment and emotion analysis of memes, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [4] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, GoEmotions: A dataset of fine-grained emotions, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 4040–4054. URL: <https://aclanthology.org/2020.acl-main.372>. doi:10.18653/v1/2020.acl-main.372.
- [5] Z. Jianqiang, G. Xiaolin, Z. Xuejun, Deep convolution neural networks for twitter sentiment analysis, *IEEE Access* 6 (2018) 23253–23260.
- [6] N. Alvarez-Gonzalez, A. Kaltenbrunner, V. Gómez, Uncovering the limits of text-based emotion detection, in: *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2560–2583. URL: <https://aclanthology.org/2021.findings-emnlp.219>.
- [7] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.
- [8] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1415–1420. URL: <https://aclanthology.org/N19-1144>. doi:10.18653/v1/N19-1144.
- [9] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, P. Nakov, SOLID: A large-scale

- semi-supervised dataset for offensive language identification, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 915–928. URL: <https://aclanthology.org/2021.findings-acl.80>. doi:10.18653/v1/2021.findings-acl.80.
- [10] R. Cao, R. K.-W. Lee, T.-A. Hoang, DeepHate: Hate speech detection via multi-faceted text representations, in: 12th ACM Conference on Web Science, 2020, pp. 11–20.
- [11] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, arXiv preprint arXiv:2004.06465 (2020).
- [12] A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: The figurative language of social media, *Data & Knowledge Engineering* 74 (2012) 1–12. URL: <https://www.sciencedirect.com/science/article/pii/S0169023X12000237>. doi:<https://doi.org/10.1016/j.datak.2012.02.005>, applications of Natural Language to Information Systems.
- [13] O. Weller, K. Seppi, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3621–3625. URL: <https://aclanthology.org/D19-1372>. doi:10.18653/v1/D19-1372.
- [14] R. Ortega-Bueno, C. E. Muniz-Cuza, J. E. M. Pagola, P. Rosso, Uo upv: Deep linguistic humor detection in spanish social media, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), 2018, pp. 204–213.
- [15] J. Plepi, L. Flek, Perceived and intended sarcasm detection with graph attention networks, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4746–4753. URL: <https://aclanthology.org/2021.findings-emnlp.408>.
- [16] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, R. Mihalcea, CASCADE: Contextual sarcasm detection in online discussion forums, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1837–1848. URL: <https://aclanthology.org/C18-1156>.
- [17] D. Bamman, N. A. Smith, Contextualized sarcasm detection on twitter, in: Ninth international AAAI conference on web and social media, 2015.
- [18] L. P. Dinu, I.-B. Iordache, A. S. Uban, M. Zampieri, A computational exploration of pejorative language in social media, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3493–3498. URL: <https://aclanthology.org/2021.findings-emnlp.296>.
- [19] O. Ignat, Y.-L. Boureau, A. Y. Jane, A. Halevy, Detecting inspiring content on social media, in: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE Computer Society, 2021, pp. 1–8.
- [20] C. Caragea, L. P. Dinu, B. Dumitru, Exploring optimism and pessimism in Twitter using deep learning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018,

- pp. 652–658. URL: <https://aclanthology.org/D18-1067>. doi:10.18653/v1/D18-1067.
- [21] X. Ruan, S. Wilson, R. Mihalcea, Finding optimists and pessimists on Twitter, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 320–325. URL: <https://aclanthology.org/P16-2052>. doi:10.18653/v1/P16-2052.
- [22] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, D. Inkpen, Deep learning for depression detection of Twitter users, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, Association for Computational Linguistics, New Orleans, LA, 2018, pp. 88–97. URL: <https://aclanthology.org/W18-0609>. doi:10.18653/v1/W18-0609.
- [23] A.-M. Bucur, A. Cosma, L. P. Dinu, Early risk detection of pathological gambling, self-harm and depression using bert, CLEF (Working Notes) (2021).
- [24] R. Sawhney, H. Joshi, R. R. Shah, L. Flek, Suicide ideation detection via social and temporal user representations using hyperbolic learning, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2176–2190. URL: <https://aclanthology.org/2021.naacl-main.176>. doi:10.18653/v1/2021.naacl-main.176.
- [25] G. Coppersmith, R. Leary, E. Whyne, T. Wood, Quantifying suicidal ideation via language usage on social media, in: Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM, volume 110, 2015.
- [26] J. H. Shen, F. Rudzicz, Detecting anxiety through Reddit, in: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology – From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Vancouver, BC, 2017, pp. 58–65. URL: <https://aclanthology.org/W17-3107>. doi:10.18653/v1/W17-3107.
- [27] R. Schifanella, P. de Juan, J. Tetreault, L. Cao, Detecting sarcasm in multimodal social platforms, in: Proceedings of the 24th ACM International Conference on Multimedia, MM ’16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1136–1145. URL: <https://doi.org/10.1145/2964284.2964321>. doi:10.1145/2964284.2964321.
- [28] R. Gomez, J. Gibert, L. Gomez, D. Karatzas, Exploring hate speech detection in multimodal publications, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1470–1478.
- [29] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 32–41. URL: <https://aclanthology.org/2020.trac-1.6>.
- [30] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, T. Chakraborty, MOMENTA: A multimodal framework for detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4439–4455. URL: <https://aclanthology.org/2021.findings-emnlp.379>.
- [31] M. Miliiani, G. Giorgi, I. Rama, G. Anselmi, G. E. Lebani, Dankmemes @ evalita 2020: The memeing of life: Memes, multimodality and politics, in: Proceedings of the 7th evaluation

- campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), 2020.
- [32] C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, B. Gambäck, SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor!, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 759–773. URL: <https://aclanthology.org/2020.semeval-1.99>. doi:10.18653/v1/2020.semeval-1.99.
- [33] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, *Advances in Neural Information Processing Systems* 33 (2020).
- [34] L. Mathias, S. Nie, A. Mostafazadeh Davani, D. Kiela, V. Prabhakaran, B. Vidgen, Z. Waseem, Findings of the WOAAH 5 shared task on fine grained hateful memes detection, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online, 2021, pp. 201–206. URL: <https://aclanthology.org/2021.woah-1.21>. doi:10.18653/v1/2021.woah-1.21.
- [35] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 70–98. URL: <https://aclanthology.org/2021.semeval-1.7>. doi:10.18653/v1/2021.semeval-1.7.
- [36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: <http://arxiv.org/abs/1409.1556>.
- [38] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.
- [39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [40] S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Memotion 2: Dataset on sentiment and emotion analysis of memes, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
- [41] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational

- Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [42] L. Bonheme, M. Grzes, SESAM at SemEval-2020 task 8: Investigating the relationship between image and text in sentiment analysis of memes, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 804–816. URL: <https://aclanthology.org/2020.semeval-1.102>. doi:10.18653/v1/2020.semeval-1.102.
- [43] V. Keswani, S. Singh, S. Agarwal, A. Modi, IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of Internet memes, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1135–1140. URL: <https://aclanthology.org/2020.semeval-1.150>. doi:10.18653/v1/2020.semeval-1.150.
- [44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [45] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://aclanthology.org/2020.emnlp-demos.6>. doi:10.18653/v1/2020.emnlp-demos.6.
- [46] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. URL: <https://aclanthology.org/D18-2012>. doi:10.18653/v1/D18-2012.
- [47] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
- [48] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.
- [49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
- [50] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410>. doi:10.18653/v1/D19-1410.
- [51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition,

Ieee, 2009, pp. 248–255.

- [52] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, C. Olah, Multimodal neurons in artificial neural networks, *Distill* 6 (2021) e30.
- [53] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, *Pattern Recognition Letters* 140 (2020) 325–331. URL: <http://www.sciencedirect.com/science/article/pii/S016786552030413X>. doi:<https://doi.org/10.1016/j.patrec.2020.11.008>.
- [54] D. Hendrycks, K. Gimpel, Bridging nonlinearities and stochastic regularizers with gaussian error linear units, *CoRR* abs/1606.08415 (2016). URL: <http://arxiv.org/abs/1606.08415>. arXiv:1606.08415.
- [55] L. N. Smith, No more pesky learning rate guessing games, *CoRR* abs/1506.01186 (2015). URL: <http://arxiv.org/abs/1506.01186>. arXiv:1506.01186.
- [56] B. Duan, Y. Zhu, Browallia at memotion 2.0 2022 : Multimodal memotion analysis with modified ogb strategies, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [57] G. G. Lee, M. Shen, Amazon pars at memotion 2.0 2022: Multi-modal multi-task learning for memotion 2.0 challenge, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [58] T. T. Nguyen, N. T. Pham, N. D. Nguyen, H. Nguyen, L. H. Nguyen, Y.-G. Kim, Hcilab at memotion 2.0 2022: Analysis of sentiment, emotion and intensity of emotion classes from meme images using single and multi modalities, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [59] Y. Zhuang, Y. Zhang, Yet at memotion 2.0 2022 : Hate speech detection combining bilstm and fully connected layers, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [60] K. N. Phan, G.-S. Lee, H.-J. Yang, S.-H. Kim, Little flower at memotion 2.0 2022 : Ensemble of multi-modal model using attention mechanism in memotion analysis, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [61] R. van der Goot, G. van Noord, Monoise: Modeling noise using a modular normalization system, *Computational Linguistics in the Netherlands Journal* 7 (2017) 129–144.
- [62] A.-M. Bucur, A. Cosma, L. P. Dinu, Sequence-to-sequence lexical normalization with multilingual transformers, in: *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Association for Computational Linguistics, Online, 2021, pp. 473–482. URL: <https://aclanthology.org/2021.wnut-1.53>. doi:10.18653/v1/2021.wnut-1.53.