E8-IJS@LT-EDI-ACL2022 - BERT, AutoML and Knowledge-graph backed Detection of Depression

Ilija Tavchioski[*]
Jožef Stefan Institute
Jamova 39, Ljubljana
`ilijatavchioski@gmail.com`

Boshko Koloski[*]
Jožef Stefan Institute
Jožef Stefan Institute IPS
Ljubljana, Slovenia
`boshko.koloski@ijs.si`

Blaž Škrlj[*]
Jožef Stefan Institute
Jamova 39, Ljubljana
`blaz.skrlj@ijs.si`

Senja Pollak
Jožef Stefan Institute
Jamova 39, Ljubljana
`senja.pollak@ijs.si`

Abstract

Depression is a mental illness that negatively affects a person's well-being and can, if left untreated, lead to serious consequences such as suicide. Therefore, it is important to recognize the signs of depression early. In the last decade, social media has become one of the most common places to express one's feelings. Hence, there is a possibility of text processing and applying machine learning techniques to detect possible signs of depression. In this paper, we present our approaches to solving the shared task titled *Detecting Signs of Depression from Social Media Text*. We explore three different approaches to solve the challenge: fine-tuning BERT model, leveraging AutoML for the construction of features and classifier selection and finally, we explore latent spaces derived from the combination of textual and knowledge-based representations. We ranked 9th out of 31 teams in the competition. Our best solution, based on knowledge graph and textual representations, was 4.9% behind the best model in terms of Macro F1, and only 1.9% behind in terms of Recall.

1 Introduction

Depression is a type of mental illness that affects a large part of our society and is one of the most complex challenges facing our humanity. Since depression is a disease that, if left untreated, can lead to serious consequences such as suicide over time, its early detection is crucial. Since people with depression typically do not open up in person very often, they often see social media as a way to express their thoughts and feelings (Steger and Kashdan, 2009). This trend increased rapidly with

the COVID-19 pandemic due to restrictive measures that encouraged people to use social media as a means of expression. As the number of posts on social media has increased rapidly in recent years, there is a need to process them automatically to extract valuable information such as signs of depression. For this task, detecting signs of depression is a standard multi-class classification problem, where each post can be assigned to one of three classes ("non-depressive", "moderate", and "severe"). Increasing predictive accuracy may be critical for psychiatrists to detect the early signs of major depression to prevent further consequences. The remainder of this article is organized as follows: in Section 2, we discuss approaches to solve multi-class text classification problems and related work with data sets of social media posts on depression, in Section 3, we present the statistics of our given data, in Section 4, we explain the methods we used to solve the given problem, in Section 5, we present and analyze the results, and in Section 6, we provide the conclusion and present our plans for future work.

2 Related work

In related work, we can find various approaches to detecting depression from textual data including various social media.

One of the frequently used sources is **Twitter.** In one of the earlier studies of data from Twitter users who have attempted to take their life, Coppersmith et al. (2016) propose a logistic regression classifier using character *n*-gram character features. Leis et al. (2019) tackled the task of detecting depression in Spanish tweets and used occurrences of negative and positive words determined by Spanish Sentiment Lexicon (Pérez-Rosas et al., 2012) and

---

[*] Equal contribution.

the Spanish SentiCon Lexicon (Cruz et al., 2014) as additional features. For Arabic, Almouzini et al. (2019) addressed the task of classifying Twitter posts as depressed or non-depressed. After text preprocessing and cleaning, the authors extract sparse features from tweets to construct feature vectors. The classification is done using four popular models: Random Forest (Liaw and Wiener, 2001), Naïve Bayes (Shukla and Shukla, 2015), AdaBoostM1 (Wang et al., 2014) and LibLinear SVM (Fan et al., 2008). Recently, there were also deep learning based approaches proposed for depression detection on Twitter data. For example, Mathur et al. (2020) use a Bidirectional Long Short Term Memory Recurrent Neural Network with Attention (Zhou et al., 2016) which produces a high performance on a data set consisting of over 30,000 English tweets. In the last years, also multimodal approaches have been explored. For example, Gui et al. (2019) propose the combination of visual and textual information in order to achieve better results. They model the problem as Markov Decision Process, solving it in reinforcement learning manner. For the text feature extraction, they consider learning custom embeddings via a bidirectional GRU network, for images they used the pre-trained VGGNet (Simonyan and Zisserman, 2014).

