LT3 at SemEval-2020 Task 8: Multi-Modal Multi-Task Learning for Memotion Analysis

Pranaydeep Singh, Nina Bauwelinck and Els Lefever
LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
pranaydeeps@gmail.com, nina.bauwelinck, els.lefever@ugent.be

Abstract

Internet memes have become a very popular mode of expression on social media networks today. Their multi-modal nature, caused by a mixture of text and image, makes them a very challenging research object for automatic analysis. In this paper, we describe our contribution to the SemEval-2020 Memotion Analysis Task. We propose a Multi-Modal Multi-Task learning system, which incorporates "memebeddings", viz. joint text and vision features, to learn and optimize for all three Memotion subtasks simultaneously. The experimental results show that the proposed system constantly outperforms the competition's baseline, and the system setup with continual learning (where tasks are trained sequentially) obtains the best classification F1-scores.

1 Introduction

While internet memes initially seemed to be restricted to users of specific communities online, memes have rapidly spread among the wider user base, having now claimed its status among emojis and reaction gifs as a common mode of expression on a variety of online platforms. Their user base is now so varied, it ranges from youngsters, boomers to marketers. Despite the lack of consensus on the exact meaning of the original term "meme", as coined by Dawkins (1989), we follow the growing consensus in Communications research in employing the specified definition of internet memes, as "amateur media artifacts, extensively remixed and recirculated by different participants on social media networks" (Milner, 2012).

Internet memes can take on many different roles in the online communicative sphere. They may be used for entertainment purposes. They have also been described as a form of visual rhetoric (Huntington, 2013), functioning as persuasive devices, all the while disguising their message under a layer of humour (Shifman, 2013). Memes also play a significant role in the practice of online trolling, in which they are often the preferred mode of expression because of their potential for spreading provocative and attention-grabbing humour (Leaver, 2013). They have been described both as speech acts (Grundlingh, 2018) and performative acts, involving a conscious decision to either support or reject an ongoing social discourse (Gal et al., 2016). In their capacity as speech acts, the range of functions the memes can perform, extend to the range of functions of any other type of speech act, such as the use of memes to question something (Grundlingh, 2018).

In terms of their form, up to 13 different types of memes have been identified: ranging from simple text-based memes to quotes, rage comics and drawings (Milner, 2012). In their most recognizable form, memes often follow specific image macros (Milner, 2012), or templates with a recognizable image and a text-overlay which changes in individual occurrences of the meme. This multimodal aspect of memes causes meaning to be created on various levels: within each mode and in the interaction between them (Jewitt, 2013). Recent research focusing on automatic genre classification of (mostly political) memes (Theisen et al., 2020; Beskow et al., 2020) will help facilitate data collection, an important aim given the current lack of gold standard datasets of memes.

The multimodal aspect has caused research on the automatic detection of internet memes to be divided in two separate fields: NLP research on memes tends to focus only on the textual aspect, whereas the

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1155–1162
Barcelona, Spain (Online), December 12, 2020.

field of Computer Vision primarily takes into account the visual aspect. The SemEval 2020 Memotion task (Sharma et al., 2020) aims to introduce the novel task of emotion detection of internet memes to the research fields of NLP and Computer Vision. The task aims to answer the need for broadening the focus of the study of social media data to a more hybrid perspective (Highfield and Leaver, 2016) and to bridge the gap between this separation of modalities.

In this paper, we describe our contribution to the SemEval-2020 Memotion Analysis Task. We propose a Multi-Modal Multi-Task learning system, which incorporates "memebeddings", viz. joint text and vision features, to learn and optimize for all Memotion subtasks simultaneously.

2 Related Research

The wide variety of communicative functions has made memes a topic of interest for various research fields. The cohesive effect of internet memes on communities has been studied in sociolinguistics (Procházka, 2016) as well as their potential to generate awareness for social and political issues (Phillips and Milner, 2017). The impact of certain types of humour on the potential of a meme going viral is of particular interest within Communication Studies (Taecharungroj and Nueangjamnong, 2014; Mina, 2019). Internet memes have also been a subject of study in the field of computational creativity with attempts to automate the meme generation process (Peirson et al., 2018; Oliveira et al., 2016) and in the field of information retrieval which is interested in the possibility of personalized searches for memes (Milo et al., 2019).

