Holder and Target Identification on Opinion Text Using Deep Neural Networks

Mohamad Mirza Maulana Ikhsan and
Fariska Zakhralativa Ruskanda

October 26, 2022

Holder and Target Identification on Opinion Text Using Deep Neural Networks

[1]Moh. Mirza Maulana Ikhsan, [2]Fariska Zakhralativa Ruskanda
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
[1]mirzaikhsan181@gmail.com, [2]fariska.zr@staff.stei.itb.ac.id

***Abstract*—The development of social media platforms has made it possible for everyone to be able to express their opinions online. Therefore, various techniques have been developed to extract the information in opinion texts. Opinion role labeling (ORL) aims to identify opinion holder and opinion target within documents. We propose the deep learning models to identify opinion holder and opinion target given opinion text. Based on the experiments using MPQA as training data, we report that the use of Convolutional Neural Network (CNN) architecture for character level feature extraction can increase the F1-score of the BERT-BiLSTM-CRF base as baseline model by 3%. In addition of an opinion expression feature on the model can significantly increase the F1-score of the baseline model by 20%.**

Keywords*—*deep learning, opinion role labeling, opinion holder, opinion target.

I. INTRODUCTION

In 2020, Digital Global Overview report that there was an increase in internet users by 7% which caused the current number of users to reach around 3.8 billion internet users. Moreover, there are currently quite a number of social media platforms, such as Facebook, Twitter, and Instagram which we commonly use as a place to express opinions online. This has become the factors in the development of various techniques that can be used to extract the information within the opinion text.

Opinion Role Labeling (ORL) is one of the tasks of Opinion Mining (OM) which aims to identify the opinion holder and opinion target within documents. There are three types of opinion entities:

- Opinion holder, *H*, is an entity that provides an opinion on a particular claim.
- Opinion target, *T*, is the entity in the opinion text that which is the target of the opinion text.
- Expression of opinion, *DS*, is a marker of an opinion text that can be explicit and implicit.

Figure 1 shows an example of opinion role labeling task

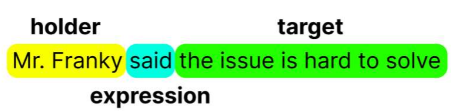

Figure 1. Example of opinion role labeling

In Figure 1, there are labels for opinion holder (H), opinion targets (T), and opinion expressions (DS). The word "said" indicates that "Mr. Franky" is consider that some "issue is hard to solve".

The deep learning approach for opinion role labeling is still understudied [10]. There are four research for opinion role labeling based deep learning, Katiyar, et al., 2016, Marasovic, et al., 2018, Zhang, et al., 2019, and Quan, et al., 2019. Therefore, this research uses deep learning approach because this approach has various potentials that can be explored further.

Our main contribution in this research is the use of character level features using a Convolutional Neural Network (CNN) and opinion expression features to improve the performance of the BERT-BiLSTM-CRF model. The use of the CNN architecture is expected so that the model can learn the character level features of each sentence so that it can overcome errors in the prediction results that only consist of a few characters, such as the words I, You, We, They, Mr., Mrs., etc. Beside the use of the character level feature, there are other features used in this research, that is the opinion expression feature. Opinion expression has a relationship between the opinion holder and the opinion target so that it can make the model to make easier to make such predictions on the opinion holder and opinion target [5].

II. RELATED WORK

A. Opinion Role Labeling

Opinion Role Labeling is one of the Opinion Mining task which aims to identify opinion holder and opinion target within the documents. This task is categorized into sequential labeling problem where each word will have its own label and our task is to predict the label for each word.

In 2016, Katiyar, et al. [5], proposed a deep learning model using the BiLSTM architecture and the opinion expression relation feature to predict the label of each word. By using the MPQA 2.0 corpus as training data and validation data, it is found that the use of the opinion expression relation feature has succeeded in improving the performance of the model.

In 2019, Quan et al. [11] successfully integrated the use of BERT into the BiLSTM-CRF model architecture for opinion role labeling tasks. Based on this research, Quan managed to get a fairly good F1 result, but the recall obtained was still below 50%.

B. Character Level Features

The use of character level features using CNN have been used in some NLP research to get the information from every single character in a word.

In 2014, Kim [6] have studied the use of character level feature by using CNN built on top of word2vec for sentence classification. They report that the use of a simple CNN with one layer of convolution can performs remarkably well.

