Combining Humor and Sarcasm for Improving Political Parody Detection

Xiao Ao[α] **Danae Sánchez Villegas**[α] **Daniel Preoţiuc-Pietro**[β] **Nikolaos Aletras**[α]
[α] Computer Science Department, University of Sheffield, UK
[β] Bloomberg
{xao3,dsanchezvillegas1,n.aletras}@sheffield.ac.uk
dpreotiucpie@bloomberg.net

Abstract

Parody is a figurative device used for mimicking entities for comedic or critical purposes. Parody is intentionally humorous and often involves sarcasm. This paper explores jointly modelling these figurative tropes with the goal of improving performance of political parody detection in tweets. To this end, we present a *multi-encoder* model that combines three parallel encoders to enrich parody-specific representations with humor and sarcasm information. Experiments on a publicly available data set of political parody tweets demonstrate that our approach outperforms previous state-of-the-art methods.[1]

Twitter Handle	@Queen_UK
Parody tweet	Boris Johnson on the phone. Very smug that #myspace hasn't gone down. Says he's always backed winning technologies #whatsappdown #instagramdown #FacebookIsDown

Table 1: Example of a *parody* tweet[3] by the Twitter handle @Queen_UK. Humor and sarcasm are expressed simultaneously.

1 Introduction

Parody is a figurative device which imitates entities such as politicians and celebrities by copying their particular style or a situation where the entity was involved (Rose, 1993). It is an intrinsic part of social media as a relatively new comedic form (Vis, 2013). A very popular type of parody is political parody, which is used to express political opposition and civic engagement (Davis et al., 2018).

One of the hallmarks of parody expression is the deployment of other figurative devices, such as humor and sarcasm, as emphasized on studies of parody in linguistics (Haiman et al., 1998; Highfield, 2016). For example, in Table 1 the text expresses sarcasm about Myspace[2] being a 'winning technology', while mocking the fact that three more popular social media sites were unavailable. This example also highlights the similarities between parody and real tweets, which may pose issues to misinformation classification systems (Mu and Aletras, 2020).

These figurative devices have so far been studied in isolation to parody. Previous work on modeling humor in computational linguistics has focused on identifying jokes, i.e., short comedic passages that end with a hilarious line (Hetzron, 1991), based on linguistic features (Taylor and Mazlack, 2004; Purandare and Litman, 2006; Kiddon and Brun, 2011) and deep learning techniques (Chen and Soo, 2018; Weller and Seppi, 2019; Annamoradnejad and Zoghi, 2020). Similarly, computational approaches for modeling sarcasm (i.e., a form of verbal irony used to mock or convey content) in texts have been explored (Davidov et al., 2010; González-Ibáñez et al., 2011; Liebrecht et al., 2013; Rajadesingan et al., 2015; Ghosh et al., 2020, 2021), including multi-modal utterances, i.e. texts, images, and videos (Cai et al., 2019; Castro et al., 2019; Oprea and Magdy, 2020). Recently, parody has been studied with natural language processing (NLP) methods by Maronikolakis et al. (2020) who introduced a data set of political parody accounts. Their method for automatic recognition of posts shared by political parody accounts on Twitter is solely based on vanilla transformer models.

In this paper, we hypothesize that humor and sarcasm information could guide parody specific text encoders towards detecting nuances of figu-

[1]Code is available here https://github.com/iamoscar1/Multi_Encoder_Model_for_Political_Parody_Prediction

[2]https://myspace.com

[3]https://twitter.com/Queen_UK/status/1445103605355323393?t=FGMNsMVFF_G2tABYxFmkFw&s=07

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1800 - 1807
July 10-15, 2022

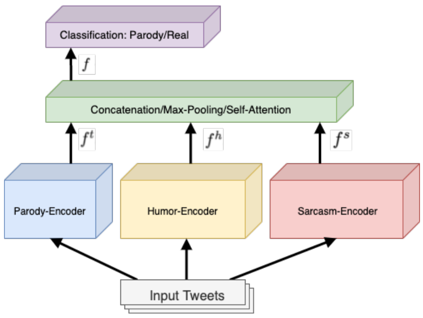

Figure 1: The structure of our *multi-encoder* model for combining humor and sarcasm information for political parody prediction.