Another source of data is **Reddit**, as in our shared task. Trifan et al. (2020) use Term-Frequency Inverse-Document-Frequency features and various classification models, such as Support Vector Machine with Stochastic Gradient Descent and Multinominal Naive Bayes (Ahmed and Ghafir, 2019). Reddit was also the source of a study by Wolohan (2020) about the linguistic characteristics of depression during global pandemics. The authors analyze the increase of depression on the social platform and model the problem with FastText (Bojanowski et al., 2016) embeddings and employ LSTM networks to learn to detect depression in the data.

Next, the authors have also used **Facebook** data. For example, Wu et al. (2018) decided to generate content features by LSTM Neural Network expanding the quantity of information represented in the feature vectors. Merging these features with additional content, behavior, and living-based features are used for the construction of a standard deep neural network.

Finally, the authors also use **blogs**. Yuka Niimi (2021) tackled the problem of depression detection

Label	Train Set	Development Set
Not Depressed	1971 (22 %)	1830 (41 %)
Moderate	6019 (68 %)	2306 (51 %)
Severe	901 (10 %)	360 (8 %)
Size	8891	4496

Table 1: Label Distribution

in Japanese blogs. They firstly filter out the documents without significant topics via LDA and later produce LSA representation of the space and apply SVD to build classifiers.

3 Data description

The data set (Kayalvizhi et al., 2022) that is provided by the task's organizers consists of English posts from the Reddit social media platform, which includes more textual data compared to other social media platforms (Kayalvizhi and Thenmozhi, 2022). The posts belong to one of three given classes: "not depressed", "moderate" and "severe".

We use three data splits one for training, one for development, and one for testing. We used the development set for the internal evaluation of various models.

4 Methodology

In the following section, we will present the methods that were used along with the evaluation measures. For the given task we have developed three independent methods.

4.1 BERT

We opted to fine-tune large pre-trained models based on BERT (Devlin et al., 2018), which often produce *state-of-the-art* results for various tasks. We tested several pre-trained BERT variants that we then fine-tune on the depression detection data provided by the organizers. We investigate **BERT End to end**, which is the base BERT model. Next, we experiment with a faster and smaller **distil-BERT** model (Sanh et al., 2019).Finally, we consider **RoBERTa** (Liu et al., 2019), which is a robustly optimized BERT pretraining approach trained over more data and on longer sequences.

For official submission, we opted for RoBERTa model with a train batch size of 32 in 10 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) which is the Adam optimizer (Kingma

and Ba, 2015) enriched with weight decay, also it is worth to notice that this particular model is case sensitive. We choose to use RoBERTa model over the other two BERT distributions due to its larger pretraining data and better performance when evaluated on the development set.

4.2 autoBOT

In our work for the second method we have considered for Automated Machine Learning, more precisely autoBOT (Automated Bag of Tokens) proposed by Škrlj et al. (2021). autoBOT is a system that can learn from different document representations while iteratively re-weighting the joined representation space. The core of the autoBOT system is the representation evolution, in which by re-weighting different document representations, including token, sub-word, and sentence-level features (contextual and non-contextual) the system is obtaining the final representation for the given task. There are two user inputs for this system: the amount of time for evolution and the kind of document representation. For our task, we have used autoBOT's configuration that it is using both symbolic and sub-symbolic features. The symbolic features are a set of features that are based on words, characters, part-of-speech tags, and keywords. The sparsity parameter for this configuration was 0.05 which implies that the dimension of symbolic subspaces would be 10,250, because the default dense dimension is set to 512 and the sparsity presents the quotient of dense dimension and final dimension. We set the time constraint to 8 hours.

4.3 Knowledge Graphs

Knowledge-backed representation of documents has proven to be useful in text classification tasks (Koloski et al., 2022). We explore how these representations perform in the problem of the detection of depression. We first follow the original idea of the authors and generate standalone text and knowledge graph based representations.

4.3.1 Knowledge-graph features

We use the WikiData5m (Wang et al., 2021) dataset and match the concepts appearing both in the documents and the KG. Based on which representations the data catches, we utilize 6 different knowledge graph representations: transE (Bordes et al., 2013), rotatE (Sun et al., 2019), complEx (Trouillon et al., 2016), distmult (Yang et al., 2015), simplE (Kazemi and Poole, 2018), and quate (Zhang

et al., 2019). We generate them from the pretrained embeddings with the GraphVite library (Zhu et al., 2019). The distribution of most-frequent concepts is shown in Figure 1.

