Language or Paralanguage, This is the Problem: Comparing Depressed and Non-Depressed Speakers Through the Analysis of Gated Multimodal Units

Nujud Aloshban[1]*, Anna Esposito*[2] *and Alessandro Vinciarelli*[1]

[1]University of Glasgow, Glasgow (UK)
[2]Università degli Studi della Campania "Luigi Vanvitelli", Caserta (Italy)
n.aloshban.1@research.gla.ac.uk, Anna.Esposito@unicampania.it,
Alessandro.Vinciarelli@glasgow.ac.uk

Abstract

Speech-based depression detection has attracted significant attention over the last years. A debated problem is whether it is better to use language (what people say), paralanguage (how they say it) or a combination of the two. This article addresses the question through the analysis of a Gated Multimodal Unit trained to weight modalities according to how effectively they account for the condition of a speaker (depressed or non-depressed). The experiments involved 29 individuals diagnosed with depression and 30 non-depressed participants. Besides an accuracy of 83.0% (F1 score 80.0%), the results show that the Gated Multimodal Unit tends to give more weight to paralanguage. However, the relative contribution of language tends to be higher, to a statistically significant extent, in the case of non-depressed speakers.

Index Terms: Computational paralinguistics, depression detection, social signal processing, Gated Multimodal Units

1. Introduction

According to the estimates of the World Health Organization, depression affects roughly 4.4% of the world's population [1]. However, only a limited number of patients receives appropriate medical attention, partly because depressed people try to avoid the stigma associated to the pathology and escape treatment, partly because resources for depression diagnosis, an expensive and time-consuming process, are not always available [2]. For these reasons, the computing community has made major efforts towards the development of depression detection technologies (see, e.g., [3] for an extensive survey). Proposed approaches use a wide spectrum of modalities as input (e.g., facial expressions [4], social media posts [5], etc.). However, it is still unclear whether any modality is more likely than the others to carry depression-relevant information. This article addresses such a problem, at least to a partial extent, by testing whether depression traces are more likely to appear in *language* (what people say) or *paralanguage* (how they say it).

The experiments involved 59 participants, including 29 people diagnosed with depression. They were recorded while answering questions about everyday life (e.g., "What did you do in the last week end?", "Do you have a family?", etc.) and the total amount of material corresponds to roughly 4 hours. Following the approaches typical of Social Signal Processing [6] and Computational Paralinguistics [7], the recordings were converted into sequences of feature vectors expected to account for paralinguistic aspects of speech. Furthermore, the data were manually transcribed and segmented into clauses, atomic linguistic units including a noun, a verb and a complement. In such a way, it was possible to jointly model language and paralanguage, while comparing their respective contributions to depression detection.

Transcriptions and sequences of feature vectors were fed to a multimodal approach based on Bidirectional Long Short-Term Memory Networks (BiLSTM) [8] and Gated Multimodal Units (GMU) [9]. These latter were trained to weight the two input modalities according to how likely they led to the correct classification of a speaker. The accuracy was up to 83.0% (F1 score 80.0%). However, the key-result of the work is that the ratio $w = w_l/w_p$ (w_l and w_p are the weights that the GMU assigns to language and paralanguage, respectively) tends to be higher, to a statistically significant extent, in the case of control participants. In other words, the relative contribution of language tends to be more important in the case of control participants, thus suggesting that these manifest their condition through language more than depressed ones.

To the best of our knowledge, this is one of the first works highlighting the difference above. The main reason why such a result is important is that the question of whether depression detection should focus on language, paralanguage or the combination of the two is still open. Some works show that the best results can be achieved by using only one of the two modalities (e.g., language was shown to lead to the best performance in [10]). Others suggest that language and paralanguage should always be jointly modeled through multimodal approaches (see, e.g., [11]). However, these appear to be particularly effective when including a component (e.g., a Gated Multimodal Unit [9]) capable to select only one of the two input modalities, thus suggesting that the other is unlikely to convey enough depression-relevant information (see, e.g., [12, 13]). Finally, it was shown that the multimodal approaches based on language and paralanguage require sometimes extra input to be effective (e.g., the point in time a sentence was uttered during a conversation [14]). Overall, this brief state-of-the-art suggests that the relative contributions of language and paralanguage are still to be clarified and, for this reason, this article presents experiments aimed at investigating such a problem.