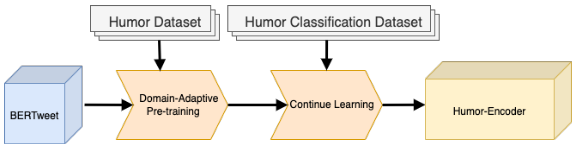

Figure 2: *Humor Encoder*.

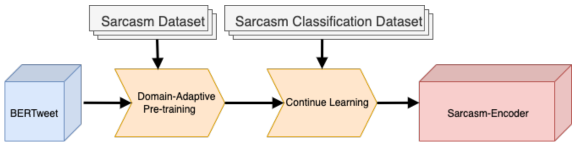

Figure 3: *Sarcasm Encoder*.

rative language. For this purpose, we propose a *multi-encoder* model (§2) consisting of three parallel encoders that are subsequently fused for parody classification. The first encoder learns parody specific information subsequently enhanced using the representations learned by a humor and sarcasm encoder respectively.

Our contributions are: (1) new state-of-the-art results on political parody detection in Twitter, consistently improving predictive performance over previous work by Maronikolakis et al. (2020); and (2) insights on the limitations of neural models in capturing various linguistic characteristics of parody from extensive qualitative and quantitative analyses.

2 Multi-Encoder Model for Political Parody Prediction

Maronikolakis et al. (2020) define political parody prediction as a binary classification task where a social media post T, consisting of a sequence of tokens $T = \{t_1, ..., t_n\}$, is classified as real or parody. Real posts have been authored by actual politicians (e.g., `realDonaldTrump`) while parody posts come from their corresponding parody accounts (e.g., `realDonaldTrFan`).

Parody tends to express complex tangled semantics of both humor and sarcasm simultaneously (Haiman et al., 1998; Highfield, 2016). To better exploit this characteristic of parody, we propose a *multi-encoder* model that consists of three parallel encoders, a feature-fusion layer and a parody classification layer depicted in Fig.1.[4].

[4]Early experiments with multi-task learning did not result in improved performance. The results of these experiments can be found in Appendix A.

2.1 Text Encoders

Parody As a task-specific parody encoder, we use the vanilla pretrained BERTweet (Nguyen et al., 2020), a BERT (Devlin et al., 2019) based model pre-trained on a corpus of English Tweets and fine-tuned on the parody data set (§3.1).

Humor To capture humor specific characteristics in social media text, we use the data set introduced by Annamoradnejad and Zoghi (2020) which contains humorous and non-humorous short texts collected from Reddit and Huffington Post. First, we adapt BERTweet using domain-adaptive pre-training (Sun et al., 2020a; Gururangan et al., 2020) on 10,000 randomly selected humor-only short texts with masked language modeling. Subsequently, we use a continual learning strategy (Li and Hoiem, 2018; Sun et al., 2020b) to gradually learn humor-specific properties by further fine-tuning BERTweet on a humor classification task (i.e., predicting whether a text is humorous or not) by using 40,000 randomly selected humorous and non-humorous short texts from the humor corpus described above (see Figure 2).

Sarcasm Similar to humor, we extract sarcasm-related semantic information from a post T by using sarcasm annotated data sets from Oprea and Magdy (2020) and Rajadesingan et al. (2015). The first data set consists of 777 and 3,707 sarcasm and non-sarcasm posts from Twitter and the second data set consists of 9,104 sarcasm and more than 90,000 non-sarcasm posts from Twitter. We first perform domain-adaptive pre-training of BERTweet on all sarcastic posts with masked language modeling. Then, we fine-tune the model on a sarcasm classification task, similar to the humor encoder (see Figure 3). For the fine-tuning step, we use the 9,881 sarcastic tweets and 10,000 randomly sampled non-sarcasm tweets from the two data sets (i.e., 3,707 from the first and 6,293 from the second).

We compute parody f^t, humor f^h, and sarcasm f^s representations by extracting the 'classification' [CLS] token from each encoder respectively, where $f \in \mathbf{R}^{768}$.

2.2 Combining Encoders

We explore three approaches to combine f^t, f^h, and f^s representations.

Concatenation First, the three text representations are simply concatenated to form a combined representation $f \in \mathbf{R}^{768\times3}$.

Self-Attention We also use a 4-head self-attention[5] mechanism (Vaswani et al., 2017) on f^t, f^h, f^s. The goal is to find correlations between representations and learn the contribution of each encoder in the final representation.

