MasterMind: A Novel Multi-Output Model Approach to Detecting Mental Illnesses through Natural Language Processing

Kabilan Prasanna

Lightridge High School

Virginia

Abstract

Identifying and diagnosing mental health conditions such as depression, anxiety, and suicidal thoughts is growing in importance, with Natural Language Processing (NLP) emerging as a viable key to do so. By utilizing publicly available, labeled data obtained from platforms like X (formerly Twitter) and Reddit, this research focuses on developing individual NLP models using Naïve Bayes, Random Forest, XGBoost, and KNN Classifiers. These models are then integrated into a comprehensive multi-output model, aimed to optimize efficiency. Although the individual models displayed a tendency to overfit, they demonstrated high accuracy and effectiveness in mental health detection, specifically the Random Forest Classifier. The multi-output model, although not surpassing the individual models in performance, still showed high accuracy as well. This research underscores the potential of NLP in mental health analysis and its practical application in medical and psychiatric practices to accurately address the needs of those showing signs of these conditions in their writing.

MasterMind: A Novel Multi-Output Model Approach to Detecting Mental Illnesses through Natural Language Processing

Mental health is not a silent struggle; in recent years, the awareness of mental health conditions and their profound impact on family, friends, and society is increasing as those suffering under these conditions, like depression, anxiety, and suicidal thoughts, have only grown in size. In a 2023 survey conducted by Gallup, the percentage of adults who have depression in the United States grew roughly 10% since 2015 at a whopping 29%, or almost 1 in every 3 American adults (Witters, 2023). Additionally, in 2021, the CDC estimated roughly 5.2 million Americans planned and/or attempted suicide, while almost 12.3 million Americans seriously thought about it (CDC, 2023). Similarly, roughly 1 in every 3 Americans (31.2%) are also predicted to have an anxiety disorder sometime throughout their life (NIH, n.d.).

In a world that is becoming increasingly dependent on social media and other forms of digital communication, the texts and words posted online can have great insight into the emotions and mental state of the person they're from. To utilize this fact, as the field of Natural Language Processing (NLP) grows, this research utilizes various techniques such as Naïve Bayes, Random Forest, and other classifiers have proven useful in detecting sentiment, which can be applied to the detection of these mental conditions as well. Can a multi-output model and other techniques trained on this textual data, derived from social media sources such as X (formerly Twitter) and Reddit, be used to accurately predict the presence of depression, suicidal thoughts, and anxiety in text?

By combining the results of three separate optimized models into data for a final multi-output model, this research hopes to enhance the accuracy and precision of the model and compile a comprehensive assessment of what the text may indicate about the person who said it. Furthermore, being able to predict and identify mental health conditions before they have a major impact is an incredibly helpful ability, and can offer an opportunity to intervene and support individuals who may be suffering. It can not only be used personally, but by concerned parents, counselors, friends, and other individuals in the person's life. The possibilities and applications of

this research are endless.

Related Work

Machine Learning. There have been machine learning models created in the past in order to predict the presence of mental health crises through the usage of electronic health records, such as the work completed by Garriga et al. (2022). Similarly, work using facial recognition with convolutional neural networks (CNNs) have also been used to predict the presence of mental health issues, such as the research completed by Shafiei et al. (2020).

Natural Language Processing and Sentiment Analysis. The precursor to most classification models using NLP is sentiment analysis, which classifies textual data into having positive, negative, or neutral sentiment. However, this led to other applications of sentiment analysis past the emotion expressed in the sentence, such as the detection of conditional sentences done by Narayanan et al. (2009), the usage of sentiment analysis to analyze relationships between financial risk and the language used in financial reports by Wang et al. (2013), or work done for other languages like the detection of sarcasm in Arabic by Abu Farha and Magdy (2020) and El Mahdaouy et al. (2021). These additional uses of classification through NLP paved the way for mental health detection.