The rest of this article is organized as follows: Section 2 describes the data used in the experiments, Section 3 describes the depression detection approach, Section 4 presents experiments and results, and Section 5 draws some conclusions.

2. The Data

The experiments have involved 59 participants split into two groups, namely 29 individuals diagnosed with depression by professional psychiatrists and 30 control individuals that have never experienced mental health issues. Table 1 shows the distribution of gender, age and education level across the two groups. According to a χ^2 test, there is no difference between

	F	M	Avg. Age	Range	P	S
Depressed	21	8	45.7	23-69	16	13
Control	21	9	44.0	23-68	12	18
Total	42	18	44.4	23-69	28	31

Table 1: *The table shows the demographic information available about the participants. The letters* F *and* M *refer to gender and stand for female and male, respectively. The expression* Avg *and* Range *refer to age and stand for average and range, respectively. Finally, the letters* P *and S refer to the education level and stand for primary (8 years of study at most) and superior (13 years of study at least), respectively.*

depressed and control participants in terms of gender and education level. Similarly, according to a two-tailed t-test, there is no difference in terms of age. This is important because it means that observable differences in paralanguage and language, the two modalities used in this work, depend on the condition of a participant (depressed or non-depressed) and not on other factors. Therefore, the proposed approach actually detects depression and not other individual characteristics.

Every participant was recorded while being interviewed about everyday life aspects such as activities in the week end or interaction with family members (interviewers posed always the same questions and always in the same order). The result is a corpus in which the average duration of the recordings is 242.2 seconds, corresponding to a total amount of 3 hours, 58 minutes and 10 seconds. The average length of the interviews is 267.1 seconds for control participants and 216.5 seconds for depressed ones. According to a one-tailed t-test, such a difference is statistically significant ($p < 0.01$). The interviewers were instructed to speak as little as possible and, on average, they acount for 10.0% of the interview duration. However, such a percentage is 5.0% and 14.7% when taking into account only control or only depressed participants. Such a difference is statistically significant with $p < 0.01$ according to a two-tailed t-test.

The interviews were manually transcribed and segmented into clauses, atomic linguistic units that include a noun, a verb and a complement. The transcription is synchronized with the speech signal so that it is possible to know what are the words being uttered in correspondence of a given signal segment. This allows one to apply multimodal approaches that jointly model acoustic and linguistic aspects of speech (see Section 3). The average number of clauses per participant is 114.0 and there is a statistically significant difference ($p < 0.05$ according to a one-tailed t-test) between control and depressed participants for which the average number of clauses is 126.9 and 100.8, respectively. However, the difference between the average number of words per participant, 429.7 for depressed and 463.9 for control, is not statistically significant. This suggests that depressed participants tend to use more words per clause.

The recordings were collected in three mental health centers in Italy and all participants accepted to be involved on a fully voluntary basis. Each of them signed a an informed consent letter formulated in accord with the privacy and data protection procedures established by the Italian and European laws. The ethical committee of the Department of Psychology at Universitá degli Studi della Campania "Luigi Vanvitelli", responsible for the data collection, provided the ethical clearance with protocol number 09/2016.