The aspect of multimodality makes the internet meme an especially challenging research object for automatic detection. Nevertheless, internet memes play a significant role in users' online expression and thus constitute a wealth of possible information on uncovering user sentiment. The automatic detection of sentiment expressed in memes may help advance research into the viral spread of internet phenomena, specifically regarding the principle of emotional contagion (Guadagno et al., 2013). Despite the current lack of research in automatic sentiment detection of internet memes, an exception being French (2017), who researched the extraction of its inherent sentiment by linking the memes to the surrounding user comments, the focus of the emerging sentiment detection research promisingly lies in multimodal classification (Verma et al., 2020). While great efforts have been made for the sentiment classification on textual data (Joshi et al., 2017b), recent attempts to incorporate image-based features from the field of Computer Vision like OCR and face recognition have uncovered the need for improvements in the text analysis of memes due to their short, complex nature (Verma et al., 2020). Sentiment classification has been applied to image data for the purposes of automatic tag predictions for images uploaded on social media (Gajarla and Gupta, 2015), which in turn will help optimize image search algorithms by providing a large collection of tagged image data.

The importance of joining together the insights gained in the study of the two modalities of memes becomes evident when we consider the complexity such multimodal communication modes add to tasks such as the automatic detection of offensive discourse online (Williams et al., 2016; Lee et al., 2018) when compared to offensive language detection in textual data (Zampieri et al., 2019). The SemEval Memotion task not only aims to contribute to the field of automatic detection of sentiment in internet memes, but also to the extraction of more fine-grained information such as sarcasm, humour and offensiveness. Distinguishing these three types of humour is not an easy task, since even the type of humour typically found in non-offensive memes, tends to skirt the boundaries of what is acceptable. It is in this often politically incorrect capacity that their cohesive potential lies (Procházka, 2016). Due to the ambiguous nature of the categories to be identified, the task extends to the detection of the degree (not, slightly, mildly, very) to which the humour type identified is present in the meme. As the meme generation experiments of Oliveira et al. (2016) showed, the perceived humorous nature of memes is often based on the macros used and the meaning these already carry. Since these macros are multimodal, the automatic detection of the meaning they carry must rely on a combination of text-and image-based features. The characteristics of sarcasm forming the basis of the features used in automatic detection methods, such as incongruity, shared knowledge, plausibility and ridicule (Joshi et al., 2017a), originate from traditional theories of humour, such as Raskin's (1989) three sources of humour: incongruity, arousal-safety and disparagement. Sarcasm detection on data that is not purely textual, has only been performed on typographic memes

(memes consisting only of text, but often formatted with a variety of possible fonts) using a Multi Layer perceptron, resulting in an accuracy score of 88% (Kumar and Garg, 2019). While the detection of textual sarcasm has achieved some important advances, among which the use of semi-supervised pattern extraction, lexico-semantic knowledge bases and data-driven methods to identify implicit sentiment (Joshi et al., 2017a; Van Hee et al., 2018), the detection of sarcasm in multimodal environments will need to take into account new features for sarcasm as produced by the image and the interaction between text and image.

The goal of the Memotion Analysis task is to investigate methods suited to the detection of finegrained information in memes, such as the type of humor or offense present. The purposes of the automatic detection of sentiment and humour types may aid the general aim of detection tasks on social media data, namely gaining a better understanding of online communities, but may also help gain important insights on the link between popular ("viral") memes and types of humour they represent to help inform communication strategies online.

3 System Architecture

This section describes the general system architecture we designed for the SemEval 2020 Memotion Analysis Task, which comprises the following subtasks:

- Task A - Sentiment Classification: Given an Internet meme, the first task is to classify it as a positive, negative or neutral;
- Task B - Humor Classification: Given an Internet meme, the system has to identify the type of humor expressed. The categories are sarcastic, humorous, offensive and motivation meme. A meme can have more than one category label;
- Task C - Scales of Semantic Classes: The third task consists in quantifying the extent to which a particular effect is being expressed.