Natural Language Processing and Mental Health. Due to the digital presence of individuals with mental health conditions on the Internet, much research goes into the usage of collecting data from publicly available sources of textual information, namely social media. Platforms such as X (formerly Twitter), Facebook, and Reddit contain millions of lines of data from people sharing their experiences with mental conditions and their emotions through text, allowing them to be great sources to pull data from (CALVO et al., 2017). A review done by Malgarejo et al. (2023) shows how detection of mental health conditions (usually by individual conditions) can be used to intervene and promote psychological well being.

Variables

In this study, the independent variables are the specific NLP models used for each mental condition, namely the NaïveBayesClassifier, RandomForestClassifier,

K-Nearest-NeighborsClassifier, and XGBoostClassifier models. The dependent variable is the overall model performance, including the precision, accuracy, F1 score, and recall for the individual models and the area under the receiver operating characteristic (ROC) curve for the multi-output model. Additionally, the overall effectiveness of the final multi-output model in classifying text generated on the spot is a qualitative dependent variable being analyzed.

To complete this research, standard machine learning Python libraries and additional libraries containing specific models were necessary, including:

- nltk
- numpy
- seaborn
- scikit-learn (sklearn)
- pickle
- pandas
- matplotlib
- xgboost

Data Preparation

For natural language processing, dataset quality is one of the most important factors that contribute to the model's accuracy and overall results. With blurred or ambiguous data points, it can be hard for the models to precisely label a sentence with the correct tag. Furthermore, the size of the dataset can significantly affect the model as well, since it will be exposed to different styles of writing, allowing for less overfitting. Therefore, it is important to select and use datasets that not only consist of meaningful sentences, but ones that can optimize the model's efficiency.

Dataset Collection

The data for each of the models trained was collected from a publicly available source, specifically Kaggle. Many of the sentences collected for Kaggle come from social media websites, such as various subReddits relating to the mental illness being detected as well as X (formerly Twitter). Taking data from general media addresses the ethical dimension of mental health prediction, emphasizing the responsible handling of this sensitive information in a private and anonymous way.

These datasets were specifically handpicked for their pre-labeled data: for example, the depression dataset was already split into depressed (represented by the positive Boolean value 1)

and not-depressed tweets (represented by the negative value 0), which is necessary when training any machine learning model. Additionally, the datasets range over a large timeframe, especially those consisting of posts scraped from Reddit. For example, the dataset used for suicidal thoughts consisted of posts ranging from January 1, 2009, to Jan 2, 2021.

As the computer used was not capable of running models on hundreds of thousands of lines of data, most of the datasets were truncated at 50,000 lines or less. See Table 1 summarizing the number of lines for each individual model (excluding the final multi-output model).

Table 1

Number of lines in each dataset used for the various mental conditions being classified. The number in parenthesis indicates the specific trial number of that mental condition's experiment.

Dataset Overviews	
Mental Condition	Number of Lines
Depression	16,035
Anxiety (1)	50,000
Anxiety (2)	42,706
Suicidal Thoughts (1)	30,000
Suicidal Thoughts (2)	37,666

Dataset Pre-Processing

The datasets used in each model underwent basic natural language processing pre-preprocessing steps. In chronological order, the following changes were applied:

1. Converting all characters to lowercase
2. Removing all non-letter characters using RegEx
3. Removing all stop words (imported from the Natural Language Toolkit, or NLTK)
4. Tokenizing the data points using NLTK
5. Lemmatizing the data points using NLTK
6. Vectorizing the data

Tokenizing refers to the process of splitting data into multiple tokens, which can either be words, sentences, or sub-words. In this case, the sentences were split into words. After the tokenization was complete, each of the tokens were lemmatized. Lemmatization is the process of stripping words down to their most basic form, which includes removing suffixes or undoing verb conjugation. Both of these methods, alongside the other steps used in text pre-processing, allow the model to rely less on smaller, repetitive morphemes and more on important words (Khyani et al., 2010). After tokenization and lemmatization, the data is vectorized into a machine readable format and the vectorizer settings used for that specific dataset are stored for later usage.

