Machine Learning for Mental Health: A Study of Reddit Discussions and Class Imbalance

Beatriz Billón, Hugo Pérez de Albéniz, Javier Sánchez-Prieto

IE University

November 30, 2024

Introduction

Examining user-generated content on websites like Reddit provides insightful information on language usage and human behavior, particularly in delicate areas like mental health. There are numerous forums on Reddit, referred to as subreddits, where people freely discuss their individual and group experiences with mental health. For people dealing with mental health issues, these internet forums are frequently a vital source of support. By analyzing these discussions, this project aims to provide a deeper understanding of how mental health issues are discussed in online spaces and how machine learning can aid in better identifying and categorizing mental health-related content.

The dataset used for this analysis, The Reddit Mental Health Dataset, was collected from 28 subreddits, including 15 dedicated to specific mental health conditions (such as anxiety, depression, and schizophrenia) and 13 from broader categories, including general mental health support and non-health related topics (Pushshift API, 2020). Over 826,000 posts from more than 826,000 unique users were included, covering discussions from 2018 to 2020. This period includes the onset of the COVID-19 pandemic, making it an opportune time to investigate how global crises might influence online mental health discourse (Smith et al., 2021). This dataset was curated through the Pushshift API, which allowed for a large-scale collection of both post metadata and textual content. By including posts from both mental health-specific and general subreddits, the dataset provides a comparative perspective on the nature of mental health discussions within different community settings.

The features extracted from the dataset for analysis include readability indicators, which measure the complexity and accessibility of posts (Kincaid et al., 1975), and TF-IDF (Term Frequency-Inverse Document Frequency) representations, which identify the most significant terms in the posts by evaluating their frequency and importance across the dataset (Ramos, 2003). These features are essential for uncovering patterns in the language used to discuss mental health and identifying key themes that may reflect different mental health issues.

To achieve the goal of classifying Reddit posts by their associated subreddit, which correlates to specific mental health categories, we applied three machine learning techniques:

1. **Decision Trees**, which create clear and interpretable decision rules for classifying posts based on their text features (Quinlan, 1986).

2. **Gradient Boosting**, which builds an ensemble of decision trees sequentially, where each tree corrects the errors of the previous ones, to reduce bias and variance and improve classification accuracy (Breiman, 2001).

3. **Neural Networks**, which leverage layers of interconnected nodes to capture complex relationships within the data, offering a more flexible approach to classification (LeCun et al., 2015).

We chose these approaches because they have a track record of success in text categorization challenges. Decision trees are perfect for applications where comprehending model decisions is essential since they produce outcomes that are clear and easy to understand. While Neural Networks provide a more advanced method that may identify complicated patterns and non-linear relationships in the data, Random Forests expand on this by mixing numerous decision trees to enhance performance.

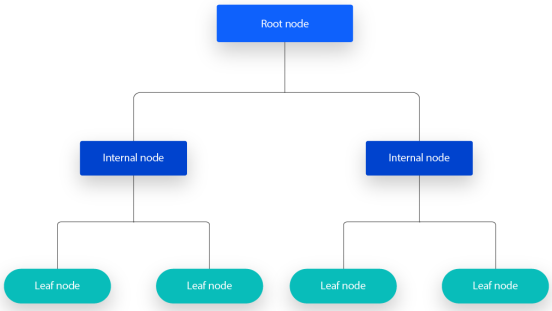
The project's insights are meant to aid in the creation of tools such as therapeutic chatbots that may categorize user inquiries according to their mental health issues. These technologies could provide tailored support by classifying conversations using machine learning algorithms, increasing the relevance and accessibility of mental health resources.

The technique used, results, and difficulties in categorizing Reddit postings will all be covered in this report. Additionally, it will consider the wider ramifications of using machine learning methods on digital mental health data, investigating how these technologies might improve online mental health services.

Background Information

Text classification is the process of classifying text in emails, documents or messages into different categories. We aim to use the text in reddit articles to classify them into different mental health conditions. There are different methods to classify text such as decision trees, gradient boosting and natural networks.