4.3.2 Textual features

In order to generate textual representations we consider using two different type of representations, based on the ones used in (Koloski et al., 2021):

Latent Semantic Analysis: The original implementation first generates n-grams of word and character with maximum of $n - feat$ features and then applies TruncatedSVD to reduce them to $dims$. We create a grid of *n-feats* and *dims*:

- *n-feat* $\in [2500, 5000, 10000, 15000]$

- *dims* $\in [128, 256, 512, 768]$

Contextual Features: We use the distilBERT (*distilbert-base-nli-mean-tokens*) (Sanh et al., 2019) implemented as sentence-transformers (Reimers and Gurevych, 2019).

4.3.3 Learning of intermediate representations

We use two strategies for merging the aforementioned representations:

- **CN** - Concatenation and normalization: we simply concatenate the generated KG and textual features, next we normalize them and finally search for a linear classifier.

- **DR** - Dimensionality reduction: we first concatenate and normalize the given representations and later apply SVD (Halko et al., 2010) to obtain a new latent-space on which we later we learn a new classifier. We search for a new space in *dims* $\in [128, 256, 512, 768, 1024, 2048]$.

4.3.4 Classifier selection

We decided for a linear classifier, based on Stochastic Gradient Descent optimizing two different loss functions *hinge* and *log*, while penalizing *elasticnet* with *alpha* $\in [0.01, 0.001, 0.0001, 0.0005]$, *l1_ratio* $\in [0.05, 0.25, 0.3, 0.6, 0.8, 0.95]$ and *power_t* $\in [0.5, 0.1, 0.9]$. We performed *10-fold* cross-validation search of the grid to obtain the best-performing model on the training split of the data.
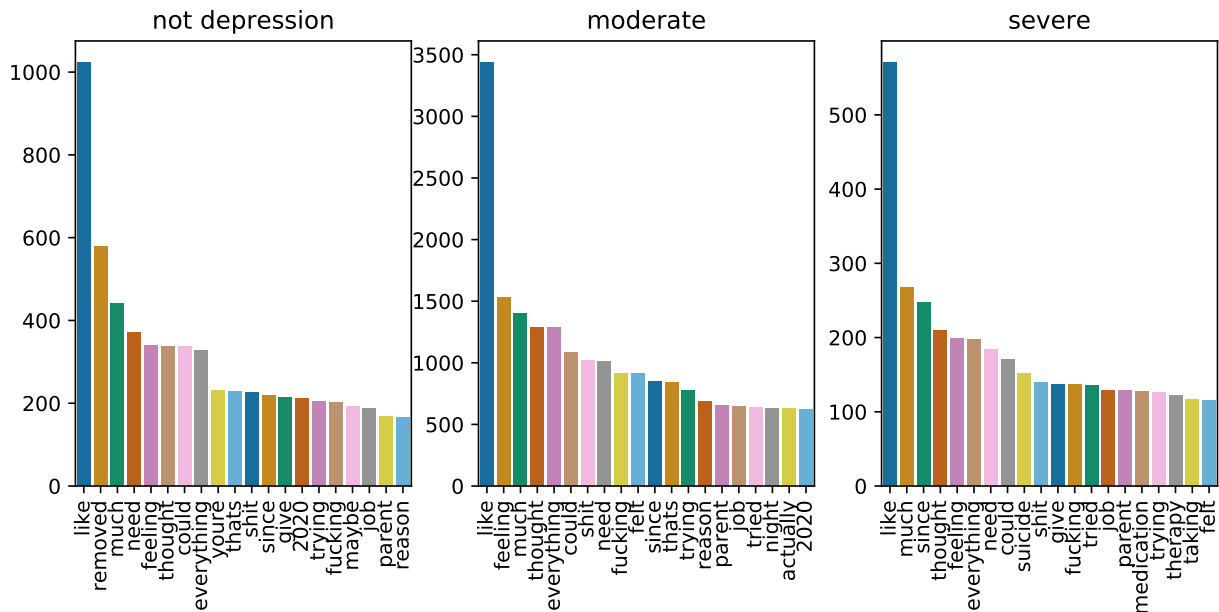
Figure 1: Distribution of the most-frequent concepts found in the WikiData5m knowledge graph and in the documents, grouped by the corresponding level of depression.

