Deep Formality: Sentence Formality Prediction with Deep Learning

Can Li
Department of Electrical Engineering & Computer Science
University of Missouri
Columbia, MO, USA
lican@mail.missouri.edu

Wenbo Wang
Department of Electrical Engineering & Computer Science
University of Missouri
Columbia, MO, USA
wwr34@mail.missouri.edu

Bitty Balducci
Carson College of Business
Washington State University
Pullman, WA, USA
bitty.balducci@wsu.edu

Lingshu Hu
Department of Business Administration
Washington and Lee University
Lexington VA, USA
0000-0003-0304-882X

Matthew Gordon
Department of English
University of Missouri
Columbia, MO, USA
gordonmj@missouri.edu

Detelina Marinova
Robert J. Trulaske, Sr. College of Business
University of Missouri
Columbia, MO, USA
marinovad@missouri.edu

Yi Shang
Department of Electrical Engr. & Comp. Science
University of Missouri
Columbia, MO, USA
shangy@missouri.edu

***Abstract*— Formality analysis is a very important task in natural language processing because formality is a dimension of linguistic variation that language users draw on strategically to achieve effective communications under different situations. However, to our best knowledge, there has been no research that conducts formality prediction with deep learning techniques. Most existing work captures formality with statistical methods based on lexical features and human perception of formality. To fill in this gap, our work focuses on automatic text formality prediction with deep learning methods. In this paper, we proposed two deep learning models, Formality-LSTM and Formality-BERT, for formality prediction that do not need feature engineering. Formality-LSTM is a LSTM-based deep neural network that takes the text and corresponding part-of-speech tag as inputs and outputs the formality score. Formality-BERT is a BERT-based end-to-end deep neural network that takes the original text as input and outputs the formality score. Instead of using different statistical lexical features, the two proposed methods use the sentence content and context to predict formality. We applied both Formality-LSTM and Formality-BERT on a public dataset that contains four genres of text and the results of these models outperform state-of-the-art results for all genres. Formality-BERT outperforms existing models by 14 points on the Spearman correlation between predicted formality and human-labeled formality for four genres.**

***Keywords*— sentence, formality, linguistic, deep learning**

I. Introduction

Formality has been studied by linguistic researchers systematically for many decades. Formality is one dimension or aspect of variation between different language styles. People tend to intuitively form judgements about the formality of a sentence, but may find it difficult to verbalize, define it or quantify it. Some genres typically involve formal styles, such as journalism and professional emails, while others feature relatively informal language such as blogs and online discussion boards [1]. Heylighen and Dewaele define formality as avoidance of ambiguity by minimizing the context-dependence and fuzziness of expressions and propose a way to calculate formality based on the frequencies of different word classes [2]. Sociolinguists and anthropologists use the concept of formality to describe social settings as well as the language and other behavior associated with them [3]. Some research treats formality as closely tied to other characteristics of a social situation, including seriousness, politeness, and respect [4] [5] [6].

Abundance of past studies have shed light on the importance of formality. Reference [3] suggests that formality serves as a useful framework for analyzing characteristics of language along with social actions and structures in the context of discourse events. Other work has shown that using different degrees of formality when expressing the same idea may have different impacts on the listeners' understanding of the sentence [7]. The formality level of language in communication is a very important cue for the speaker's opinions, communication purpose, and familiarity with other speakers [8]. High formality can improve the engagement of online study and participants' attention relative to using casual language [9]. In machine translation application, better results have been achieved by using a lexical formality model to control for the formality level of machine translation output [10]. Study [11] investigates a formality style transfer task that converts informal sentences to formal sentences in order to improve downstream NLP tasks performance. Formality recognition has been applied in different applications: dialogue systems can integrate formality recognition models to improve the interaction [12]. Impact of formality is also considered for text extractive summarization [13].

This paper investigates automatic detection of sentence formality with deep learning rather than relying on a definition of formality. The contributions of this paper are as follows: 1) We propose two models for formality prediction without intense feature engineering: Formality-LSTM and Formality-BERT. To the best of our knowledge, this paper is the first to use deep learning for formality prediction. 2) Our results outperform state-of-the-art model findings by an average of 14 points on the Spearman correlation between predicted formality and human-labeled formality for four genres. Of special note, Formality-BERT improves the results by 33 points on the Spearman correlation with respect to email formality. 3) We release our source code publicly on GitHub.