We propose a unified Multi-Modal Multi-Task System that learns meaningful representations of memes. Independent networks for each task, while being better at encoding information about the particular task, do not capture a lot of context about the memes outside of the task. A transformer, for example, which is independently trained to predict sarcasm in memes, while excelling at the task, will not encode anything meaningful about memes in the context of other values like humour and motivation. To address this issue, we propose a joint Multi-Task Network for memes which learns embeddings and optimizes for all three tasks simultaneously. For training the system, these three tasks were further divided into the following five sub-tasks:

1. Sentiment: Predicting the sentiment (Positive, Neutral or Negative)
2. Humour: Predicting the presence and degree of Humour (Not funny, Funny, Very funny, Hilarious)
3. Sarcasm: Predicting the presence and degree of Humour (Not sarcastic, General, Twisted, Very twisted)
4. Predicting the presence and degree of Offensive Content (Not offensive, Slightly offensive, Very offensive, Hateful)
5. Predicting the presence of Motivation (Not motivational, Motivational)

Figure 1 shows the global architecture of our Multi-Modal Multi-Task System. First, the memes are encoded with the help of BERT (Devlin et al., 2019) and visual features oriented at understanding the contents of the image, other than the text (See Section 4). After obtaining these "memebedding" encodings, viz. the joint embeddings from text and image, this information is passed along to the 5 heads for classification for each sub-task. Cross-Entropy Loss is computed for each head independently and the weighted average of the combined loss is jointly optimized with the Adam optimizer. The following section describes the encoding and featurization in more detail.

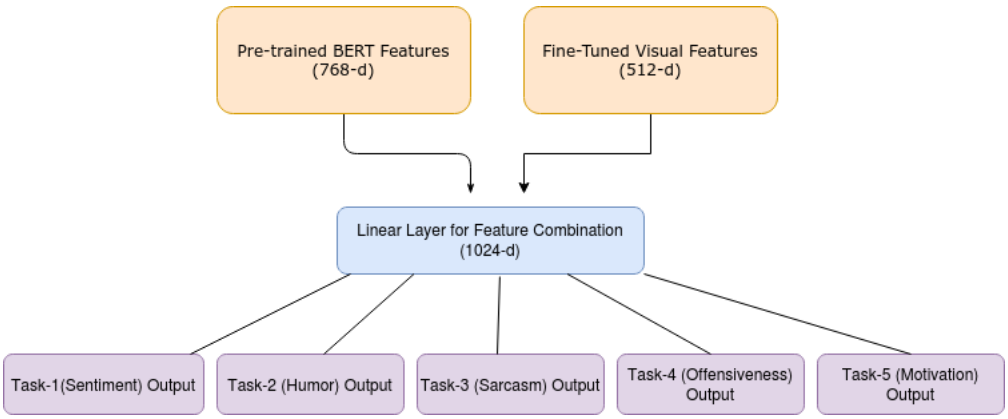

Figure 1: Architecture of the Multi-Modal Multi-Task Learning Setup

4 Experimental Setup

The encoding step derives information from 2 different sources, a pre-trained BERT Masked LM, and a visual feature extractor. Table 1 summarizes the features used.

Feature	Embedding Size	Source
BERT Contextual Embeddings	768	Pre-trained BERT-base-uncased
ResNet-18 Feature Extractor	512	Fine-tuned on Sub-reddit data

Table 1: Overview of the "Memebedding" features used for training

4.1 BERT Features for Text

BERT, the Bidirectional Encoder Representations for Transformers (Devlin et al., 2019), leverages the Bidirectional Transformer for Masked Language Modelling. It demonstrates that a language model which is bidirectionally trained has a deeper, more salient understanding of language than the previous uni-directional language models. Fine-tuning BERT has been leveraged in many downstream tasks and is the state-of-the-art for 11 such tasks like GLUE (Wang et al., 2018), SQuAD (Rajpurkar et al., 2016) and MultiNLI (Williams et al., 2017). We use the standard base-uncased BERT pre-trained model available from the Hugging Face Transformers package[1], to encode the meme text into a 768-dimensional vector.

4.2 Visual Feature Extractor

While BERT features capture essentially everything about the text of the meme in a broader sense, the visual features aim to add some context from the world of memes and information based on the image itself. To this end, Reddit features were obtained from a classifier trained on reddit memes. We collected around 3980 memes from Reddit, from 8 different subreddits, such as /r/MemeEconomy, /r/dankememes, /r/GetMotivated, etc. We then proceeded to fine-tune a pre-trained ResNet-18 (He et al., 2015) feature extractor to predict the subreddit a meme was picked from. Using a large Vision model, fine-tuning it on domain specific data and using it as a feature extractor is fairly common practice in application areas like medical diagnosis (Habibzadeh et al., 2018), self-driving cars (Jung et al., 2017) and person identification (Lu et al., 2018). The initial pre-training on large datasets encodes fundamental concepts into the model and domain-specific data, memes in this case, can be easily understood better and faster by fine-tuning with significantly lesser samples. We believed this would capture salient features of a meme since the sub-reddit a meme belongs to represents the broader category of the meme. The ResNet-18 Classifier was trained with Cross-Entropy loss and optimized with SGD. We followed standard fine-tuning practices where all the CNN layers were taken from a network pre-trained on ImageNet, the layers were frozen and only the final linear aggregator was discarded and retrained with the reddit data.

