

Multi-Task Bidirectional Transformer Representations for Irony Detection

Chiyu Zhang and Muhammad Abdul-Mageed

Natural Language Processing Lab
The University of British Columbia
chiyuzh@mail.ubc.ca, muhammad.mageeed@ubc.ca

Abstract. Supervised deep learning requires large amounts of training data. In the context of the FIRE2019 Arabic irony detection shared task (IDAT@FIRE2019), we show how we mitigate this need by fine-tuning the pre-trained bidirectional encoders from transformers (BERT) on gold data in a multi-task setting. We further improve our models by further pre-training BERT on ‘in-domain’ data, thus alleviating an issue of dialect mismatch in the Google-released BERT model. Our best model acquires 82.4 macro F_1 score, and has the unique advantage of being feature-engineering free (i.e., based exclusively on deep learning).¹

Keywords: irony detection, Arabic, social media, BERT, multi-task learning

1 Introduction

The proliferation of social media has provided a locus for use, and thereby collection, of figurative and creative language data, including irony [18]. According to the Merriam-Webster online dictionary,² *irony* refers to “the use of word to express something other than and especially the opposite of the literal meaning.” A complex, controversial, and intriguing linguistic phenomenon, irony has been studied in disciplines such as linguistics, philosophy, and rhetoric. Irony detection also has implications for several NLP tasks such as sentiment analysis, hate speech detection, fake news detection, etc [18]. Hence, automatic irony detection can potentially improve systems designed for each of these tasks. In this paper, we focus on learning irony. More specifically, we report our work submitted to the FIRE 2019 Arabic irony detection task (IDAT@FIRE2019).³ We focus our energy on an important angle of the problem—the small size of training data.

Deep learning is the most successful under supervised conditions with large amounts of training data (tens-to-hundreds of thousands of examples). For most

¹ Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). FIRE 2019, 12-15 December 2019, Kolkata, India.

² <https://www.merriam-webster.com/dictionary/irony>.

³ <https://www.irit.fr/IDAT2019/>

real-world tasks, we hard to obtain labeled data. Hence, it is highly desirable to eliminate, or at least reduce, dependence on supervision. In NLP, pre-training language models on unlabeled data has emerged as a successful approach for improving model performance. In particular, the pre-trained multilingual **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) [15] was introduced to learn language regularities from unlabeled data. Multi-task learning (MTL) is another approach that helps achieve inductive transfer between various tasks. More specifically, MTL leverages information from one or more source tasks to improve a target task [10, 11]. In this work, we introduce Transformer representations (BERT) in an MTL setting to address the data bottleneck in IDAT@FIRE2019. To show the utility of BERT, we compare to a simpler model with gated recurrent units (GRU) in a single task setting. To identify the utility, or lack thereof, of MTL BERT, we compare to a single task BERT model. For MTL BERT, we train on a number of tasks simultaneously. Tasks we train on are *sentiment analysis*, *gender detection*, *age detection*, *dialect identification*, and *emotion detection*.

Another problem we face is that the BERT model released by Google is trained only on Arabic Wikipedia, which is almost exclusively Modern Standard Arabic (MSA). This introduces a language variety mismatch due to the irony data involving a number of dialects that come from the Twitter domain. To mitigate this issue, we further pre-train BERT on an in-house dialectal Twitter dataset, showing the utility of this measure. To summarize, we make the following contributions:

- In the context of the Arabic irony task, we show how a small-sized labeled data setting can be mitigated by training models in a multi-task learning setup.
- We view different varieties of Arabic as different domains, and hence introduce a simple, yet effective, ‘in-domain’ training measure where we further pre-train BERT on a dataset closer to task domain (in that it involves dialectal tweet data).

2 Methods

2.1 GRU

For our baseline, we use gated recurrent units (GRU) [12], a simplification of long-short term memory (LSTM) [21], which in turn is a variation of recurrent neural networks (RNNs). A GRU learns based on the following:

$$\mathbf{h}^{(t)} = \left(1 - \mathbf{z}^{(t)}\right) \mathbf{h}^{(t-1)} + \mathbf{z}^{(t)} \tilde{\mathbf{h}}^{(t)} \quad (1)$$

where the *update state* $\mathbf{z}^{(t)}$ decides how much the unit updates its content:

$$\mathbf{z}^{(t)} = \sigma \left(W_z \mathbf{x}^{(t)} + U_z \mathbf{h}^{(t-1)} \right) \quad (2)$$

where W and U are weight matrices. The candidate activation makes use of a *reset gate* $\mathbf{r}^{(t)}$:

$$\tilde{\mathbf{h}}^{(t)} = \tanh \left(W \mathbf{x}^{(t)} + \mathbf{r}^{(t)} \odot \left(U \mathbf{h}^{(t-1)} \right) \right) \quad (3)$$

where \odot is a Hadamard product (element-wise multiplication). When its value is close to zero, the reset gate allows the unit to *forget* the previously computed state. The reset gate $\mathbf{r}^{(t)}$ is computed as follows:

$$\tilde{\mathbf{r}}^{(t)} = \sigma \left(W_r \mathbf{x}^{(t)} + U_r \mathbf{h}^{(t-1)} \right) \quad (4)$$