Modeling and Experimentation

Individual Models

After the vectorization is complete, each trial for the individual depression, anxiety, and suicidal thoughts model had a similar format to experiment with the Voting Classifier, Random Forest Classifier, XGBoost Classifier, and K-Nearest-Neighbors Classifier (KNN).

To find the best model for detecting depression, a Voting Classifier, Random Forest Classifier, and XGBoost Classifier were used. The Voting Classifier used a Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and Multinomial Naïve Bayes model as estimators, all taken from the scikit-learn library. For the voting style, the "soft" approach was taken, in which each estimator returns a probability rather than boolean. The final prediction is made from an average of these three predictions, rather than simply taking the majority boolean value, allowing for less overfitting and bias. The Random Forest classifier used the random state 1 to recreate results, alongside 100 estimators, also known as the "trees" of the model. As this classifier is made up of trees trained on randomized subsets of data, there is diversity and, consequently, less overfitting like the Naïve Bayes models (Pedregosa et al., 2011). Lastly, the XGBoost classifier was also created with the random state 1 and 100 estimators. This classifier is built on the idea of gradient boosting and also uses decision trees as its base model, similar to the Random Forest Classifier. Unlike the latter, however, the model builds a new tree that predicts the gradients of the loss with respect to the current model as it runs gradient descent to minimize the loss. Its penalization of

complex models by pruning trees that contribute the least while running other regularization methods (such as L1 and L2 regularization) allows the XGBoost model to reduce overfitting as well (Chen & Guestrin, 2016).

There were two separate trials done for the suicidal thoughts model, each using the same three models but with a variation in the data they were trained on. Similar to the depression model, a Voting Classifier using scikit-learn's Gaussian, Bernoulli, and Multinomial Naïve Bayes models were used as estimators, with the voting set to "soft." Additionally, a Random Forest Model was also trained, using random state 1 but only five estimators in comparison to the depression model's 100. Initially, the first trial used a Random Forest model trained with 100 estimators, but it was found to have overfit to the training data. Lastly, a K-Nearest Neighbors model with five estimators was used for the first trial, but not used for the second due to the processing constraints of the computer and the size of the second dataset used. This classifier analyzes the classification of a datapoint's "neighbor" embeddings to make its decision regarding the value of said datapoint based on the majority of classifications surrounding it. Although analyzing a larger number of "neighbors" would increase the accuracy of the model by reducing overfitting, it is computationally expensive: therefore, only five neighbors were chosen to be analyzed for each of the training data points (Pedregosa et al., 2011). Additionally, the Minkowski distance was chosen as the metric to determine these five neighbors, and is useful in NLP applications such as this due to its ability to adapt to vector space models (Grootendorst, 2021).

Like the suicidal thoughts model, two separate trials were also completed for the anxiety classification model, each with varying datasets. For both trials, a Voting Classifier with the Guassian, Bernoulli, and Multinomial Naïve Bayes models was created with "soft" voting. Additionally, a Random Forest Classifier and XGBoost classifier were also created, each with five estimators and at the random state 1.

Multi-Output Model

To experiment with a multi-output model, 5 new datasets were created using various combinations of datasets from the previous individual models. For example, the first trial

consisted of using all three datasets used for the depression, anxiety, and suicidal thoughts classification. Before the three datasets were concatenated together, the best models of the other two conditions were utilized to create two new columns on the original dataset for those conditions; if the anxiety dataset was used (which already contains a column with anxiety/not-anxiety labels), the depression and suicidal thoughts models were used to classify the sentences into their respective condition and attached as two additional columns to the dataset. Afterwards, the updated datasets were all concatenated.

After this had been done with different combinations to create five separate trials, multi-output models were created using a Voting Classifier and Random Forest Classifier for each trial. Unlike the individual models, however, these models were put into scikit-learn's MultiOutputClassifier method to adapt it to the new format.

Results

Individually, each model performed better than the collective multi-output model; however, all scores, including the accuracy, precision, recall, and F1 score, were still considerably high. See Table 2 on the next page for an overview of the scores for the best trials.