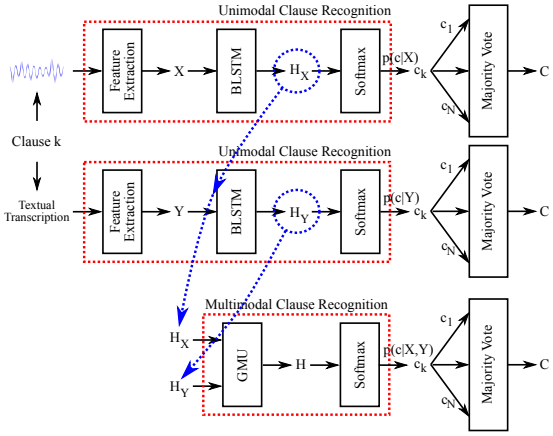

Figure 1: *The figure shows the scheme of the unimodal and multimodal recognition approaches. The decisions made at the level of individual clauses are aggregated through a majority vote. The expression c_k is the class assigned to clause k of a given speaker.*

3. The Approach

Section 2 shows that the interviews used in this work were segmented into clauses, short sentences that act as analysis units in the recognition experiments. The proposed approach (see Figure 1) classifies every clause as being uttered by a depressed or control speaker. The decisions made at the level of individual clauses are then aggregated through a majority vote (the speaker is assigned to the class her or his clauses are most frequently assigned to). The classification is performed using individual modalities (see Section 3.1) or their combination (see Section 3.2).

3.1. Unimodal Classification

The unimodal clause classification approach (see Figure 1) includes three main steps:

- *Feature Extraction*: conversion of the two input streams (speech signal and its transcription) into sequences of feature vectors X and Y;
- *Representation*: conversion of X and Y into sequences of hidden representations H_X and H_Y;
- *Recognition*: assignment of representation H_X or H_Y to one of the two possible classes (depressed or control).

In the case of the speech signal, the feature extraction process segments the signal into 25 ms long analysis windows that start at regular time steps of 10 ms (two consecutive windows overlap by 15 ms). After the segmentation, the signal intervals enclosed in the individual windows are mapped into feature vectors where the components correspond to the first 39 Mel Frequency Cepstral Coefficients (MFCCs). Such a representation is common in Computational Paralinguistics [7] and it has been shown to be effective in capturing social and psychological phenomena, including depression [15]. As a result, the feature extraction process converts the original speech signal into a sequence $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ of feature vectors, each accounting for one of the analysis windows.

In parallel to the speech signal, the feature extraction process converts the transcription of every clause into a sequence $Y = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)$, where every vector corresponds to a word. The individual vectors $\boldsymbol{y}_k$ are the output

Modality	Level	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Text	C	60.4 ± 0.003	56.1 ± 0.005	46.5 ± 0.007	51.0 ± 0.005
Text	P	72.9 ± 0.020	100.0 ± 0.000	44.8 ± 0.040	61.9 ± 0.040
Audio	C	70.0 ± 0.006	65.1 ± 0.008	65.0 ± 0.008	65.0 ± 0.007
Audio	P	74.6 ± 0.021	76.9 ± 0.032	69.0 ± 0.306	72.7 ± 0.021
Multimodal	C	63.0 ± 0.004	58.1 ± 0.005	54.5 ± 0.010	56.2 ± 0.007
Multimodal	P	83.0 ± 0.024	95.2 ± 0.024	69.0 ± 0.049	80.0 ± 0.032
Baseline	C	50.1	43.4	43.4	43.4
Baseline	P	50.0	49.1	49.1	49.1

Table 2: *The table shows, at both clause (C) and person (P) level, accuracy, Precision, Recall and F1 measure. Since all experiments are repeated 30 times, the values are accompanied by their respective standard errors.*

of *Wikipedia2Vec* [16], a *Word Embedding* (WE) approach [17]. The core idea underlying WE is that it is possible to train a shallow network (only one hidden layer) to map every word n of a training text into its following word $n+1$, where the words are represented with one-hot vectors (the dimension of the vector is the size of the dictionary and all components are set to zero except the one that corresponds to the word to be represented).