Max-Pooling Finally, we perform a max-pooling operation on each dimension of f^t, f^h, f^s to obtain a representation $f \in \mathbf{R}^{768}$. The aim is to use the most dominant features learned by each encoder.

2.3 Classification

Finally, we pass the combined representation f to a classification layer with a sigmoid activation function for predicting whether a post is a parody or not. Three encoders are fine-tuned simultaneously on the parody data set (§3.1).[6]

3 Experimental Setup

3.1 Data

We use the data set introduced by Maronikolakis et al. (2020) which contains 131,666 tweets written in English, with 65,956 tweets from political parody accounts and 65,710 tweets posted by real politician accounts. The data set is publicly available[7] and allows us to compare our results to state-of-the-art parody detection methods.

We use the three data splits provided: (i) *Person Split*, each split (train, dev, test) contains tweets from different real – *parody* account pairs; (ii) *Gender Split*, two different splits based on the gender of the politicians (i.e., female accounts in train/dev and male in test, and male accounts in train/dev and female in test); *Location Split*, data is split according to the location of the politicians in three groups (US, UK, Rest of the World or RoW). Each group is assigned to the test set and the other two groups to the train and dev sets.

3.2 Baselines

We compare our *multi-encoder* models with transformers for parody detection (Maronikolakis et al., 2020): **BERT** (Devlin et al., 2019) and **RoBERTa** (Liu et al., 2019). Also, we compare our models to **BERTweet** (Nguyen et al., 2020).

3.3 Implementation details

Humor Encoder For adaptive pre-training, the batch-size is set to 16 and the number of training epochs is set to 3 with a learning rate of $2e^{-5}$. For humor classification, we use batch size of 128 and the number of epochs is set to 2 with a learning rate of $3e^{-5}$.

Sarcasm Encoder We pretrain using a batch-size of 16 over 5 epochs with a learning rate of $2e^{-5}$. For fine-tuning on a sarcasm classification task, we use the $9,881$ sarcasm tweets and $10,000$ randomly sampled non-sarcasm tweets from the two data sets (i.e., $3,707$ from the first and $6,293$ from the second) using the same hyperparameters to the humor-specific encoder.

Multi-encoder For the complete *multi-encoder* model, we use a batch size of 128 and the learning rate is set to $2e^{-5}$. The entire model is fine-tuned for 2 epochs.

3.4 Evaluation

We evaluate the performance of all models using F1 score as Maronikolakis et al. (2020). Results are obtained over 3 runs using different random seeds reporting average and standard deviation.

4 Results

4.1 Predictive Performance

Table 2 shows the results for parody detection on the *Person Split*. We observe that *BERTweet* has the best performance (F1: 90.72) among transformer-based models (*BERT*, *RoBERTa*, *BERTweet*), outperforming previous state-of-the-art by Maronikolakis et al. (2020). This is due to the fact that *BERTweet* has been specifically pre-trained on

[5]Early experimentation with larger attention heads did not improve results in the dev set.

[6]Early experimentation with humor and sarcasm encoders frozen during the fine-tuning process did not show any performance improvement.

[7]https://archive.org/details/parody_data_acl20

Person	
Model	**F1**
Single-Encoder	
BERT**	87.65 ± 0.18
RoBERTa**	89.66 ± 0.33
BERTweet	90.72 ± 0.31
Multi-encoder (Ours)	
Concatenation	88.99 ± 0.17
Self-Attention	**91.19 ± 0.31**
Max-Pooling	91.05 ± 0.30

Table 2: F1-scores for parody detection on the *Person Split*. ** Results from Maronikolakis et al. (2020). Best results are in bold.

Gender		
Model	**M→F**	**F→M**
Single-Encoder		
BERT**	85.85 ± 0.28	84.40 ± 0.35
RoBERTa**	87.11 ± 0.31	84.87 ± 0.38
BERTweet	88.01 ± 0.29	85.57 ± 0.27
Multi-encoder (Ours)		
Concatenation	86.84 ± 0.15	84.21 ± 0.22
Self-Attention	**89.97 ± 0.34**	**88.56 ± 0.39**
Max-Pooling	88.39 ± 0.27	86.89 ± 0.56

Table 3: F1-scores on the *Gender Split*. ** Results from Maronikolakis et al. (2020). Best results are in bold.