For the depression model, the Random Forest Classifier was clearly the best, with a training accuracy of 0.9996, a testing accuracy of 0.89888, and a precision, recall, and F1 in the high 0.80s and low 0.90s. However, as the training accuracy is near perfect and the testing accuracy is relatively much lower, the model shows clear signs of overfitting to the training data.

In both trials done for the suicidal thoughts model, the Random Forest Classifier was also the most accurate. Trial 1 had a training accuracy of near 1, similar to the depression model, but a drop with a testing accuracy of 0.86867. Trial 2, on the other hand, shows less signs of overfitting, with a lower training accuracy of 0.98742 but higher testing accuracy of 0.86989. Additionally, the precision, recall, and F1 scores were all higher in trial 2 as well. This may have been due to the greater variation present in the second dataset used for trial 2 and the slightly larger size.

For the anxiety model, the Random Forest Classifier was once again the most accurate of the three tested. On the other hand, there was a difference in data quality that made the accuracy

and other metrics better in Trial 2 in comparison to Trial 1. For example, the Random Forest Classifier in Trial 1 had a training accuracy of 0.965 but a much lower testing accuracy of 0.6894. For Trial 2, however, there was a training accuracy of 0.97634 but a better testing accuracy of 0.74918, albeit the score is still low. Compared to the other mental conditions, the lexicon used in the anxiety dataset is very ambiguous, and due to the word-independent nature of the models tested, they don't accurately detect these subtle relationships.

Table 2

Various metrics for the most accurate models of each mental condition (all Random Forest Classifiers).

Individual Random Forest Classifiers: Metrics			
Metric	Depression	Suicidal Thoughts	Anxiety
Training Acc.	0.99961	0.98742	0.97634
Testing Acc.	0.89888	0.86989	0.74918
Precision	0.90476	0.87966	0.75341
Recall	0.87297	0.89046	0.75824
F1	0.88858	0.88503	0.75582

For the final model, the general training score, test score, precision, recall, and F1 score cannot be utilized to determine the performance of the model due to its multi-output nature. For this reason, the area under the receiver operating characteristic curve, or ROC curve, is utilized, as a higher area (AUC; area under the curve) indicates great performance (Yang et al., 2021). From the five separate trials of the multi-output model, the greatest ROC-AUC score of 0.93 for the Trial 2 Voting Classifier indicates that model does very well at determining the labels of the testing data, as 0.5 would indicate random chance. However, as the model does not do as well when labeling some ambiguous user-given input, it may have overfit to the textual style present in the training data, similar to what may have occurred to the individual models. See Figure 1 to view the ROC-AUC curve of the model described above.