In 2016, Jason, et al. [2] proposed hybrid architecture by using the BiLSTM and CNN architecture for task named entity recognition. The used of CNN architecture is to extract the information from every single character and then combine the results between character level feature with word level feature. They report the increasing of F1-score by 7%. Not only that, the used of character level feature using CNN can also deal with out-of-vocabulary (OOV) problem or unknown words.

III. METHODOLOGY

The model architecture used in this research consists of four architectures, namely BERT [3], Bidirectional Long Short-Term Memory (BiLSTM) [15], Conditional Random Field (CRF) [16], and Convolutional Neural Network (CNN) [17]. Our model accept word tokens and character tokens as input. BERT is used to get the vector representation for each word in a sentence, while CNN is used to extract character level features for each word. The BiLSTM architecture is used so that the model can learn the information from two directions without losing the context and meaning of a sentence. In addition, the CRF architecture is also used to overcome the bias problem caused by the majority label during the learning process.

We introduce our four BERT based models in detail below:

A. BERT-BiLSTM-CRF

This model is used as baseline model (Figure 2) which combines BERT, BiLSTM, and CRF architectures. This model accepts input in the form of word tokens into BERT layer so that a vector representation of each word is obtained, then enters the BiLSTM layer so that the context is learning from two directions, and finally into the CRF layer to process the opinion holding entities and opinion targets.

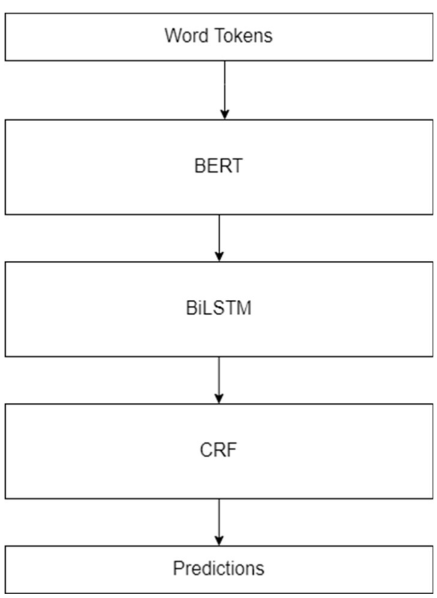

Figure 2. Baseline model architecture

B. BERT-CNN-BiLSTM-CRF

This model accepts input in the form of word tokens and character tokens (Figure 3). The use of CNN architecture in this model is to extract character level features (Figure 4). The use of this character level feature can improve the prediction results of opinion entities that only consist of a few characters. Furthermore, the model combines the character level features with the representation of each word.

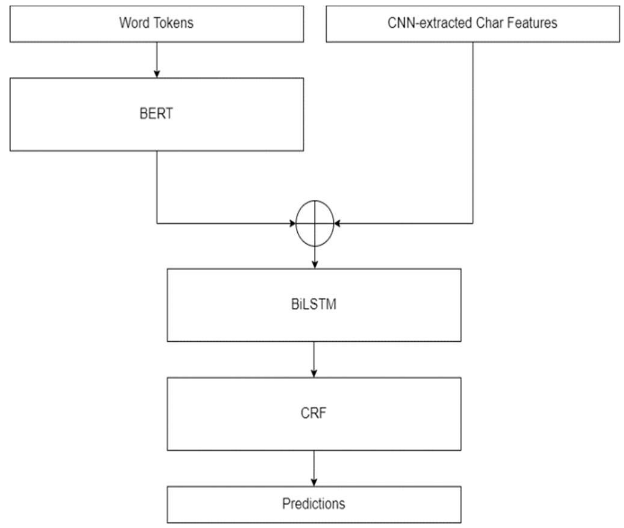

Figure 3. Model architecture with character level features

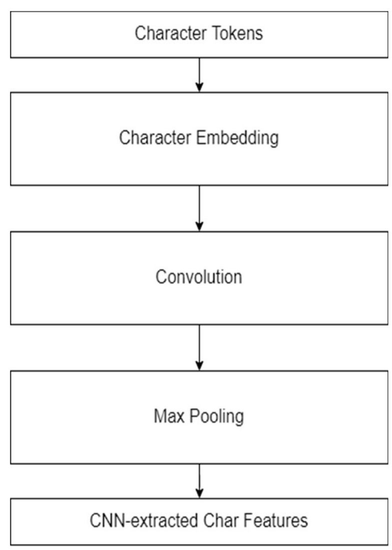

Figure 4. CNN model architecture for character level feature extraction

C. BERT-BiLSTM-CRF with Opinion Expression

This model accepts input in the form of word tokens and opinion expression features obtained from the corpus (Figure 5). The use of the opinion expression feature has a relation with the opinion holder and opinion target so that it can be used as an additional feature in making predictions on the opinion entity. The result of the word representation vector is then combined with the opinion expression feature.