978-1-6654-6603-5/22/$31.00 ©2022 IEEE
DOI 10.1109/IRI54793.2022.00014

II. RELATED WORK

Formality has been studied by prior work on different levels. Some work focuses on formality at the genre level which contains relatively large chunks of text [2] [14], while other studies address formality of smaller units of text, such as at the sentence-level [1] [15] [16], or word-level [17] [18] [19].

Formality has been treated as a binary classification problem for formal-informal document classification [20], email classification [21] and web text classification [22]. As linguistic research notes, however, language is not either formal or informal but varies continuously along this dimension, and formality is best treated as a continuum and measured by multiple levels of scores [2] [3].

All previous research about formality detection is based on feature engineering from human perception of formality where the methods used can be categorized into mathematical formulas, statistical models, and machine learning models. Heylighen and Dewaele proposed the F-score -which uses the percentage of part-of-speech tags to calculate formality score with nouns, adjectives, articles and prepositions as positive terms, and adverbs, verbs and interjections as negative terms [2]. The CF-score, a variation of the F-score is a combination of five scores: narrativity, referential cohesion and deep cohesion, syntactic simplicity, word concreteness [14]. Statistical models, like ridge regression, have been used to predict sentence-level formality with 11 types of feature groups, including lexical features, POS tags, ngrams, embedding, etc [1]. Various machine learning methods, including Decision Trees, Naive Bayes, and Support Vector Machine, also have been applied to document formality classification [20].

There are three main disadvantages for these methods mentioned above. The first disadvantage is that they require intense feature engineering to find appropriate features for formality prediction. The second drawback is that most features they use are word-level linguistic features and the context and structure of text are not considered. The third downside is that since human's perceptions of formality is relatively subjective, featured-based models are hard to generalize to different types of datasets. This paper proposes a transformer-based deep neural network, Formality-BERT, which can solve these three problems accordingly.

III. FORMALITY DATASET

The Likert scale approach [23] is the most popular method for sentence-level formality annotation when formality is treated as a continuum. We have only found three studies that incorporate the task of sentence-level formality annotation. In study [16], 600 sentences were annotated with a Likert scale of 1-5. Lahiri extended this work and released 7,032 labeled sentences from news, blogs, and forums with a Likert scale of 1-7 for formality [24]. More recently, Pavlick and Tetreault took the news and blogs corpus from [24] and added 1,701 sentences from emails and 4,977 sentences from Yahoo Answers to form a new dataset of 1,1274 annotated sentences [1]. They used a Seven-point Likert scale -3 to 3.

Our study is conducted on the dataset published by [1]. Fig. 1 is the distribution of the number of samples for four genres. A few examples of the dataset are shown in Table I.

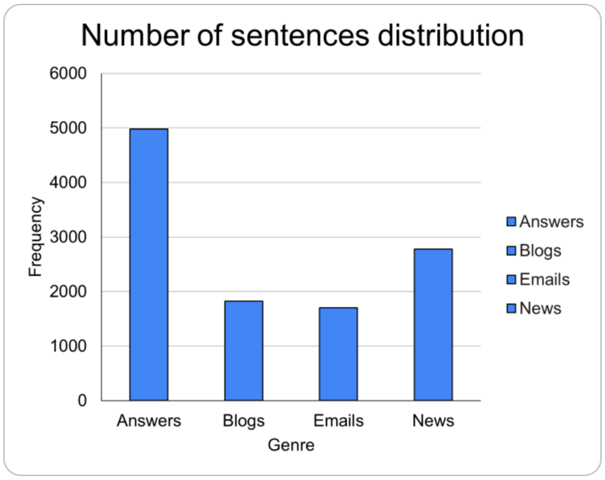


Fig. 1. Distribution of the number of sentences by genre

Each sentence has 5 formality scores labeled independently by 5 different people. Since different people may have different formality perceptions of the same sentence, there is variation in the 5 formality labels for the same sentence. A negative score means more informal, a score near zero means neutral, and a positive score is more formal. The average of the 5 scores for each sentence are used as the final formality score for the modeling process. Fig. 2 is the distribution of the final formality score for the four genres. In general, answers and blogs are more informal while emails and news are more formal.