Location			
Model	**UK+US → RoW**	**RoW+US → UK**	**RoW+UK → US**
Single-Encoder			
BERT**	86.69 ± 0.45	83.78 ± 0.19	83.12 ± 0.60
RoBERTa**	87.70 ± 0.45	85.10 ± 0.27	85.99 ± 0.61
BERTweet	88.21 ± 0.26	87.85 ± 0.24	**87.18 ± 0.41**
Multi-encoder (Ours)			
Concatenation	85.41 ± 0.26	84.74 ± 0.20	83.62 ± 0.35
Self-Attention	**88.37 ± 0.28**	**87.91 ± 0.19**	87.16 ± 0.37
Max-Pooling	88.25 ± 0.39	86.49 ± 0.33	86.54 ± 0.41

Table 4: F1-scores on the *Location Split*. ** Results from Maronikolakis et al. (2020). Best results are in bold.

Twitter text. Similar behavior is observed on the *Gender* and *Location* splits (see Table 3 and 4 respectively).

Our proposed *multi-encoder* achieves the best performance when using *Self-Attention* to combine the three parallel encoders (F1: 91.19; 89.97, 88.56; 88.37, 87.91, 87.16; for *Person*, *Gender*, and *Location* splits respectively). Moreover, it outperforms the best single-encoder model *BERTweet* in the majority of cases which corroborates that parody detection benefits from combining general contextual representations with humor and sarcasm specific information, as humor and sarcasm are important characteristics of parody (Haiman et al., 1998; Highfield, 2016). On the other hand, simply concatenating the three parallel encoders degrades the performance across different splits (*Person*: 88.99; *Gender*: 86.84, 84.21 *Location*: 85.41, 84.74, 83.62). This happens because the concatenation operation treats the three encoders as equally important. While humor and sarcasm are related to parody, they may not necessarily have the same relevance as indicators of parody.

Our best performing model (*Self-Attention*) outperforms the vanilla *BERTweet* by 3 F1 points when trained on female accounts and by almost 2 F1 points when trained on male accounts. We speculate that the additional linguistic information from the two encoders (i.e., sarcasm and humor) is more beneficial in low data settings. The number of female politicians is considerably smaller than males in the data set (see Maronikolakis et al. (2020) for more details).

4.2 Ablation Study

We also examine the effect of combining parody-specific representations with humor and sarcasm information by running an ablation study. We compare performance of four models: using parody representations only (P), and combining parody representations with humor (P+H), or sarcasm (P+S) information, as well as with both (P+S+H). The results of this analysis are depicted in Tables 5, 6 and 7. We observe that both sarcasm and humor contribute to the performance gain, but using both is more beneficial. Modelling sarcasm leads to more gains than humor and this could be attributed to the characteristics of the parody corpus, namely that it focuses primarily on the political domain, which have a high sarcastic component (Anderson and Huntington, 2017).

5 Error Analysis

Finally, we perform an error analysis to examine the behavior and limitations of our best-performing model (*multi-encoder* with Self-Attention).

The next two examples correspond to real tweets that were misclassified as parody:

(1) *Congratulations, <mention>! <url>.*

(2) *It's a shame that Boris isn't here answering questions from the public this evening.*

We speculate that the model misclassified these tweets as parody because they contain terms that

Person	
Model	**F1**
Single-Encoder	
BERTweet (P)	90.72 ± 0.31
Multi-encoder (Ours)	
Concatenation (P+S+H)	88.99 ± 0.17
Concatenation (P+S)	90.51 ± 0.26
Concatenation (P+H)	89.98 ± 0.23
Self-Attention (P+S+H)	**91.19 ± 0.31**
Self-Attention (P+S)	91.14 ± 0.40
Self-Attention (P+H)	90.98 ± 0.36
Max-Pooling (P+S+H)	91.05 ± 0.30
Max-Pooling (P+S)	91.06 ± 0.39
Max-Pooling (P+H)	90.78 ± 0.42

Table 5: F1-scores for parody detection on the *Person Split* with various settings: parody (P) representations only, and combining parody representations with humor (P+H), or sarcasm (P+S) information, as well as with both (P+S+H). Best results are in bold.