[1] https://huggingface.co/transformers/

	Macro F1			Micro F1		
	Task A	Task B	Task C	Task A	Task B	Task C
Task Baseline	0.2176	0.5002	0.3008	0.3077	0.5686	0.3328
Multi-modal No MTL	**0.3220**	0.4623	0.2852	0.5175	**0.6817**	0.4539
Multi-modal MTL Average Loss	0.2447	0.4331	0.2291	**0.5915**	0.6529	0.4640
Multi-modal MTL Weighted Average Loss	0.2477	0.4485	0.2499	**0.5915**	0.6617	**0.4652**
Multi-modal MTL Continual Learning	0.2771	**0.5077**	**0.5069**	0.5750	0.6317	0.4227

Table 2: Macro and Micro Averaged F1 for all the 3 Sub-Tasks for the competition

We combined the features from BERT and the Visual Feature Extractor with a Linear Layer to produce a 1024-dimensional "memebedding" encoding for classification for the 5 sub-tasks. Although this technique, viz. combining the embeddings from multiple modalities using a linear layer as an aggregator, is very primitive, and not the state-of-the-art in feature aggregation, we opted for it because of its simplicity and lack of complexity in training. The combined "memebeddings" are optimized for each task individually using standard Cross-Entropy Loss. The classifier used for each task is again a simple single layer linear neural network.

5 Results and Analysis

We conducted various experiments to understand the working of the Multi-Task Learning (MTL) Model better. Table 2 summarizes the results of these different experiments. A first experiment consisted in training the 5 tasks independently ("Multi-modal No MTL"), to see how this compares with the MTL system. We used Cross-Entropy Loss for every task, but since the loss needs to be combined for joint training, we used a weighted mean of the losses (Kendall et al., 2017)("Multi-modal MTL Weighted Average Loss"). Another setup consisted in experimenting with the mean loss ("Multi-modal MTL Average Loss"), in contrast to the weighted loss. We found that higher weights for the Sarcasm and Offensiveness task, average weights for the Humor and Motivation tasks, and very low weights for the Sentiment task, work better towards obtaining a model that performs well on all 3 tasks. This is understandable since the Sentiment task is less complicated, while the detection of Sarcasm and Offensiveness can be very challenging. Finally, we experimented with continual learning (Ribeiro et al., 2019), where instead of joint training, the tasks were trained sequentially, in different orders ("Multi-modal MTL Continual Learning"). We found that training Motivation and Humor first, followed by Sarcasm, then Offensiveness and finally Sentiment, gave the best results.

The "Multi-modal MTL Average Loss" flavor was our official submission for the SemEval Task. "Multi-modal MTL Weighted Average Loss" and "Multi-modal MTL Continual Learning" were evaluated afterwards on the held-out test data.

6 Conclusion

In this research, we have focused on constructing a unified network for meme analysis that incorporates both textual and visual information. The system has shown to perform well on multiple tasks at the same time. In future research, it would be worth exploring the joint "memebeddings" it creates by combining the visual and textual features, to understand what the network is able to encode. The visual features are tuned on memes and thus tailored for this domain. The BERT text embeddings, however, are generic and could be better tailored for this context as well. Another interesting direction for future research would be to add further sub-tasks to the Multi-Task Training. A task to predict the "premise" (the underlying argument or concept) of a meme would be highly interesting as it would help the network further understand the general concept behind a meme.

References

David M Beskow, Sumeet Kumar, and Kathleen M Carley. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, 57(2):102170.

Richard Dawkins. 1989. *The selfish gene*. Oxford university press.

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.

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

Vasavi Gajarla and Aditi Gupta. 2015. Emotion detection and sentiment analysis of images. *Georgia Institute of Technology*.

Noam Gal, Limor Shifman, and Zohar Kampf. 2016. "it gets better": Internet memes and the construction of collective identity. *New media & society*, 18(8):1698–1714.

Lezandra Grundlingh. 2018. Memes as speech acts. *Social Semiotics*, 28(2):147–168.