2.2 BERT

BERT [15] is based on the Transformer [36], a network architecture that depends solely on encoder-decoder attention. The Transformer attention employs a function operating on *queries*, *keys*, and *values*. This attention function maps a query and a set of key-value pairs to an output, where the output is a weighted sum of the values. *Encoder* of the Transformer in [36] has 6 attention layers, each of which is composed of two sub-layers: (1) *multi-head attention* where queries, keys, and values are projected h times into linear, learned projections and ultimately concatenated; and (2) fully-connected *feed-forward network (FFN)* that is applied to each position separately and identically. *Decoder* of the Transformer also employs 6 identical layers, yet with an extra sub-layer that performs multi-head attention over the encoder stack. The architecture of BERT [15] is a multi-layer bidirectional Transformer encoder [36]. It uses masked language models to enable pre-trained deep bidirectional representations, in addition to a binary *next sentence prediction* task captures context (i.e., sentence relationships). More information about BERT can be found in [15].

2.3 Multi-task Learning

In multi-task learning (MTL), a learner uses a number of (usually relevant) tasks to improve performance on a target task [10, 11]. The MTL setup enables the learner to use cues from various tasks to improve the performance on the target task. MTL also usually helps regularize the model since the learner needs to find representations that are not specific to a single task, but rather more general. Supervised learning with deep neural networks requires large amounts of labeled data, which is not always available. By employing data from additional tasks, MTL thus practically augments training data to alleviate need for large labeled data. Many researchers achieve state-of-the-art results by employing MTL in supervised learning settings [20, 25]. In specific, BERT was successfully used with MTL. Hence, we employ multi-task BERT (following [25]). For our training, we use the same pre-trained BERT-Base Multilingual Cased model as the

initial checkpoint. For this MTL pre-training of BERT, we use the same aforementioned single-task BERT parameters. We now describe our data.

3 Data

The shared task dataset contains 5,030 tweets related to different political issues and events in the Middle East taking place between 2011 and 2018. Tweets are collected using pre-defined keywords (i.e. targeted political figures or events) and the positive class involves ironic hashtags such as #sokhria, #tahakoum, and #maskhara (Arabic variants for “irony”). Duplicates, retweets, and non-intelligible tweets are removed by organizers. Tweets involve both MSA as well as dialects at various degrees of granularity such as *Egyptian*, *Gulf*, and *Levantine*.

IDAT@FIRE2019 [17] is set up as a binary classification task where tweets are assigned labels from the set $\{\textit{ironic}, \textit{non-ironic}\}$. A total of 4,024 tweets were released by organizers as training data. In addition, 1,006 tweets were used by organizers as test data. Test labels were not release; and teams were expected to submit the predictions produced by their systems on the test split. For our models, we split the 4,024 released training data into 90% TRAIN ($n=3,621$ tweets; ‘ironic’=1,882 and ‘non-ironic’=1,739) and 10% DEV ($n=403$ tweets; ‘ironic’=209 and ‘non-ironic’=194). We train our models on TRAIN, and evaluate on DEV.

Our multi-task BERT models involve six different Arabic classification tasks. We briefly introduce the data for these tasks here:

- **Author profiling and deception detection in Arabic (APDA).** [28]⁴. From APDA, we only use the corpus of author profiling (which includes the three profiling tasks of *age*, *gender*, and *variety*). The organizers of APDA provide 225,000 tweets as training data. Each tweet is labelled with three tags (one for each task). To develop our models, we split the training data into 90% *training* set ($n=202,500$ tweets) and 10% *development* set ($n=22,500$ tweets). With regard to age, authors consider tweets of three classes: $\{\textit{Under 25}, \textit{Between 25 and 34}, \textit{and Above 35}\}$. For the Arabic varieties, they consider the following fifteen classes: $\{\textit{Algeria}, \textit{Egypt}, \textit{Iraq}, \textit{Kuwait}, \textit{Lebanon-Syria}, \textit{Lybia}, \textit{Morocco}, \textit{Oman}, \textit{Palestine-Jordan}, \textit{Qatar}, \textit{Saudi Arabia}, \textit{Sudan}, \textit{Tunisia}, \textit{UAE}, \textit{Yemen}\}$. Gender is labeled as a binary task with $\{\textit{male}, \textit{female}\}$ tags.
- **LAMA+DINA Emotion detection.** Alhuzali et al. [7] introduce *LAMA*, a dataset for Arabic emotion detection. They use a first-person seed phrase approach and extend work by Abdul-Mageed et al. [4] for emotion data collection from 6 to 8 emotion categories (i.e. *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*). We use the combined LAMA+DINA corpus. It is split by the authors as 189,902 tweets *training* set, 910 as *development*, and 941 as *test*. In our experiment, we use only the training set for our MTL experiments.