Once the shallow network is trained, the weight w_{ij} (corresponding to the connection between input neuron i and hidden neuron j) can be thought of as the j^{th} component of a vector reresenting word i in the dictionary. In the particular case of this work, *Wikipedia2Vec* appears to work better than more sophisticated WE methodologies (e.g., the *Bidirectional Encoder Representations from Transformers* or BERT [18]) and, therefore, it has been preferred. One probable reason is that clauses tend to be short and do not allow sophisticated methodologies to be of full benefit.

The sequences X and Y extracted from a given clause are fed to two unimodal Bi-LSTMs with a final *softmax* layer that gives as output the probability of the clause having been uttered by a depressed speaker. Such an approach classifies every clause individually, but what matters from an application point of view is the classification of speakers. For this reason, the decisions made at the level of individual clauses are aggregated through a majority vote, i.e., a speaker is assigned to the class her or his clauses are most frequently assigned to:

$$\hat{c} = \arg\max_{c \in \mathcal{C}} n(c), \tag{1}$$

where $n(c)$ is the number of clauses assigned to class c and $\mathcal{C}$ is the set of all classes (depressed and control in the experiments of this work).

3.2. Multimodal Combination

Sequences X and Y are fed to two unimodal BiLSTMs that give as output two sequences of hidden representations H_X and H_Y (see Figure 1). These are then combined through a *Gated Multimodal Unit* [9] providing as output a representation H that, fed to a softmax layer, leads to the probability of the input clause having been uttered by a depressed speaker. The main motivation behind the use of the GMU is that, besides providing a multimodal representation combining language and paralanguage, it weights the unimodal inputs according to how likely they are to convey information relevant to the condition of the speaker. In particular, H_X is assigned a weight w_p and H_Y is assigned a weight w_l, with $w_p + w_l = 1$. This allows one to assess the contribution of individual modalities to the final classification outcome and, indirectly, it gives insight on whether depression manifests itself through what people say or through how they say it. Like in the unimodal case, the classifications made at the level of individual clauses are aggregated through a majority vote (see above).

4. Experiments and Results

The experiments were performed according to a k-fold experimental setup ($k = 5$). The data corresponding to every participant were randomly assigned to one of the folds. In such a way, the data of each participant appears in one fold only and, as a consequence, the same participant never appears in both training and test set. This guarantees that the experiments are *person-independent*, i.e., that the approaches recognize the condition of the participants and not simply their voice or identity. Every experiment was repeated 30 times and, at every repetition, the weights of the networks were initialized randomly. Correspondingly, the recognition results are reported in terms of average and standard deviation across the repetitions.

The optimal value of the hyperparameters was found through crossvalidation and the search space was defined by taking into account a set of values that the literature considers to be standard. In the case of the learning rate, the values were 10^{-3}, 3×10^{-3}, 10^{-2}, and 10^{-1}. The number of training epochs was set to 30, 50 or 80. The number of hidden neurons in the Bi-LSTMs was set to 32, 64 or 128. The padding values for the sequences of feature vectors extracted from the speech signals were 40, 50, 60, 70, 80, 100 and 120. Finally, the padding values for text were all integers between 9 and 14 included. The models were trained through backpropagation by using the Adam optimizer [19] and categorical cross-entropy as a loss function [20]. All models and training methodologies were implemented with *Tensorflow*.

Table 2 shows the results at both clause level (effectiveness at classifying individual clauses) and person level (effectiveness at classifying participants through the application of a majority vote over their clauses). The low standard deviation values suggest that the classification outcomes do not change substantially with the initialization of the networks. According to a two-tailed t-test ($p << 0.01$), all systems improve over a random classifier that assigns an unseen sample to a class c according to its a-priori probability $p(c)$. The accuracy α of such a classifier can be estimated as follows:

$$\alpha = \sum_{c \in \mathcal{C}} p(c)^2, \tag{2}$$

where $\mathcal{C}$ is the set of all classes (depression and control in the experiments of this work).