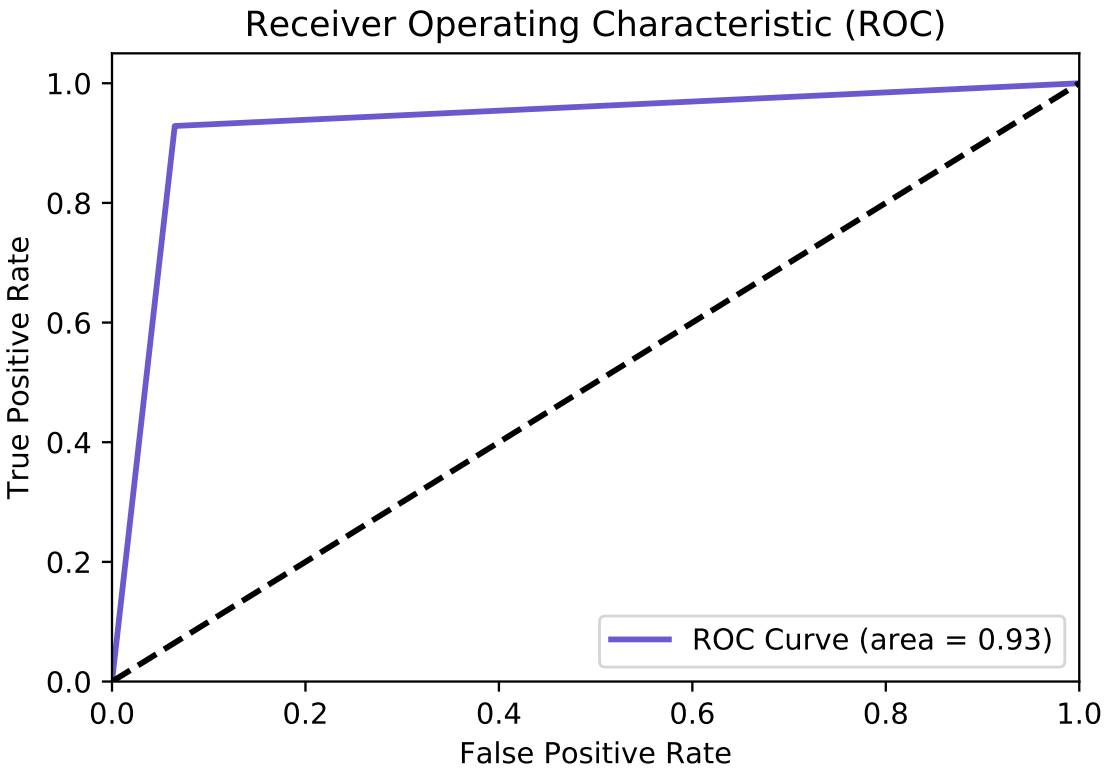

Figure 1

ROC-AUC Curve for the Trial 2 Voting Classifier of the Multi-Output Model. The score of 0.93 indicates high and accurate performance on the testing data.

Conclusion

In this research using Natural Language Processing techniques, three individual models were optimized to classify for their respective mental condition (depression, anxiety, or suicidal thoughts) and used together to create a multi-output model. It was found that a Random Forest Classifier regularly does the best on each individual model, but tends to overfit to the training data. Additionally, it was found that the multi-output model was optimized with a Voting Classifier and did very well according to the ROC-AUC curve, however due to the slight inaccuracies present in the individual models, it could be better.

Overall, the research shows that Natural Language Processing techniques such as Naïve Bayes, Random Forests, XGBoost, and KNN-Classifiers can accurately be used to classify text

into three separate mental conditions using a multi-output format. This has major implications for fields that this is useful for, such as the medical or psychiatric fields; by reducing the three individual models into a singular one, it can save time and be efficient for those who are in need of a quick check in.

Future Work

Improving Current Model Performance

As of now, the NLP techniques utilized do not inherently understand the meaning of the word it's given as training data, but treat each word as an individual feature and operate solely on numerical features. For text classification cases, such as this research, many different variables and assumptions made by the model may impact the accuracy.

For example, the Naïve Bayes models assume each word is independent of those around it, so although text labeled for depression may contain the words "not happy", the model may assume that "happy" is what indicates the text is depressed. Additionally, order is not considered in these models, so different sentences with the same words would be considered the same, even if their meaning is incredibly different (Vikramkumar et al., 2014). The Random Forest and XGBoost Classifiers handle independence a little better; they don't explicitly assume features are independent and are able to capture some relations between the features (Chen & Guestrin, 2016; Louppe, 2015).

Implementing a Language Model

Additionally, the usage of a "frozen" language model such as a frozen RoBERTa language model will be able to further capture the relationships between the words used in the data. As language models are able to consider the semantic meaning of the words, which are stored within the vector embeddings, they would most likely be more accurate in their classification of depression, anxiety, or suicidal thoughts (Lee et al., 2019; Liu et al., 2019).

Expanding to Multiple Languages

This concept can be expanded to multiple languages as well, not just English. Natural language processing, especially language models, have come a long way and many include multilingual capabilities (known as Multilingual Language Models). An example of these publicly available models is the multilingual BERT (mBERT) model released by Devlin et al. (2019), including languages like Basque, Tamil, Chinese, Swahili, and more. This can expand the market significantly and open up the mental health field to those who may not currently have access to it.