IV. PROPOSED METHODS

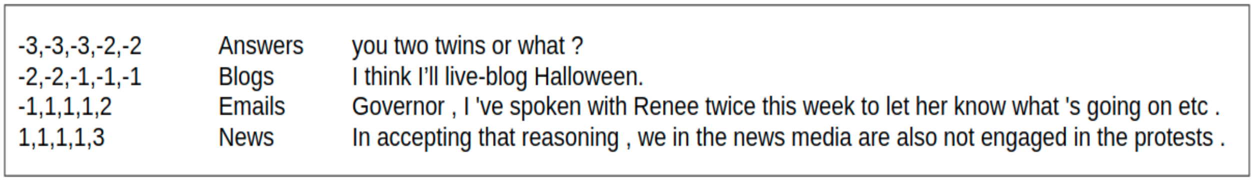
The sentence-level formality score is a continuous decimal number, so we can treat this task as a regression problem. For continuous formality detection, the state-of-the-art method is a statistical regression model fitted on various linguistic features extracted from the text of sentences [1]. But these linguistic features can only represent discrete information and may not be able to capture the underlying context information of the whole sentence and dependencies between different words. In this paper we propose two deep learning methods that can solve this problem.

A. Formality-LSTM

LSTM, a variant of RNN, is able to catch long-term dependencies by using internal mechanisms called gates that can regulate the flow of information [25]. Our proposed Formality-LSTM consists of two LSTM components as shown in Fig. 3.

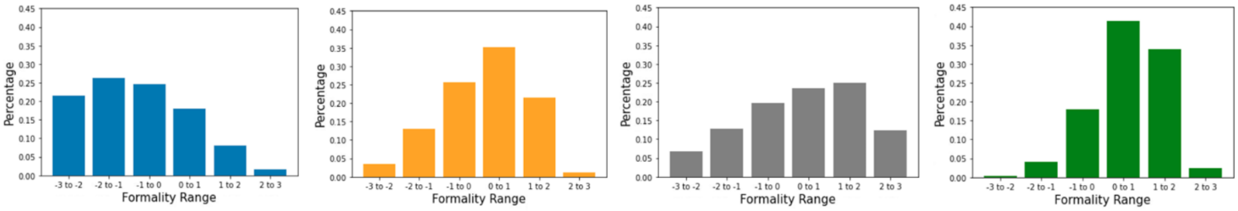
The first LSTM component is responsible for understanding the text content of the sentence. We extract a 768-dimensional word embedding from the pre-trained BERT-base-cased [26] model for each token in the sentence. Then the embedding array of each sentence is padded to have the same length as the maximum number of tokens of all

TABLE I. EXAMPLES FOR FORMALITY DATASET



-3,-3,-3,-2,-2	Answers	you two twins or what ?
-2,-2,-1,-1,-1	Blogs	I think I'll live-blog Halloween.
-1,1,1,1,2	Emails	Governor , I 've spoken with Renee twice this week to let her know what 's going on etc .
1,1,1,1,3	News	In accepting that reasoning , we in the news media are also not engaged in the protests .

This table shows some examples formality dataset. The first column contains 5 formality scores rated by 5 different people. The second column is the genre type. The third column is the text of the sentence.



(a) Answer (b) Blog (c) Email (d) News

Fig. 2. Distribution of sentence-level formality scores by genre. Answers and blogs are relatively informal than emails and news.

sentences and the padded tokens are masked. The padded and masked embedding is subsequently fed into a LSTM which outputs the content feature for the whole sentence.

The goal of the second component is to capture the structure information of the sentence from the POS tags. First, we use Stanford CoreNLP [27] toolkit to extract POS tags for each sentence. The same padding and masking process introduced in the first component is applied to the POS tags. Followed is an embedding layer that is used to learn a 256-dimension embedding for each POS tag. Then a LSTM is employed to extract the structure feature for the sentence.