Decision Trees: Decision trees are hierarchical structures that split the feature space based on values. Decision trees start with a root node. A root node represents the entire population data that is being analyzed, hence it has no branches above it. From it, the population is divided into sub-groups (or decision nodes). All the data is divided according to different features. The nodes conduct analyses to split the population heterogeneously based on the features. Once a node cannot be further split (or it's decided not to) it is called a leaf node or terminal node. The leaf nodes represent all the possible outcomes within the dataset.



However, sometimes too much splitting can mean the tree is overfitting the training data and will not perform well on the actual population we want analyzed. Pruning is the process of removing nodes. Pruning is an important aspect of decision trees as it is used to reduce overfitting and optimize the model. There are 2 types of pruning. Pre-pruning consists of stunting the growth of the tree before it is too complex while Post-pruning is the opposite. Pre-pruning helps the tree remain simple and prevent overfitting while post-pruning simplifies the tree while conserving the accuracy. Pruning leads to robust models that perform better on data. Decision trees are clear and easy to understand which make them great applications when the objective is to interpret the data as well as prediction. However, they might prove to be too simple to deal with complex data.

Gradient Boosting: Decision trees are likely to overfit. There are many methods that can be used to make the models more accurate and robust. Boosting consists of various 'weak trees' which perform better than simply random. The combination of many sequential weak learners eventually leads to strong learners as the errors of the previous models are corrected after every iteration. Both random forests and boosting are used to improve the accuracy of a model through individual trees. While random forests train trees independently and aim to get a prediction by joining the results from the trees, boosting algorithms use a different approach. Boosting uses weak learners in a sequence where each model is tasked with correcting the previous tree's mistakes.

Gradient boosted decision trees are a variant of boosting that focuses on gradient descent. Gradient descent, once again, uses weak learners to reach a prediction. The initial tree used in gradient boosting is the base learner. The next decision tree looks at the mistakes the base learner made in order to account for them. All trees are created additively based on the previous mistakes. A residual is the difference between the predicted value and the actual value. Residuals are used to calculate how precise the model is and they are added to score the model with a loss function. A loss function is what is used to measure the precision of a model. Gradient descent is a very effective algorithm to minimize the loss function.

Gradient boosting has many advantages such as being able to handle a lot of data and the process being effective since it can be run in parallel. Gradient boosting is also very effective at preventing overfitting which one of the main issues in text classification.

Neural Networks: Neural network is a model that attempts to mimic human decisions by copying the way neurons work to weight options, recognize patterns and reach conclusions. Neural Networks consist of multiple nodes (which mimic the human neurons), these are located in different layers includ-

ing an input layer, output layer and many more hidden ones. Every node is connected to others and has an individual threshold and weight. The threshold means that if the output of a node is above a specific number the node is activated and it sends the data to the next layer of nodes. If the threshold is not met then the data will not reach the next layer. Each node is essentially a linear regression model with its own data, weight, threshold and output. Neural networks require training to improve the accuracy over time. If a neural network is trained properly it will classify text at a high speed such as with Google.

Methodology

This section details the procedure we followed in our attempt to predict the mental health category of a given message. We applied 3 algorithms (Cost Complexity Pruned Decision Trees, Gradient Boosted Trees, and Neural Networks) to the dataset and utilised hyperparameter tuning to achieve optimal prediction accuracy. All three algorithms were trained on 3 stages of data preparation: 1. The raw data, using all provided NLP and TF-IDF features, 2. class balancing to 4000 posts per target class, 3. reduction of classes with a combined class "Other" of the left out subreddits. The first was discarded due to incredibly poor results accross all models caused by the data imbalance. The second was an improvement but still proved to be too complex for our implementations and did not achieve acceptable accuracy. Thus we opted to simplify the problem with step 3. The remainder of this section will focus on the third stage but we will mention the results of the first two later in the discussion.