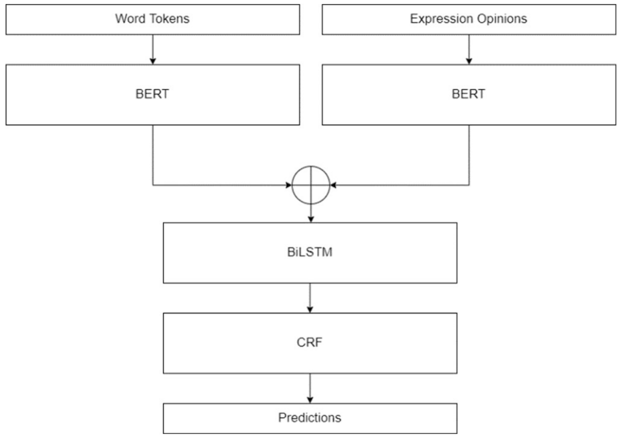

Figure 5. Model architecture with opinion expression features

D. BERT-CNN-BiLSTM-CRF with Opinion Expression

This model is a combination of BERT, CNN, BiLSTM, and CRF architectures with additional features in the form of opinion expression (Figure 6). This model accepts input in the form of word tokens, character tokens, and opinion expression features. The combined results of the three features will then be forwarded to the BiLSTM layer for contextual learning from two directions, and finally into the CRF layer for the prediction process for the opinion holder and opinion target entities.

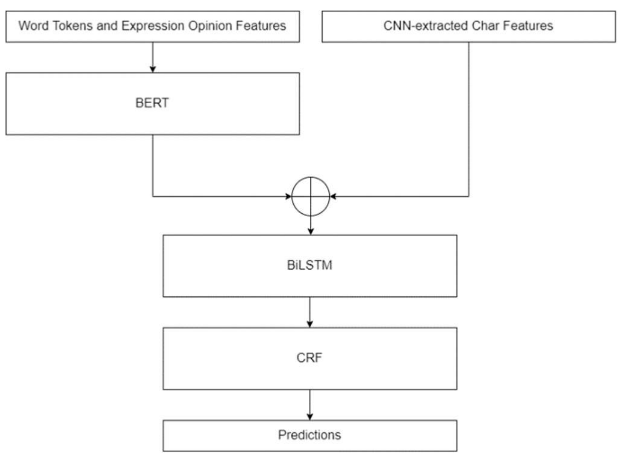

Figure 6. Model architecture with opinion expression features and character level features.

IV. Experiments and Results

A. Datasets

The Multi-Purpose Question Answering (MPQA) corpus is the first annotated corpus that has annotations for opinion holder, opinion target, and opinion expressions. It consists of various news articel that contain facts and opinions. The annotation format used in this corpus is based on a range of words or sentences (also called span based) that are the opinion holder, opinion target, and expression of opinion. For 10-fold CV and pre-processing the corpus, we used the approach from Marasovic and Frank [7].

B. Training Details

We use tensorflow framework for building the proposed model. Training and inference are done on a per-sentence level. All lookup tables for character embedding are randomly initialized, while the LSTM are initialize with zero vector.

We fine-tuned the pre-trained BERT during training in order to make the model learn the task specifically. For BERT model we use the $BERT_{BASE}$, uncased provided by Google Research. Table 1 shows the training configurations for all models.

Table 1. Training configuration

batch size	32
learning rate	1e-3
epochs	50
optimizer	Adam
LSTM units	75
CRF units	7
Embedding output dimension	30
Convolution kernel	3
Convolution filter	30
Dropout	0.5

C. Evaluation Metrics

Evaluation metrics used for span-based identification are binary overlap (Figure 7) and proportional overlap (Figure 8) for precision, recall, and F1-score, respectively. Binary overlap evaluation is an evaluation method that will count the number of overlapping matches between the predicted label sequence and the actual label sequence. As long as the predicted range overlaps with the actual label sequence, the binary overlap evaluation will assume that the predicted results match to the actual label sequence. To improve the results of the binary overlap evaluation, the proportional overlap will take into account the correct length of the overlapping sequence according to the actual label. After calculating the proportional overlap evaluation, then the performance evaluation metrics are calculated using proportional precision, proportional recall, and proportional F1-score for each entity.