After we get both content feature and structure feature, a merge layer will fuse them together by concatenating them. The merged feature will go through a few dense and dropout layers. Finally, a formality score will be generated by the output layer. ReLU activation function is used for all internal dense layers and linear activation for the output layer.

B. Formality-BERT

Transformer [28] has been the driving force for the recent breakthrough in the NLP field. BERT, a transformer-based model, uses masked language training mode to learn deep bidirectional language representation. It has been widely used in language understanding and NLP downstream tasks, like sentiment analysis, question answering, and conversation understanding [29]. We proposed an end-to-end automatic formality prediction model Formality-BERT, a BERT-based model as shown in Fig. 4, which can use the whole-sentence context to predict formality instead of word-level linguistic features. The input of Formality-BERT is the raw sentence. On top of the input layer is a BERT tokenizer. The tokenized sentence will be fed into a BERT-base-cased encoder to get the sentence embedding. Then we put 5 blocks of dense layer and dropout layer on top of it. The final output layer is a single neuron to predict the formality score. All the internal dense layers use ReLU as activation function and the output layer uses linear activation.

V. EXPERIMENT SETUP

The dataset introduced in the formality dataset section is used to train the models. We train one different model for each genre. The state-of-the-art regression model for formality prediction is ridge regression in [1]. We will use it as a baseline to compare with our proposed methods.

A. Experiments

Ridge Regression (baseline) A ridge regression model is fitted on 11 types of feature groups, including length, case, POS tags, ngrams, punctuation, etc. This model focuses on

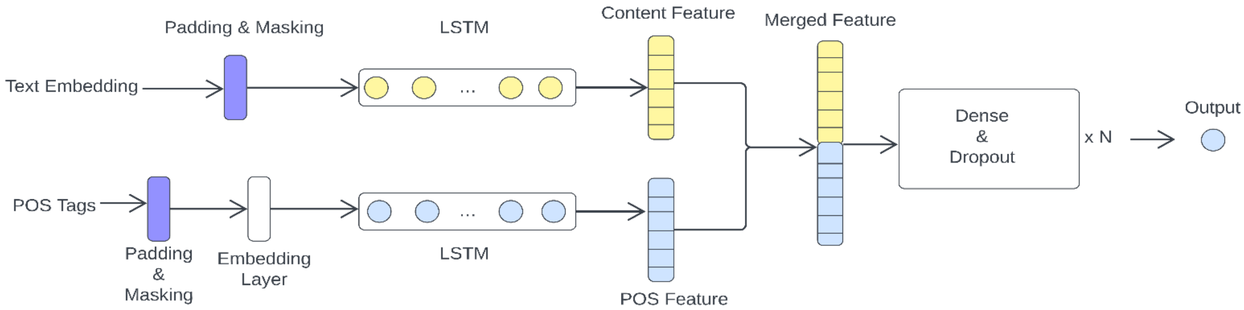


Fig. 3. Formality-LSTM architecture. Formality-LSTM has two LSTM components. The first LSTM extracts content features from text embedding. The second LSTM captures POS features from POS tags. Then content features and POS features are merged and fed through a few dense and dropout layers to finally output a formality score.

statistical linguistic features and ignores the sequence and context information.

Formality-LSTM This model extracts the content information from the text and structure information from the POS tags, then fuses them together to predict the formality. But this model can only use the left-to-right context information of the sentence. We implemented this method on text only, POS tags only, and both.

Formality-BERT A BERT-based model which uses both left-to-right and right-to-left context formation to predict formality.

B. Training Setup

For each genre, we use 80% of the data for training, 10% as validation, and 10% for testing. Since this is a regression problem, we use mean squared error (MSE) as loss function. For performance metric, we want to pick something that can measure how the model behavior algins with human perceptions of formality and the relative formality rank correlation. Therefore, Spearman correlation, as shown in (1), between predicted formality score and human labeled formality score is used as performance measurement. Also, it will be consistent when comparing our performance with paper [1] since it uses spearman correlation as well. The optimizer is Adam [30] and the learning rate is 3e^{-5}. Early stopping with a patience of 10 epochs is used to prevent overfitting.

$$r_s = \frac{cov(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}, \qquad (1)$$

where cov(R(X), R(Y)) is the covariance of the rank variables,
$\sigma_{(R(X))}$ and $\sigma_{(R(Y))}$ are the standard deviation of the rank variables.