Rebalancing Classes

The raw dataset, as mentioned, had an extremely skewed balance of classes (posts per subreddit) as can be seen in Figure 1.
To deal with this we opted to include Imbalanced Learn's (Imbalanced Learn) RandomUnderSample to undersample all classes (excluding "COVID19_support" because it was too small and irrelevant) to 4000 posts. To further simplify the problem the 6 most populated classes were chosen for prediction and all others combined into an equally sampled class named "Other". In this way we are able to more accurately predict the few selected classes ('adhd', 'anxiety', 'bpd', 'depression', 'suicidewatch', and 'mentalhealth') versus the remaining subreddits. The rebalancing procedure was only applied to training data after performing a train-test split with test-size=0.1 (more than sufficient, given the original large number of posts and the loss of samples during undersampling).

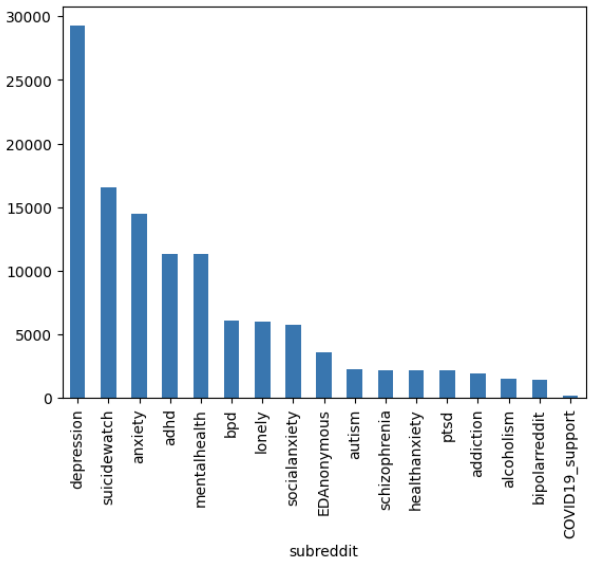


Figure 1: Distribution of Posts by Subreddit

Cost Complexity Pruned Tree

The CCP tree was created using Sklearn's "tree" module, with the "tree.DecisionTreeClassifier" class and corresponding "cost_complexity_pruning_path" function to determine viable values of α (the pruning parameter in the cost function). Said function generated 5077 viable αs, 102 of which were tested and evaluated as per Figure 2.
The optimal value for α is that which maximizes validation accuracy. In this case it is approximately

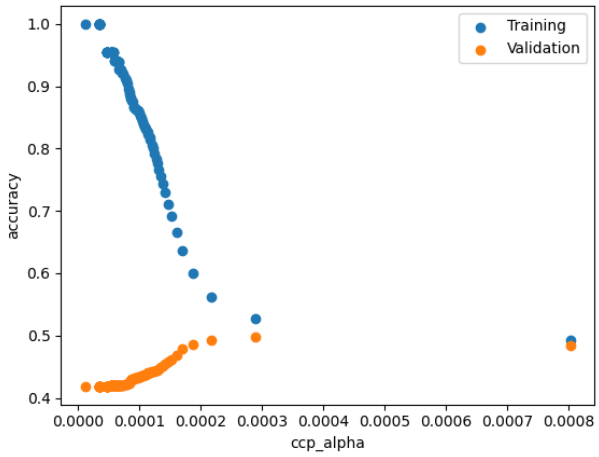


Figure 2: Accuracy of Trees for each CCP value

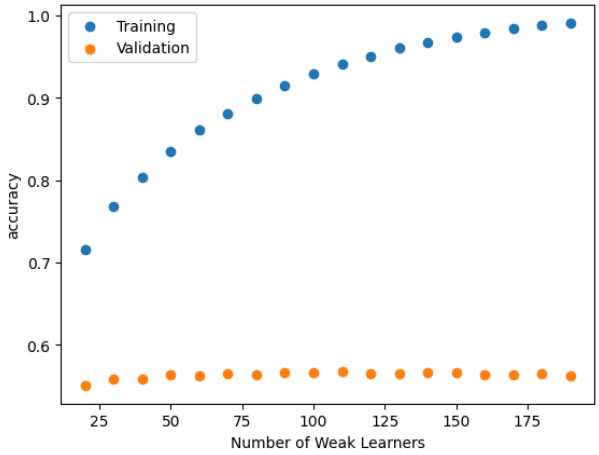


Figure 3: Training and Validation Accuracy - GBTree

2.9×10^{-4} which yielded an accuracy of 45.2% on the test set.