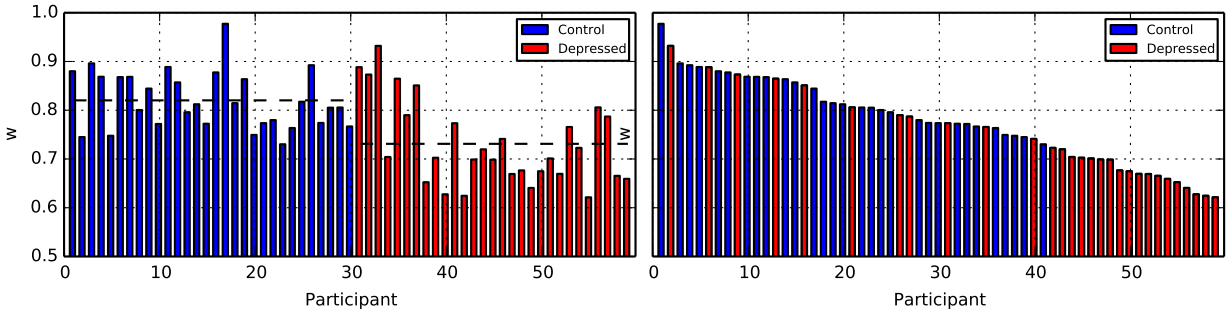

Figure 2: *The left chart shows the w ratio for all participants (the horizontal dashed lines correspond to the average w values for control and depressed participants). The right chart shows the same w values in descending order.*

According to a two-tailed t-test ($p < 0.01$), the best performance at the clause level can be observed in correspondence of the unimodal paralanguage-based approach. However, it is the multimodal combination of language and paralanguage that leads to the best results at the person level, the one that actually matters from an application point of view. In particular, the multimodal approach improves over the best unimodal approach to a statistically significant extent ($p < 0.01$ according to a two-tailed t-test). This seems to suggest that the two unimodal classifers are *diverse*, i.e., they tend to make different mistakes over different samples [21].

One possible explanation behind the diversity observed above is that participants belonging to a given class tend to manifest their condition through one modality, while those belonging to the other class tend to do it through the other modality. For this reason, the left chart in Figure 2 shows, for every participant, the value of the ratio $w = w_l/w_p$, where w_l and w_p are the weights that the GMU assigns to language and paralanguage, respectively. The higher such a ratio, the more the GMU considers language to convey reliable information and vice versa. The value of w is always lower than 1, thus suggesting that paralanguage tends to play a more important role than language in depression detection (at least for the data of this work). However, the average w value for control participants is 0.82, while it is 0.73 for depressed ones. Such a difference is statistically significant ($p < 10^{-5}$ according to a two-tailed t-test) and this suggests that, on average, language plays a more important role in the case of control participants than in the case of depressed ones.

The right chart of Figure 2 shows the w values in descending order and further confirms the observations above. In particular, the chart shows that the lowest 18 values correspond to depressed participants, thus suggesting that roughly two thirds of these latter (18 out of the total 29) can be correctly identified by simply finding the speakers for which w is below or equal to a threshold corresponding to the 18^{th} value from the bottom. In other words, the w value can possibly be used as a confidence score when a speaker is classified as depressed. The remaining 11 depressed speakers distribute roughly uniformly across the rest of the chart. However, it can be observed that 15 of the speakers corresponding to the top 20 w values are non-depressed, thus confirming the tendency of the GMU to assign higher weights to language in the case of control participants.

5. Conclusions

This article presents experiments on automatic depression detection, a task that was performed with accuracy up to 83% (F1 score 80%) over a corpus of 59 interviews involving both depressed and non-depressed speakers. In addition, the article shows the analysis of the weights that a Gated Multimodal Unit [9] attributed to language and paralanguage, the two modalities used in the experiments. The goal was to identify the modality that contributes most to depression detection and the results show that, at least for the data used in this work, it is paralanguage to consistently be assigned the higher weight. One possible explanation is that the proposed approach is based on the recognition of clauses, sentences that include only a few words (less than 10, on average). Therefore, the input texts might be too short for text modeling approaches to achieve their best results. However, the most interesting observation is that the ratio w between the weights of language and paralanguage is higher, to a statistically significant extent, in the case of non-depressed speakers. This suggests that the role of language is likely to be more important in the case of control participants than in the case of depressed ones.