VI. RESULTS

Table II reports the results of Spearman correlation between human labeled formality scores for all models on each genre. Both of our two proposed deep learning methods outperform the previous state-of-the-art results generated by ridge regression. Formality-BERT achieves the best results, surpassing ridge regression by an average of 14 points on all genres and 33 points on news. This indicates that the context of the sentence is very important for understanding formality and word-level statistical linguistic features are not enough for predicting formality.

Comparing the results of Formality-LSTM on POS, Text, and POS + Text, we see both the text content and POS features are useful for formality prediction, but the content of the sentence plays a bigger role than POS tags.

Formality-BERT performs better than Formality-LSTM on all the genres because Formality-BERT uses the whole context information whereas Formality-LSTM only considers left-to-right information.

VII. CONCLUSION

In this paper, we propose two deep learning methods for automatic sentence-level formality prediction without intense feature engineering work. Both methods achieve state-of-the-art performance and Formality-BERT improves the previous results by an average of 14 points for four genres. Our results show that the context of the sentence is very important for understanding formality and word-level statistical linguistic features are not enough for formality prediction. Given the lack of such studies, we hope our findings can provide guidance to other scholars interested in using deep learning to predict formality judgments.

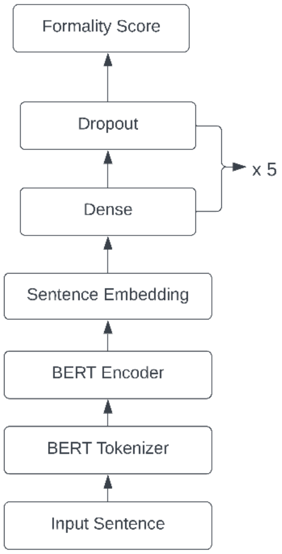
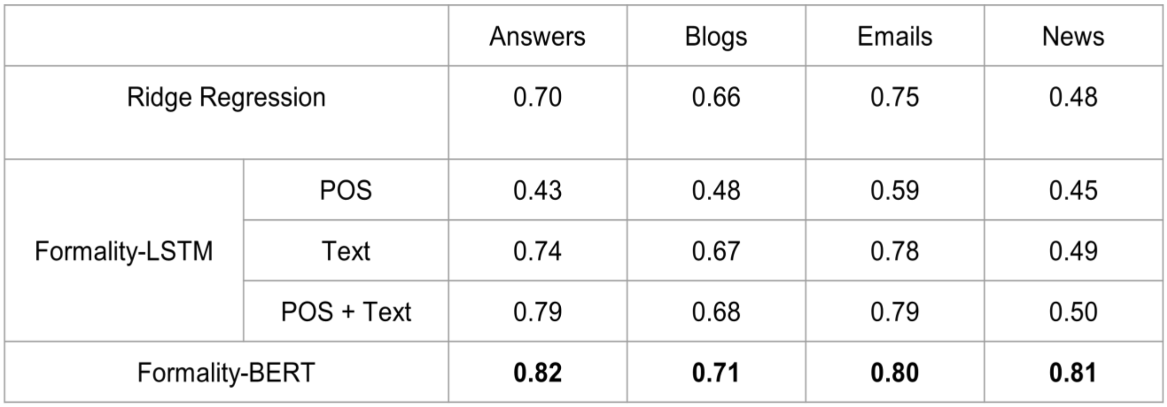


Fig. 4. Formality-BERT architecture. Formality-BERT consists of BERT tokenizer, BERT encoder, and 5 dense and dropout layer blocks.

TABLE II. SPEARMAN CORRELATION FOR ALL MODELS ON EACH GENRE



		Answers	Blogs	Emails	News
Ridge Regression		0.70	0.66	0.75	0.48
Formality-LSTM	POS	0.43	0.48	0.59	0.45
	Text	0.74	0.67	0.78	0.49
	POS + Text	0.79	0.68	0.79	0.50
Formality-BERT		**0.82**	**0.71**	**0.80**	**0.81**

REFERENCES

[1] E. Pavlick and J. Tetreault, "An empirical analysis of formality in online communication," *Transactions of the Association for Computational Linguistics,* vol. 4, pp. 61-74, 2016.