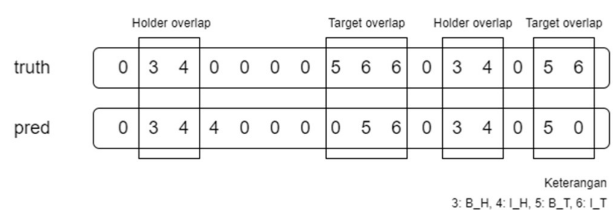

Figure 7. Illustration of binary overlap

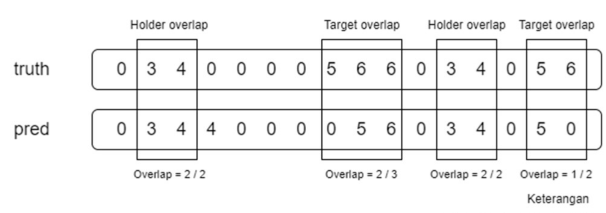

Figure 8. Illustration of proportional overlap

D. Results

Table 2 shows the performance of the model in predicting the opinion role labeling using binary overlap, while Table 3 shows the performance of the model using proportional overlap. The results shown are the average of precision, recall, and F1-score using the 10-fold CV scheme of the resulting model and the standard deviation in superscript in the table with best performances in bold.

In the binary overlap evaluation (Table 2), we can see that our BERT-CNN-BiLSTM-CRF model with opinion expression feature outperforms all the baseline in F1 on labeling opinion holder and target. The use of character level feature are able to improve up to 3% on recall and F1-score. Furthermore the use of opinion expression feature successfully boost the precision, recall, and F1-score up to 20%.

A similar trend can be seen in the proportional overlap evaluation (Table 3), the BERT-CNN-BiLSTM-CRF model with opinion expression feature also outperforms all the baseline in F1 on labeling opinion holder and target. The use of character level feature are able to improve up to 3% on recall and F1-score. Furthermore the use of opinion expression feature successfully boost the precision, recall, and F1-score up to 20%.

Models in opinion holder labeling task generally performs better than in opinion target labeling task in terms of binary overlap and proportional overlap. It's because the opinion holders are usually short and less ambiguous, while opinion targets are longer and sometimes it's challenging for human to annotate or predict.

Comparing both evaluation metrics, binary overlap has higher scores than proportional overlap measurement since the algorithm for proportional overlap are more strict than binary overlap. From both tables we can see that the use of character level feature and opinion expression feature can improves the F1-score for baseline model.

Table 2. Binary overlap 10-fold CV results

Method	Opinion Holder			Opinion Target		
	P	R	F1	P	R	F1
BERT-BiLSTM-CRF	$59.61^{3.90}$	$47.09^{5.27}$	$52.42^{3.53}$	$61.96^{3.04}$	$43.94^{4.81}$	$51.28^{3.68}$
BERT-BiLSTM-CRF + Opinion Expression	$81.41^{3.03}$	$68.46^{3.85}$	$74.32^{2.90}$	$\mathbf{81.35^{3.82}}$	$57.09^{4.37}$	$66.93^{2.75}$
BERT-CNN-BiLSTM-CRF	$59.17^{3.25}$	$49.73^{4.64}$	$54.00^{3.92}$	$60.08^{3.74}$	$47.22^{5.34}$	$52.73^{4.06}$
BERT-CNN-BiLSTM-CRF + Opinion Expression	$\mathbf{82.47^{3.57}}$	$\mathbf{72.42^{3.82}}$	$\mathbf{76.54^{2.43}}$	$80.58^{3.17}$	$\mathbf{63.31^{7.72}}$	$\mathbf{70.63^{4.82}}$