Gender		
Model	**M→F**	**F→M**
Single-Encoder		
BERTweet (P)	88.01 ± 0.29	85.57 ± 0.27
Multi-encoder (Ours)		
Concatenation (P+S+H)	86.84 ± 0.15	84.21 ± 0.22
Concatenation (P+S)	86.93 ± 0.40	83.70 ± 0.41
Concatenation (P+H)	86.58 ± 0.31	83.34 ± 0.38
Self-Attention (P+S+H)	**89.97 ± 0.34**	**88.56 ± 0.39**
Self-Attention (P+S)	89.49 ± 0.37	88.23 ± 0.44
Self-Attention (P+H)	88.71 ± 0.42	87.62 ± 0.50
Max-Pooling (P+S+H)	88.39 ± 0.27	86.89 ± 0.56
Max-Pooling (P+S)	88.36 ± 0.46	86.55 ± 0.49
Max-Pooling (P+H)	88.14 ± 0.52	86.53 ± 0.53

Table 6: F1-scores for parody detection on the *Gender Split* with various settings: parody (P) representations only, and combining parody representations with humor (P+H), or sarcasm (P+S) information, as well as with both (P+S+H). Best results are in bold.

are related to sarcastic short texts such as user mentions, punctuation marks (*!*), and negation (*isn't*) (González-Ibáñez et al., 2011; Highfield, 2016).

The following two examples correspond to parody tweets that were misclassified as real:

(3) *Hey America, it's time to use your safe word.*

(4) *I fully support the Digital Singles Market.*

Example (3) is a call-to-action message, while Example (4) is a statement expressing support for a particular subject. These statements are written in a style that is similar to political slogans or campaign speeches (Fowler et al., 2021) that the model fails

Location			
Model	**UK+US → RoW**	**RoW+US → UK**	**RoW+UK → US**
Single-Encoder			
BERTweet (P)	88.21 ± 0.26	87.85 ± 0.24	**87.18 ± 0.41**
Multi-encoder (Ours)			
Concatenation (P+S+H)	85.41 ± 0.26	84.74 ± 0.20	83.62 ± 0.35
Concatenation (P+S)	85.92 ± 0.24	85.67 ± 0.18	84.09 ± 0.39
Concatenation (P+H)	85.39 ± 0.29	85.33 ± 0.26	83.75 ± 0.44
Self-Attention (P+S+H)	**88.37 ± 0.28**	**87.91 ± 0.19**	87.16 ± 0.37
Self-Attention (P+S)	88.24 ± 0.33	87.88 ± 0.23	86.47 ± 0.32
Self-Attention (P+H)	88.13 ± 0.35	87.05 ± 0.28	85.36 ± 0.40
Max-Pooling (P+S+H)	88.25 ± 0.39	86.49 ± 0.33	86.54 ± 0.41
Max-Pooling (P+S)	88.28 ± 0.42	87.83 ± 0.39	86.56 ± 0.36
Max-Pooling (P+H)	88.22 ± 0.52	86.44 ± 0.42	85.96 ± 0.45

Table 7: F1-scores for parody detection on the *Location Split* with various settings: parody (P) representations only, and combining parody representations with humor (P+H), or sarcasm (P+S) information, as well as with both (P+S+H). Best results are in bold.

to recognise. As a result, in addition to humor and sarcasm semantics, the model might be improved by integrating knowledge from the political domain such as from political speeches.

6 Conclusion

In this paper, we studied the impact of jointly modelling figurative devices to improve predictive performance of political parody detection in tweets. Our motivation was based on studies in linguistics which emphasize the humorous and sarcastic components of parody (Haiman et al., 1998; Highfield, 2016). We presented a method that combines parallel encoders to capture parody, humor, and sarcasm specific representations from input sequences, which outperforms previous state-of-the-art proposed by Maronikolakis et al. (2020).

In the future, we plan to combine information from other modalities (e.g., images) for improving parody detection (Sánchez Villegas and Aletras, 2021; Sánchez Villegas et al., 2021).

Acknowledgements

We would like to thank all reviewers for their valuable feedback. DSV is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1.

References

Ashley A Anderson and Heidi E Huntington. 2017. Social media, science, and attack discourse: How twitter discussions of climate change use sarcasm and incivility. *Science Communication*, 39(5):598–620.

Issa Annamoradnejad and Gohar Zoghi. 2020. Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. pages 41–48.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. 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 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Jenny L Davis, Tony P Love, and Gemma Killen. 2018. Seriously funny: The political work of humor on social media. *New Media & Society*, 20(10):3898–3916.