[2] F. Heylighen and J.-M. Dewaele, "Formality of language: definition, measurement and behavioral determinants," *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel,* vol. 4, 1999.

[3] J. T. Irvine, "Formality and informality in communicative events," *American anthropologist,* vol. 81, no. 4, pp. 773-790, 1979.J. Fischer, "The stylistic significance of consonantal sandhi in Trukese and Ponapean," American Anthropologist, 67(6), 1965, pp.1495-1502.

[4] J. L. Fischer, "The stylistic significance of consonantal sandhi in Trukese and Ponapean," *American Anthropologist,* vol. 67, no. 6, pp. 1495-1502, 1965.

[5] S. Ervin-Tripp, "On sociolinguistic rules: Alternation and co-occurrence," *Directions in sociolinguistics,* vol. 2, pp. 213-250, 1972.

[6] J. A. Fishman, "Sociolinguistics: A brief introduction," 1970.

[7] E. Hovy, "Generating natural language under pragmatic constraints," *Journal of Pragmatics,* vol. 11, no. 6, pp. 689-719, 1987.

[8] B. Endrass, M. Rehm, and E. André, "Planning small talk behavior with cultural influences for multiagent systems," *Computer Speech & Language,* vol. 25, no. 2, pp. 158-174, 2011.

[9] T. August and K. Reinecke, "Pay attention, please: Formal language improves attention in volunteer and paid online experiments," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1-11.

[10] X. Niu, M. Martindale, and M. Carpuat, "A study of style in machine translation: Controlling the formality of machine translation output," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2814-2819.

[11] K. Chawla and D. Yang, "Semi-supervised formality style transfer using language model discriminator and mutual information maximization," *arXiv preprint arXiv:2010.05090,* 2020.

[12] F. Mairesse, "Learning to adapt in dialogue systems: data-driven models for personality recognition and generation," University of Sheffield, 2008.

[13] P. Sidhaye and J. C. K. Cheung, "Indicative Tweet Generation: An Extractive Summarization Problem?," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 138-147.

[14] H. Li, Z. Cai, and A. C. Graesser, "Comparing two measures for formality," in *The Twenty-Sixth International FLAIRS Conference*, 2013.

[15] S. Lahiri, P. Mitra, and X. Lu, "Informality judgment at sentence level and experiments with formality score," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2011: Springer, pp. 446-457.

[16] S. Lahiri and X. Lu, "Inter-rater agreement on sentence formality," *arXiv preprint arXiv:1109.0069,* 2011.

[17] J. Brooke, T. Wang, and G. Hirst, "Automatic acquisition of lexical formality," in *Coling 2010: Posters*, 2010, pp. 90-98.

[18] J. Brooke and G. Hirst, "Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 2172-2183.

[19] E. Pavlick and A. Nenkova, "Inducing lexical style properties for paraphrase and genre differentiation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 218-224.

[20] F. A. Sheikha and D. Inkpen, "Automatic classification of documents by formality," in *Proceedings of the 6th international conference on natural language processing and knowledge engineering (nlpke-2010)*, 2010: IEEE, pp. 1-5.

[21] K. Peterson, M. Hohensee, and F. Xia, "Email formality in the workplace: A case study on the Enron corpus," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 86-95.

[22] A. Mosquera and P. Moreda, "Smile: An informality classification tool for helping to assess quality and credibility in web 2.0 texts," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2012, vol. 6, no. 3, pp. 2-7.

[23] R. Likert, "A technique for the measurement of attitudes," *Archives of psychology,* 1932.

[24] S. Lahiri, "SQUINKY! A corpus of sentence-level formality, informativeness, and implicature," *arXiv preprint arXiv:1506.02306,* 2015.

[25] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput,* vol. 9, no. 8, pp. 1735-1780, 1997.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[27] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55-60.

[28] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[29] C. Li, W. Wang, B. Balducci, D. Marinova and Y. Shang, "Predicting Conversation Outcomes Using Multimodal Transformer," *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1-6, doi: 10.1109/IJCNN52387.2021.9533935.

[30] D. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.