⁴ <https://www.auroras.net/APDA/>

- **Sentiment analysis in Arabic tweets.** This dataset is a shared task on Kaggle by Motaz Saad ⁵. The corpus contains 58,751 Arabic tweets (46,940 *training*, and 11,811 *test*). The tweets are annotated with positive and negative labels based on an emoji lexicon.

4 Models

4.1 GRU

We train a baseline GRU network with our irony TRAIN data. This network has only one layer unidirectional GRU, with 500 unites and a linear, output layer. The input word tokens are embedded by the trainable word vectors which are initialized with a standard normal distribution, with $\mu = 0$, and $\sigma = 1$, i.e., $W \sim N(0, 1)$. We use Adam [23] with a fixed learning rate of $1e - 3$ for optimization. For regularization, we use dropout [33] with a rate of 0.5 on the hidden layer. We set the maximum sequence sequence in our GRU model to 50 words, and use all 22,000 words of training set as vocabulary. We employ batch training with a batch size of 64 for this model. We run the network for 20 epochs and save the model at the end of each epoch, choosing the model that performs highest on DEV as our best model. We report our best result on DEV in Table 1. Our best result is acquired with 12 epochs. As Table 1 shows, the baseline obtains *accuracy* = 73.70% and $F_1 = 73.47$.

4.2 Single-Task BERT

We use the BERT-Base Multilingual Cased model released by the authors [15] ⁶. The model is trained on 104 languages (including Arabic) with 12 layers, 768 hidden units each, 12 attention heads. The entire model has 110M parameters. The model has 119,547 shared WordPieces vocabulary, and was pre-trained on the entire Wikipedia for each language. For fine-tuning, we use a maximum sequence size of 50 tokens and a batch size of 32. We set the learning rate to $2e - 5$ and train for 20 epochs. For single-task learning, we fine-tune BERT on the training set (i.e., TRAIN) of the irony task exclusively. We refer to this model as *BERT-ST*, ST standing for ‘single task.’ As Table 1 shows, BERT-ST unsurprisingly acquires better performance than the baseline GRU model. On accuracy, BERT-ST is 7.94% better than the baseline. BERT-ST obtains 81.62 F_1 which is 7.35 better than the baseline.

4.3 Multi-Task BERT

We follow the work of Liu et al. [25] for training an MTL BERT in that we fine-tune the afore-mentioned BERT-Base Multilingual Cased model with different tasks jointly. First, we fine-tune with the three tasks of author profiling and the

⁵ <https://www.kaggle.com/mksaad/arabic-sentiment-twitter-corpus>

⁶ <https://github.com/google-research/bert/blob/master/multilingual.md>.

irony task simultaneously. We refer to this model trained on the 4 tasks simply as BERT-MT4. BERT-MT5 refers to the model fine-tuned on the 3 author profiling tasks, the emotion task, and the irony task. We also refer to the model fine-tuned on all six tasks (adding the sentiment task mentioned earlier) as BERT-MT6. For MTL BERT, we use the same parameters as the single task BERT listed in the previous sub-section (i.e., *Single-Task BERT*). In Table 1, we present the performance on the DEV set of only the irony detection task.⁷ We note that all the results of multitask learning with BERT are better than those with the single task BERT. The model trained on all six tasks obtains the best result, which is 2.23% accuracy and 2.25% F_1 higher than the single task BERT model.