Limitations

The size of the methods (Naïve Bayes, RandomForest, etc) utilized in the training of these models often limits their accuracy. For example, much of the data was concatenated from hundreds of thousands of data points to tens of thousands in order to ensure the model does not fail to run at any point throughout the training; however, even with this shortening, the model sometimes failed and the data was cut even more to prevent this from happening again. As the difference between the language used in the training data of each mental illness is very subtle, a larger dataset with more nuance and variety would improve the performance of each model and prevent overfitting.

Applications and Ethics

In the future, mobile or web applications to alert users of possible need for diagnosis can be created, useful for any concerned family members, friends, counselors, or even individuals concerned about themselves. Additionally, this has great use in various fields, including the military, schools, and medical offices.

There are also ethical concerns that should be considered with this application, as some individuals may not feel comfortable sharing their personal data with models like this. To ensure that nobody's rights are breached, consent must be acquired from any user for future applications or trials run from this identified and private source (unlike the public and de-identified data used for training).

References

Abu Farha, I., & Magdy, W. (2020, May). From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, & H. Mubarak (Eds.), *Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection* (pp. 32–39). European Language Resource Association. https://aclanthology.org/2020.osact-1.5

CALVO, R. A., MILNE, D. N., HUSSAIN, M. S., & CHRISTENSEN, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, *23*(5), 649–685. https://doi.org/10.1017/S1351324916000383

CDC. (2023). Facts About Suicide. https://www.cdc.gov/suicide/facts/

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

El Mahdaouy, A., El Mekki, A., Essefar, K., El Mamoun, N., Berrada, I., & Khoumsi, A. (2021, April). Deep Multi-Task Model for Sarcasm Detection and Sentiment Analysis in Arabic Language. In N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, & S. Touileb (Eds.), *Proceedings of the sixth arabic natural language processing workshop* (pp. 334–339). Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.42

Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, *28*, 1240–1248.

Grootendorst, M. (2021). 9 Distance Measures in Data Science. *Towards Data Science*. https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Khyani, D., B, S., N, N., & B, D. (2010). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Journal of University of Shanghai for Science and Technology*. https://jusst.org/wp-content/uploads/2020/10/An-Interpretation-of-Lemmatization-and-Stemming-in-Natural-Language-Processing-1.pdf

Lee, J., Tang, R., & Lin, J. (2019). What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Louppe, G. (2015). Understanding Random Forests: From Theory to Practice.

Malgarejo, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, *13*.

Narayanan, R., Liu, B., & Choudhary, A. (2009, August). Sentiment Analysis of Conditional Sentences. In P. Koehn & R. Mihalcea (Eds.), *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 180–189). Association for Computational Linguistics. https://aclanthology.org/D09-1019

NIH. (n.d.). Any Anxiety Disorder. https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Shafiei, S. B., Lone, Z., Elsayed, A. S., Hussein, A. A., & Guru, K. A. (2020). Identifying mental health status using deep neural network trained by visual metrics. *Translational Psychiatry*, *10*(430).

Vikramkumar, B, V., & Trilochan. (2014). Bayes and Naive Bayes Classifier.

Wang, C.-J., Tsai, M.-F., Liu, T., & Chang, C.-T. (2013, October). Financial Sentiment Analysis for Risk Prediction. In R. Mitkov & J. C. Park (Eds.), *Proceedings of the sixth international joint conference on natural language processing* (pp. 802–808). Asian Federation of Natural Language Processing. https://aclanthology.org/I13-1097

Witters, D. (2023). U.S. Depression Rates Reach New Highs. https://news.gallup.com/poll/505745/depression-rates-reach-new-highs.aspx

Yang, Z., Xu, Q., Bao, S., Cao, X., & Huang, Q. (2021). Learning with Multiclass AUC: Theory and Algorithms.