Table 3. Proportional overlap 10-fold CV results

Method	Opinion Holder			Opinion Target		
	P	R	F1	P	R	F1
BERT-BiLSTM-CRF	$58.33^{3.83}$	$46.10^{5.36}$	$51.31^{3.64}$	$57.48^{3.32}$	$40.73^{4.32}$	$47.55^{3.38}$
BERT-BiLSTM-CRF + Opinion Expression	$78.67^{3.51}$	$66.12^{3.34}$	$71.79^{2.75}$	$\mathbf{75.10^{2.51}}$	$54.34^{6.38}$	$63.17^{4.62}$
BERT-CNN-BiLSTM-CRF	$58.09^{3.12}$	$48.82^{4.57}$	$53.01^{3.84}$	$55.66^{3.51}$	$43.77^{5.27}$	$48.87^{4.05}$
BERT-CNN-BiLSTM-CRF + Opinion Expression	$\mathbf{79.76^{4.29}}$	$\mathbf{67.08^{3.62}}$	$\mathbf{72.70^{2.63}}$	$74.78^{2.91}$	$\mathbf{56.32^{5.31}}$	$\mathbf{63.71^{3.93}}$

P, R, F1 defines respectively for precision, recall, and F1-score. The standard deviation written in superscript form.

Other than quantitative results, we also analyze the quantitative results in Table 4 and Table 5. Table 4 shows the comparison results between BERT-BiLSTM-CRF model and BERT-CNN-BiLSTM-CRF, while Table 5 shows the comparison results between BERT-BiLSTM-CRF with opinion expression feature and BERT-CNN-BiLSTM-CRF with opinion expression feature. The prediction results are indicated by square brackets and subscript letters. The green color indicates that the prediction results is in accordance with the actual label, the yellow color indicates that the predicted result doesn't match the actual label in the form of opinion holder, while the blue color indicates that the prediction result doesn't match the actual label for opinion target.

From table 4, we can see that the prediction result of BERT-BiLSTM-CRF model are failed to predict "we" as opinion holder and "he" as opinion target. These failures may happened be cause the small size of dataset. While in the BERT-CNN-BiLSTM-CRF model, it succeed predicting the opinion holder and target. It happened because the model learn the word "we" and "he" from the character level fearture. Similar problem happened again in S2 for BERT-BiLSTM-CRF model. It failed to predict the word "u.s" as opinion holder and "it's" as opinion target. But this problem solved by making the model learn from character level feature. In S3 the BERT-BiLSTM-CRF failed to predict the word "U.S." as one word, instead of "U" and "S". This happened because the tokenization process split the word U.S. into "U" and "S" so the model are unable to predict the label for the word "U". But this problem solved in the BERT-CNN-BiLSTM-CRF.

From Table 5 we can see that most problem have been solved by adding opinion expression feature except for S2. Is S2, the model BERT-BiLSTM-CRF failed to predict word "it's" as opinion target. But in BERT-CNN-BiLSTM-CRF this problem got solved by using it's character level feature and succeesfully predict the word "it's" as opinion target.

Table 4. Quantitative results on BERT-BiLSTM-CRF and BERT-CNN-BiLSTM-CRF.

No	BERT-BiLSTM-CRF	BERT-CNN-BiLSTM-CRF
S1	"If Osama bin Laden is still alive, we suspect he is giving the civil rights advocates a hearty thumbs up."	"[If Osama bin Laden is still alive]$_T$, [we]$_H$ suspect [he]$_T$ is giving the civil rights advocates a hearty thumbs up."
S2	On the other hand, u.s. assurances that it's treating the prisoners humanly don't mean that's so.	On the other hand, [u.s.]$_H$ assurances that it's treating the prisoners humanly don't mean that's so.
S3	U[.S. officials]$_H$ have partially endorsed [that view]$_T$, pointing to rebel leaders such as Shamil Basayev and the Jordanian – born Omar ibn Al Khattab, who are believed to have financial and other ties to Osama bin Laden.	[U.S. officials]$_H$ have partially endorsed [that view]$_T$, pointing to rebel leaders such as Shamil Basayev and the Jordanian – born Omar ibn Al Khattab, who are believed to have financial and other ties to Osama bin Laden.