Table 1. Model Performance

Model	Acc	F1
GRU	0.7370	0.7347
BERT-ST	0.8164	0.8162
BERT-MT4	0.8189	0.8187
BERT-MT5	0.8362	0.8359
BERT-MT6	0.8387	0.8387
BERT-1M-MT5	0.8437	0.8434
BERT-1M-MT6	0.8362	0.8360

4.4 In-Domain Pre-Training

Our irony data involves dialects such as Egyptian, Gulf, and Levantine, as we explained earlier. The BERT-Base Multilingual Cased model we used, however, was trained on Arabic Wikipedia, which is mostly MSA. We believe this dialect mismatch is sub-optimal. As Sun et al. [34] show, further pre-training with domain specific data can improve performance of a learner. Viewing dialects as constituting different domains, we turn to dialectal data to further pre-train BERT. Namely, we use 1M tweets randomly sampled from an in-house Twitter dataset to resume pre-training BERT before we fine-tune on the irony data.⁸ We use BERT-Base Multilingual Cased model as an initial checkpoint and pre-train on this 1M dataset with a learning rate of $2e - 5$, for 10 epochs. Then, we fine-tune on MT5 (and then on MT6) with the new *further-pre-trained* BERT model. We refer to the new models as BERT-1M-MT5 and BERT-1M-MT6, respectively. As Table 1 shows, BERT-1M-MT5 performs best: BERT-1M-MT5

⁷ We do not list acquired results on other tasks, since the focus of this paper is exclusively the IDAT@FIRE2019 shared task.

⁸ A nuance is that we require each tweet in the 1M dataset to be > 20 words long, and so this process is not entirely random.

obtains 84.37% accuracy (0.5% less than BERT-MT6) and 83.34 F_1 (0.47% less than BERT-MT6).

4.5 IDAT@FIRE2019 Submission

For the shared task submission, we use the predictions of BERT-1M-MT5 as our first submitted system. Then, we concatenate our DEV and TRAIN data to compose a new training set (thus using all the training data released by organizers) to re-train BERT-1M-MT5 and BERT-MT6 with the same parameters. We use the predictions of these two models as our second and third submissions. Our second submission obtains 82.4 F_1 on the official test set, and ranks 4th on this shared task.

5 Related Work

Multi-Task Learning. MTL has been effectively used to model several NLP problems. These include, for example, syntactic parsing [26], sequence labeling [32, 30], and text classification [24].

Irony in different languages. Irony detection has been investigated in various languages. For example, Hee et al. [35] propose two irony detection tasks in English tweets. Task A is a binary classification task (*irony* vs. *non-irony*), and Task B is multi-class identification of a specific type of irony from the set $\{verbal, situational, other-irony, non-ironic\}$. They use hashtags to automatically collect tweets that they manually annotate using a fine-grained annotation scheme. Participants in this competition construct models based on logistic regression and support vector machine (SVM) [31], XGBoost [29], convolutional neural networks (CNNs) [29], long short-term memory networks (LSTMs) [37], etc. For the Italian language, Cignarella et al. propose the IronTA shared task [13], and the best system [14] is a combination of bi-directional LSTMs, word n -grams, and affective lexicons. For Spanish, Ortega-Bueno1 et al. [27] introduce the IroSvA shared task, a binary classification task for tweets and news comments. The best-performing model on the task, [19], employs pre-trained Word2Vec, multi-head Transformer encoder and a global average pooling mechanism.

Irony in Arabic. Arabic is a widely spoken collection of languages (~ 300 million native speakers) [3, 38]. A large amount of works in Arabic are those focusing on other text classification tasks such as sentiment analysis [5, 6, 2, 1], emotion [7], and dialect identification [38, 16, 8, 9]. Karoui et al. [22] created a Arabic irony detection corpus of 5,479 tweets. They use pre-defined hashtags to obtain irony tweets related to the US and Egyptian presidential elections. IDAT@FIRE2019 [17] aims at augmenting the corpus and enriching the topics, collecting more tweets within a wider region (the Middle East) and over a longer period (between 2011 and 2018).

6 Conclusion

In this paper, we described our submissions to the Irony Detection in Arabic shared task (IDAT@FIRE2019). We presented how we acquire effective models using pre-trained BERT in a multi-task learning setting. We also showed the utility of viewing different varieties of Arabic as different domains by reporting better performance with models pre-trained with dialectal data rather than exclusively on MSA. Our multi-task model with domain-specific BERT ranks *4th* in the official IDAT@FIRE2019 evaluation. The model has the advantage of being exclusively based on deep learning. In the future, we will investigate other multi-task learning architectures, and extend our work with semi-supervised methods.

7 Acknowledgement

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences Research Council of Canada (SSHRC), and Compute Canada (www.computecanada.ca).