One of the most important consequences of the observations above is that the two modalities appear to be a source of *diversity*, the tendency of different classifiers to make different mistakes [21]. Such a property was shown to increase the chances of classifier ensembles [22] to outperform their best members [23]. Therefore, in the experiments of this work, diversity across modalities might be at the origin of the significant performance difference between the multimodal approach and the best unimodal recognizer (83.0% vs 74.6% in terms of accuracy). In this respect, the main question seems to be not whether there is a modality that is better than the others (like the state-of-the-art in Section 1 seems to suggest), but whether it is possible to find multiple modalities that can correct each other when one or some of them do not carry reliable information. Furthermore, the experiments of this work suggest that the modality carrying the most reliable information can be different for people belonging to different classes. This further confirms that the best strategy is not necessarily looking for the best modality, but for a set of modalities that cover all groups of people appearing in the data.

6. Acknowledgements

The research leading to these results has received funding from the project ANDROIDS funded by the program V:ALERE 2019 Università della Campania "Luigi Vanvitelli", D.R. 906 del 4/10/2019, prot. n. 157264,17/10/2019. The work of Alessandro Vinciarelli was supported by UKRI and EPSRC through grants EP/S02266X/1 and EP/N035305/1, respectively.

7. References

[1] WHO Document Production Services, "Depression and other common mental disorders," World Health Organization, Tech. Rep., 2017.

[2] P. Wang, M. Angermeyer, G. Borges, R. Bruffaerts, W. Chiu, G. De Girolamo, J. Fayyad, O. Gureje, J. Haro, Y. Huang, R. Kessler, V. Kovess, D. Levinson, N. Yoshibumi, M. Oakley Brown, J. Ormel, J. Posada-Villa, S. Aguilar-Gaxiola, J. Alonso, S. Lee, S. Heeringa, B. Pennell, S. Chatterji, and T. Bedirhan Üstün, "Delay and failure in treatment seeking after first onset of mental disorders in the world health organization's world mental health survey initiative," *World Psychiatry*, vol. 6, no. 3, pp. 177–185, 2007.

[3] J. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," in *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition*, S. Oviatt, B. Schuller, and P. Cohen, Eds. ACM Press, 2018, vol. 2, pp. 375–417.

[4] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Transactions on Affective Computing (to appear)*, 2019.

[5] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

[6] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[7] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[8] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Verlag, 2012.

[9] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. González, "Gated Multimodal Units for information fusion," arXiv, arXiv:1702.01992, 2017.

[10] J. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 11–18.

[11] M. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *proceedings of the IEEE Spoken Language Technology Workshop*, 2016, pp. 136–143.

[12] N. Aloshban, A. Esposito, and A. Vinciarelli, "What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech," *Cognitive Computation (to appear)*, 2021.

[13] M. Rohanian, J. Hough, and M. Purver, "Detecting depression with word-level multimodal fusion," in *Proceedings of Interspeech*, 2019, pp. 1443–1447.

[14] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proceedings of Interspeech*, 2018.

[15] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech," *arXiv preprint arXiv:1909.07208*, 2019.

[16] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Takefuji, "Wikipedia2Vec: An optimized implementation for learning embeddings from Wikipedia," *arXiv preprint arXiv:1812.06280*, 2018.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[18] 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.

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

[20] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *CoRR*, vol. abs/1103.0398, 2011. [Online]. Available: http://arxiv.org/abs/1103.0398

[21] R. Ranawana and V. Palade, "Multi-classifier systems: Review and a roadmap for developers," *International Journal of Hybrid Intelligent Systems*, vol. 3, no. 1, pp. 35–61, 2006.

[22] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[23] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.