Table 5. Quantitative results on BERT-BiLSTM-CRF with expression opinion feature and BERT-CNN-BiLSTM-CRF with expression opinion feature

No	BERT-BILSTM-CRF-DS	BERT-CNN-BILSTM-CRF-DS
S1	"If [Osama bin Laden]$_T$ is still alive, [we]$_H$ suspect [he]$_T$ is giving the civil rights advocates a hearty thumbs up."	"[If Osama bin Laden is still alive]$_T$, [we]$_H$ suspect [he]$_T$ is giving the civil rights advocates a hearty thumbs up."
S2	On the other hand, [u.s.]$_H$ assurances that it's treating the prisoners humanly don't mean that's so.	On the other hand, [u.s.]$_H$ assurances that [it's treating the prisoners humanly]$_T$ don't mean that's so.
S3	[U.S. officials]$_H$ have partially endorsed [that view]$_T$, pointing to rebel leaders such as Shamil Basayev and the Jordanian – born Omar ibn Al Khattab, who are believed to have financial and other ties to Osama bin Laden.	[U.S. officials]$_H$ have partially endorsed [that view, pointing to rebel leaders such as Shamil Basayev and the Jordanian – born Omar ibn Al Khattab]$_T$, who are believed to have financial and other ties to Osama bin Laden.

V. Conclusion

In this paper, we propose the use of character level feature and opinion expression feature to improve the performance of baseline model for opinion role labeling task. The use of character level feature from Convolutional Neural Network (CNN) can increase the performance of baseline model up to 3% for binary overlap and proportional overlap. This enhancement is good because of the low recall in baseline model. Meanwhile the use of opinion expression feature can increase the F1-score of the model significanly up to 20%.

Future work will investigates the using of CharacterBERT to compare the result between the used of character level features using CNN architecture. It's because the CharacterBERT has been trained using a large corpus Wikipedia and OpenWebText rather than using character embedding. Furthermore there is other language model that can be used, such as ELMo which has differect architecture and trained using different corpus.

References

[1] Barnes, J., Øvrelid, L., & Velldal, E. (2021*). If you've got it, flaunt it: Making the most of fine-grained sentiment annotations*. http://arxiv.org/abs/2102.00299.

[2] Chiu, J. P. C., & Nichols, E. (2015). *Named Entity Recognition with Bidirectional LSTM-CNNs*. http://arxiv.org/abs/1511.08308.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

[4] Jaariyah, N., & Rainarli, E. (2017). *Conditional Random Fields* Untuk Pengenalan Entitas Bernama Pada Teks Bahasa Indonesia. *Jurnal Ilmiah Komputer Dan Informatika (KOMPUTA)*, *6*(1).

[5] Katiyar, A., & Cardie, C. (2016). *Investigating LSTMs for Joint Extraction of Opinion Entities and Relations*. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 919–929. https://doi.org/10.18653/v1/P16-1087.

[6] Kim. (2014). *Convolutional Neural Networks for Sentence Classification*.

[7] Kim, S. M., & Hovy, E. (2004). *Determining the Sentiment of Opinions*. https://doi.org/10.3115/1220355.1220555.

[8] Marasovic, A. M., & Frank, A. (2018). *SRL4ORL: Improving Opinion Role Labeling Using Multi-Task Learning With Semantic Role Labeling*. http://barbar.cs.lth.se:8081.

[9] Opini. KBBI Daring. https://kbbi.web.id/opini.

[10] Permana, A. (2019, Oktober 8). Dosen ITB Kembangkan Teknologi Natural Language Processing. https://www.itb.ac.id/berita/detail/57286/dosen-itb-kembangkan-teknologi-natural-language-processing.

[11] Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). *Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research*. http://arxiv.org/abs/2005.00357.

[12] Quan, W., Zhang, J., & Hu, X. T. (2019). *End-to-End Joint Opinion Role Labeling with BERT*.

[13] Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). *A Review On The Long Short-Term Memory Model*. *Artificial Intelligence Review*, *53*(8), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1.

[14] Welianto, A. (2020, Maret 13). Fakta dan Opini: Arti dan Ciri-cirinya. https://www.kompas.com/skola/read/2020/03/13/080000869/fakta-dan-opini--arti-dan-ciri-cirinya.

[15] Sepp Hochreiter and J ̈urgen Schmidhuber. 1997. Longshort-term memory.Neural Comput., 9(8):1735–1780, November.

[16] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Condi-tional random fields: Probabilistic models for segmenting and labelingsequence data. 2001.

[17] Phil Blunsom, Edward Grefenstette, Nal Kalchbrenner,et al. 2014. A convolutional neural network for mod-elling sentences. InProceedings of the 52nd AnnualMeeting of the Association for Computational Linguis-tics. Association for Computational Linguistics.