References

1. Abdul-Mageed, M.: Modeling arabic subjectivity and sentiment in lexical space. *Information Processing & Management* (2017)
2. Abdul-Mageed, M.: Not all segments are created equal: Syntactically motivated sentiment analysis in lexical space. In: *Proceedings of the third Arabic natural language processing workshop*. pp. 147–156 (2017)
3. Abdul-Mageed, M., Alhuzali, H., Elaraby, M.: You tweet what you speak: A city-level dataset of arabic dialects. In: *LREC*. pp. 3653–3659 (2018)
4. Abdul-Mageed, M., AlHuzli, H., DuaaAbu Elhija, M.D.: Dina: A multidialect dataset for arabic emotion analysis. In: *The 2nd Workshop on Arabic Corpora and Processing Tools*. p. 29 (2016)
5. Abdul-Mageed, M., Diab, M., Kübler, S.: Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language* **28**(1), 20–37 (2014)
6. Al-Ayyoub, M., Khamaiseh, A.A., Jararweh, Y., Al-Kabi, M.N.: A comprehensive survey of arabic sentiment analysis. *Information Processing & Management* **56**(2), 320–342 (2019)
7. Alhuzali, H., Abdul-Mageed, M., Ungar, L.: Enabling deep learning of emotion with first-person seed expressions. In: *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*. pp. 25–35. Association for Computational Linguistics, New Orleans, Louisiana, USA (Jun 2018). <https://doi.org/10.18653/v1/W18-1104>, <https://www.aclweb.org/anthology/W18-1104>
8. Bouamor, H., Habash, N., Salameh, M., Zaghoulani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., et al.: The madar arabic dialect corpus and lexicon. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)* (2018)
9. Bouamor, H., Hassan, S., Habash, N.: The madar shared task on arabic fine-grained dialect identification. In: *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. pp. 199–207 (2019)

10. Caruana, R.: Multitask learning: A knowledge-based source of inductive bias. In: Proceedings of the 10th International Conference on Machine Learning (1993)
11. Caruana, R.: Multitask learning. *Machine learning* **28**(1), 41–75 (1997)
12. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
13. Cignarella, A.T., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P., et al.: Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In: Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018). vol. 2263, pp. 1–6. CEUR-WS (2018)
14. Cimino, A., De Mattei, L., Dell’Orletta, F.: Multi-task learning in deep neural networks at evalita 2018. In: EVALITA@ CLiC-it (2018)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
16. Elaraby, M., Abdul-Mageed, M.: Deep models for arabic dialect identification on benchmarked data. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018). pp. 263–274 (2018)
17. Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., Rosso, P.: Idat@fire2019: Overview of the track on irony detection in arabic tweets. In: Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15 (2019)
18. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 470–478 (2015)
19. González, J., Hurtado, L.F., Pla, F.: Elirf-upv at irosva: Transformer encoders for spanish irony detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS. org, Bilbao, Spain (2019)
20. Guo, H., Pasunuru, R., Bansal, M.: Soft layer-specific multi-task summarization with entailment and question generation. arXiv preprint arXiv:1805.11004 (2018)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
22. Karoui, J., Zitoune, F.B., Moriceau, V.: Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science* **117**, 161–168 (2017)
23. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Liu, P., Qiu, X., Huang, X.: Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101 (2016)
25. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504 (2019)
26. Luong, M.T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114 (2015)
27. Ortega-Bueno, R., Rangel, F., Hernández Farias, D., Rosso, P., Montes-y Gómez, M., Medina Pagola, J.E.: Overview of the task on irony detection in spanish variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS. org (2019)

28. Rangel, F., Rosso, P., Charfi, A., Zaghouni, W., Ghanem, B., Snchez-Junquera, J.: Overview of the track on author profiling and deception detection in arabic. In: Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15 (2019)
29. Rangwani, H., Kulshreshtha, D., Singh, A.K.: Nlprl-iitbhu at semeval-2018 task 3: Combining linguistic features and emoji pre-trained cnn for irony detection in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 638–642 (2018)
30. Rei, M.: Semi-supervised multitask learning for sequence labeling. arXiv preprint arXiv:1704.07156 (2017)
31. Rohanian, O., Taslimipoor, S., Evans, R., Mitkov, R.: Wlv at semeval-2018 task 3: Dissecting tweets in search of irony. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 553–559 (2018)
32. Søgaard, A., Goldberg, Y.: Deep multi-task learning with low level tasks supervised at lower layers. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 231–235 (2016)
33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**(1), 1929–1958 (2014)
34. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? arXiv preprint arXiv:1905.05583 (2019)
35. Van Hee, C., Lefever, E., Hoste, V.: Semeval-2018 task 3: Irony detection in english tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 39–50 (2018)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*. pp. 6000–6010 (2017)
37. Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., Huang, Y.: Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 51–56 (2018)
38. Zhang, C., Abdul-Mageed, M.: No army, no navy: Bert semi-supervised learning of arabic dialects. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop. pp. 279–284 (2019)