Rosanna E Guadagno, Daniel M Rempala, Shannon Murphy, and Bradley M Okdie. 2013. What makes a video go viral? an analysis of emotional contagion and internet memes. *Computers in Human Behavior*, 29(6):2312–2319.

Mehdi Habibzadeh, Mahboobeh Jannesari, Zahra Rezaei, Hossein Baharvand, and Mehdi Totonchi. 2018. Automatic white blood cell classification using pre-trained deep learning models: Resnet and inception. In *Tenth International Conference on Machine Vision (ICMV 2017)*, volume 10696, page 1069612. International Society for Optics and Photonics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Tim Highfield and Tama Leaver. 2016. Instagrammatics and digital methods: Studying visual social media, from selfies and gifs to memes and emoji. *Communication Research and Practice*, 2(1):47–62.

Heidi E Huntington. 2013. Subversive memes: Internet memes as a form of visual rhetoric. *AoIR Selected Papers of Internet Research*, 3.

Carey Jewitt. 2013. Multimodal methods for researching digital technologies. *The SAGE handbook of digital technology research*, pages 250–265.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017a. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Megha Joshi, Purvi Prajapati, Ayesha Shaikh, and Vishwa Vala. 2017b. A survey on sentiment analysis. *International Journal of Computer Applications*, 163(6):34–38, Apr.

Heechul Jung, Min-Kook Choi, Jihun Jung, Jin-Hee Lee, Soon Kwon, and Woo Young Jung. 2017. Resnet-based vehicle classification and localization in traffic surveillance systems. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 61–67.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115.

Akshi Kumar and Geetanjali Garg. 2019. Sarc-m: Sarcasm detection in typo-graphic memes. *Available at SSRN 3384025*.

Tama Leaver. 2013. Fcj-163 olympic trolls: Mainstream memes and digital discord? *The Fibreculture Journal*, (22 2013: Trolls and The Negative Space of the Internet).

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. *arXiv preprint arXiv:1808.10245*.

Ze Lu, Xudong Jiang, and Alex Kot. 2018. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(4):526–530.

Ryan M Milner. 2012. *The world made meme: Discourse and identity in participatory media*. Ph.D. thesis, University of Kansas.

Tova Milo, Amit Somech, and Brit Youngmann. 2019. Simmeme: A search engine for internet memes. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 974–985. IEEE.

An Xiao Mina. 2019. *Memes to Movements: How the World's Most Viral Media is Changing Social Protest and Power*. Beacon Press.

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*.

V Peirson, L Abel, and E Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*.

Whitney Phillips and Ryan M Milner. 2017. Decoding memes: Barthes' punctum, feminist standpoint theory, and the political significance of# yesallwomen. In *Entertainment Values*, pages 195–211. Springer.

Ondřej Procházka. 2016. Cohesive aspects of humor in internet memes on facebook: A multimodal sociolinguistic analysis. *Ostrava Journal of English Philology*, 8(1):7–38.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Victor Raskin. 1989. Semantic mechanisms of humor dordrecht holland: Reidel. *Reagan Ronald*.

João Ribeiro, Francisco S. Melo, and Joao Dias. 2019. Multi-task learning and catastrophic forgetting in continual reinforcement learning. *ArXiv*, abs/1909.10008.

Chhavi Sharma, William Paka, Scott, Deepesh Bhageria, Amitava Das, Soujanya Poria, Tanmoy Chakraborty, and Björn Gambäck. 2020. Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, 18(3):362–377.

Viriya Taecharungroj and Pitchganut Nueangjamnong. 2014. The effect of humour on virality: The study of internet memes on social media. In *7th International Forum on Public Relations and Advertising Media Impacts on Culture and Social Communication. Bangkok, August*.

William Theisen, Joel Brogan, Pamela Bilo Thomas, Daniel Moreira, Pascal Phoa, Tim Weninger, and Walter Scheirer. 2020. Automatic discovery of political meme genres with diverse appearances. *arXiv preprint arXiv:2001.06122*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. We usually don't like going to the dentist: Using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832, December.

Devika Verma, Rohit Chandiramani, Pranay Jain, Chinmay Chaudhari, Anmol Khandelwal, Krishnanjan Bhattacharjee, S ShivaKarthik, Swathi Mithran, Swati Mehta, and Ajai Kumar. 2020. Sentiment extraction from image-based memes using natural language processing and machine learning. In *ICT Analysis and Applications*, pages 285–293. Springer.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of internet memes. *Computers in Human Behavior*, 63:424–432.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

—