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. Association for Computational Linguistics.

Erika Franklin Fowler, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2021. Political advertising online and offline. *American Political Science Review*, 115(1):130–149.

Debanjan Ghosh, Elena Musi, and Smaranda Muresan. 2020. Interpreting verbal irony: Linguistic strategies and the connection to theType of semantic incongruity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 82–93, New York, New York. Association for Computational Linguistics.

Debanjan Ghosh, Ritvik Shrivastava, and Smaranda Muresan. 2021. "laughing at you or with you": The role of sarcasm in shaping the disagreement space. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1998–2010, Online. Association for Computational Linguistics.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

John Haiman et al. 1998. *Talk is cheap: Sarcasm, alienation, and the evolution of language*. Oxford University Press on Demand.

Robert Hetzron. 1991. On the structure of punchlines.

Tim Highfield. 2016. News via voldemort: Parody accounts in topical discussions on twitter. *New Media Soc.*, 18(9):2028–2045.

Chloé Kiddon and Yuriy Brun. 2011. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 89–94, Portland, Oregon, USA. Association for Computational Linguistics.

Zhizhong Li and Derek Hoiem. 2018. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. Analyzing political parody in social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4373–4384, Online. Association for Computational Linguistics.

Yida Mu and Nikolaos Aletras. 2020. Identifying twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, 6:e325.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S*. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215, Sydney, Australia. Association for Computational Linguistics.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

Margaret A Rose. 1993. *Parody: ancient, modern and post-modern*. Cambridge University Press.

Danae Sánchez Villegas and Nikolaos Aletras. 2021. Point-of-interest type prediction using text and images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7785–7797, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danae Sánchez Villegas, Saeid Mokaram, and Nikolaos Aletras. 2021. Analyzing online political advertisements. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3669–3680, Online. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020a. How to fine-tune bert for text classification? *arXiv:1905.05583*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020b. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

J.M. Taylor and L.J. Mazlack. 2004. Humorous wordplay recognition. 4:3306–3311 vol.4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Farida Vis. 2013. Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 uk riots. *Digit. J.*, 1(1):27–47.

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 3621–3625, Hong Kong, China. Association for Computational Linguistics.

A Multitask-Learning

We also tested applying *multi-task learning* approaches (Caruana, 1993) to use either sarcasm prediction (P+S), humor prediction (P+H) or both (P+S+H) as auxiliary tasks for parody detection. We utilize BERTweet as the share encoder and independent classification layers for parody and humor or sarcasm. Three sets of weights are applied to losses from each independent classification layer and the three layers are stacked. The best results are chosen and depicted in Table 8, Table 9 and Table 10.

Person	
Model	**F1**
Single-Encoder	
BERTweet (P)	**90.72 ± 0.31**
Multi-Task	
P+S+H	87.46 ± 0.18
P+S	89.41 ± 0.31
P+H	87.41 ± 0.38

Table 8: F1-scores for parody detection on the *Person Split* using Multi-task Learning models (P: Parody, S: Sarcasm, H: Humor). Best results are in bold.

Gender		
Model	**M→F**	**F→M**
Single-Encoder		
BERTweet (P)	88.01 ± 0.29	85.57 ± 0.27
Multi-Task		
P+S+H	85.28 ± 0.29	84.10 ± 0.37
P+S	**88.13 ± 0.21**	**86.07 ± 0.44**
P+H	84.53 ± 0.31	86.07 ± 0.47

Table 9: F1-scores on the *Gender Split* using Multi-task Learning models (P: Parody, S: Sarcasm, H: Humor). Best results are in bold.

Location			
Model	**UK+US → RoW**	**RoW+US → UK**	**RoW+UK → US**
Single-Encoder			
BERTweet (P)	**88.21 ± 0.26**	**87.85 ± 0.24**	**87.18 ± 0.41**
Multi-Task			
P+S+H	86.41 ± 0.17	86.23 ± 0.20	85.13 ± 0.29
P+S	87.74 ± 0.36	87.26 ± 0.34	86.67 ± 0.43
P+H	85.54 ± 0.38	84.78 ± 0.47	84.15 ± 0.56

Table 10: F1-scores on the *Location Split* using Multi-task Learning models (P: Parody, S: Sarcasm, H: Humor). Best results are in bold.