4.3.5 Final model configuration

For the final model we have configured the combination of all of the 6 KG (6 x $512dims$) representations, LSA model from *n-feats* $= 10.000$ reduced to 512 dimensions and the sentence-transformer variant of distilBERT 768. We chose the **DR** type with the final dimensions reduced from 4352 to 512 dims via SVD. The best-performing classifier on this model was based on *log* loss function with *alpha*$= 0.001$ and *l1-ratio* of 0.08, with *power-t* of 0.05.

5 Evaluation

In this section, we represent the evaluation of our proposed approaches. We first showcase the evaluation of the development data set, followed by the evaluation of methods on the test split.

5.1 Evaluation measures

For evaluation of the methods, we used the measures that were proposed by the authors of the shared task. These include: accuracy, macro averaged recall, macro averaged precision and macro averaged F1-score.

5.2 Internal evaluation

In this subsection we describe the internal evaluation used in our approach.

5.2.1 Baseline methods

In order to evaluate the performance of our methods we introduce several baselines:

- **majority** Assign the class that has the majority in the data set.

- **char-ngrams** Best 1000 Tf-IDf features of char bigrams, trigrams, and quadgrams, in terms of term frequency.

- **word-ngrams** Best 1000 Tf-IDf features of word unigrams, bigrams, and trigrams, in terms of term frequency.

- **doc2vec** Doc2Vec (Lau and Baldwin, 2016) embeddings with vector's size of 512 and window's size of 5.

5.2.2 Results

In the Table 2 we present the results of various methods when trained on the training data and evaluated on the development set. One can see that from the BERT-based approaches, RoBERTa outperforms BERT-e2e approach. In terms of autoBOT we run only one configuration with maximum time execution of 2 hours, while for KG approach using dimensionality reduction leads to substantially better results than when using concatenation and normalization. The best setting of each method was selected for final evaluation on the official test set. At the internal evaluation, autoBOT

254

Method	Approach	Accuracy	Macro F1
baselines	majority	0.5129	0.2260
baselines	char-ngrams	0.5650	0.3727
baselines	word-ngrams	0.5472	0.3684
baselines	doc2vec	0.4466	0.4025
BERT	distlBERT	0.5261	0.4880
BERT	BERT-e2e	0.5476	0.5034
BERT	RoBERTa	0.5634	0.5287
autoBOT	autoBOT-2h	0.5723	0.5276
KG	CN	0.5827	0.4401
KG	DR	**0.7341**	**0.8627**

Table 2: Performance evaluation of models on the development set, measured by the accuracy and macro F1-score. The *Method* column represents the type of method as defined in *Section 4*. The *Approach* column represents the representation with respect to the method from *Method* column.

and RoBERTa methods achieved similar results in terms of macro-F1 metric, while the Knowledge Graph method outperformed them with a margin of 34% in terms of macro-F1 metric.

5.3 Evaluation on the official test set

In the following Table 3 we show the results obtained by each of our methods on the competition's official test set.

As expected based on our evaluation of the development set, the method based on the knowledge graph and document representations achieved the best performance in almost all of the evaluation metrics (except for the accuracy where the autoBOT method has better performance). In comparison to the best-performing model of the competition, our team was 4.9% behind the top score in terms of best macro averaged F1-score (ranking 9th/31) and is only 1.9% behind the best Recall score (ranking $5th/31$).

6 Conclusion and future work

In this paper, we explored three different approaches to address the task of detecting depression in a social media text. First, we leveraged the

BERT family models, as they represent the state of the art for many text classification problems. Next, we investigated how AutoML approaches such as autoBOT perform on this task, where features are iteratively generated and selected via evolution strategies, and finally combined in the final representation. Finally, we studied how knowledge-based representations based on knowledge graph concepts that occur in a text and textual representations such as LSA and sentence transformers perform. We show that combined knowledge and textual representations outperform any of our other combinations and perform best. Our proposed solution yields good results in terms of macro F1, but it is *4.9%* behind the leading model. For future work, we propose to develop ensembles of the previously created models to see how the combination of different feature types affects performance. Next, we propose to explore larger (like ConceptNet (Liu and Singh, 2004)) or more specific knowledge graphs (like medical knowledge graph (Li et al., 2020)) to improve the results. Finally, we propose the use of feature importance methods to see how outcomes are affected by each family of features and what insights this can give us about depression.

7 Availability

The code is available at https://gitlab.com/tavchija/acl-depression-ldi-2022.

Acknowledgements

Method	Accuracy	Recall	Precision	Weighted F1-score	Macro F1-score
First Method (BERT)	0.5208	0.4953	0.5146	0.5360	0.4738
Second Method (autoBOT)	**0.6407**	0.4525	0.3721	0.5869	0.3680
Third Method (Knowledge Graphs)	0.6015	**0.5714**	**0.5149**	**0.6140**	**0.5334**

Table 3: Final evaluation of the scores, of the best-performing models. We have submitted the best-performing model from each group of methods.

References

Feroz Ahmed and Shabina Ghafir. 2019. Linear support vector machine (svm) with stochastic gradient descent (sgd) training and multinomial nave bayes (nb) in news classification. *International Journal of Computer Sciences and Engineering*, 7:360–363.

Salma Almouzini, Maher Khemakhem, and Asem Alageel. 2019. Detecting arabic depressed users from twitter data. *Procedia Computer Science*, 163:257–265.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117, San Diego, CA, USA. Association for Computational Linguistics.

Fermín L. Cruz, José A. Troyano, Beatriz Pontes, and F. Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems With Applications*, 41(13):5984–5994.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: a library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 110–117.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2010. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.

S Kayalvizhi and D Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *arXiv preprint arXiv:2202.03047*.

S. Kayalvizhi, D. Thenmozhi, B. R. Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4289–4300.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Boshko Koloski, Timen Stepišnik-Perdih, Senja Pollak, and Blaž Škrlj. 2021. Identification of covid-19 related fake news via neural stacking. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 177–188, Cham. Springer International Publishing.

Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. 2022. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *CoRR*, abs/1607.05368.

Angela Leis, Francesco Ronzano, Miguel A Mayer, Laura I Furlong, and Ferran Sanz. 2019. Detecting signs of depression in tweets in spanish: Behavioral and linguistic analysis. *J Med Internet Res*, 21(6):e14199.

Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, and Yuting Liu. 2020. Real-world data medical knowledge graph: construction and applications. *Artificial Intelligence in Medicine*, 103:101817.

Andy Liaw and Matthew Wiener. 2001. Classification and regression by randomforest. *Forest*, 23.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

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

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Puneet Mathur, Ramit Sawhney, Shivang Chopra, Maitree Leekha, and Rajiv Ratn Shah. 2020. Utilizing temporal psycholinguistic cues for suicidal intent estimation. In *Advances in Information Retrieval*, pages 265–271, Cham. Springer International Publishing.

Verónica Pérez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3077–3081, Istanbul, Turkey. European Language Resources Association (ELRA).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Ashish Shukla and Shweta Shukla. 2015. A survey on sentiment classification and analysis using data mining. *International Journal of Advanced Research in Computer Science*, 6(7).

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Blaž Škrlj, Matej Martinc, Nada Lavrač, and Senja Pollak. 2021. autobot: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*.

Michael F Steger and Todd B Kashdan. 2009. Depression and everyday social activity, belonging, and well-being. *J. Couns. Psychol.*, 56(2):289–300.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alina Trifan, Rui Antunes, Sérgio Matos, and Jose Luís Oliveira. 2020. Understanding depression from psycholinguistic patterns in social media texts. In *European Conference on Information Retrieval*, pages 402–409. Springer.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org.

Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, and Jibao Gu. 2014. Sentiment classification: The contribution of ensemble learning. *Decision support systems*, 57(1):77–93.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

JT Wolohan. 2020. Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Min Yen Wu, Chih-Ya Shen, En Tzu Wang, and Arbee L. P. Chen. 2018. A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, 54:225–244.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yutaka Miyaji Yuka Niimi. 2021. Machine learning approach for depression detection in Japanese. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 350–357, Shanghai, China. Association for Computational Lingustics.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. 2019. Quaternion knowledge graph embeddings. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2731–2741.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *CoRR*, abs/1611.06639.

Zhaocheng Zhu, Shizhen Xu, Meng Qu, and Jian Tang. 2019. Graphvite: A high-performance cpu-gpu hybrid system for node embedding. In *The World Wide Web Conference*, pages 2494–2504. ACM.