Gradient Boosted Decision Trees

From the CCP tree, seeing that overfitting the data would not be an issue, we opted for trying Gradient Boosting as a means to improve the accuracy of the predictions. As previously explained, Gradient Boosting increases the loss for misclassified points for subsequent estimators, allowing the algorithm to focus on its previous mistakes and slowly approximate an optimum. Because of this way we thought that said algorithm would "Boost" our predictive capability.

Using the XGBoost package (XGBoost) we trained and optimised a Gradient Boosted Decision Tree. After creating a validation split and encoding the feature space to integer values, we tuned the number of weak learners used by the tree, with values ranging from 20 to 200. This yielded the training and validation scores plotted in Figure 3, from which we chose n_estimators to be 50, approximating the elbow of the curve. The corresponding tree yielded an accuracy of 53% on the test data, an improvement from the CCP tree.

Neural Networks

In an attempt to capture more information from the given features and better capture the nuances and complexities of written text, we proceeded to train several Neural Networks using Keras and its parent library TensorFlow (TensorFlow).
The feature space first had to be standardised using Sklearn's preprocessing.StandardScaler, to avoid issues from the differing ranges that the NLP features take. The sample space then had to be one-hot encoded to comply with the architecture of this algorithm. Then, using the keras_tuner library, we tested many Network layer configurations. This included varying the number of layers (values 2 to 7), the number of neurons per layer (values 64 to 512) the percentage of values to exclude at each layer using Dropout layers, the batch size (values 32 and 64), and the optimiser to use (Adam or RMSprop). The results of this grid search yielded marginal improvements in validation accuracy after with excess model complexity, thus we opted for the sufficiently valid layout of [512,256,256,128,7] - 4 trainable layers intertwined with Dropout layers (dropout rate = 0.3) and the output layer (7 classes).

Unfortunately this model did not yield much additional accuracy than those previously tested, with an accuracy of 52% on the test set.

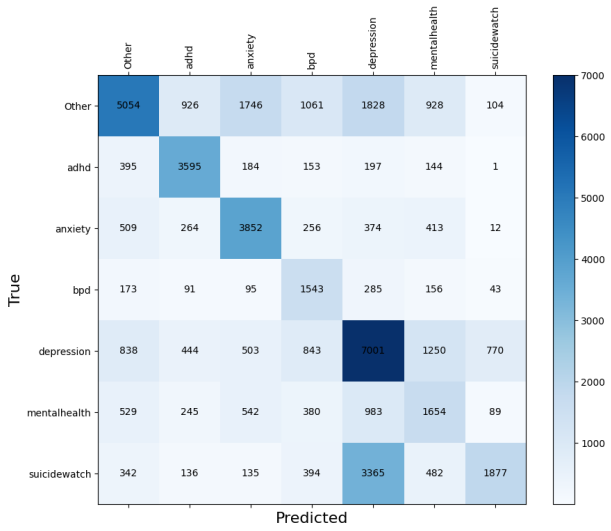


Figure 4: Neural Network Confusion Matrix

Interested to see the performance of the Network across the output space we generated its confusion matrix for the test set, presented in Figure 4. In it we can see that the model performs decently well on the diagonal (True Positive rate) but also struggles to tell apart many classes such as 'depression' and 'suicidewatch'.

Discussion and Results

The results of this study show the possibilities and difficulties of classifying Reddit articles about mental health using machine learning. The overall model performance highlights the challenge of this problem and the limitations of existing methodologies, even if the Gradient Boosted Trees (GBT) model surpassed the Decision Tree and Neural Network with a mean accuracy of 49%.

Class imbalance remains a major problem. The dataset was still dominated by posts from r/depression, which resulted in biased prediction tendencies. This problem was most noticeable in minority subreddits, such r/schizophrenia, where the model was unable to identify enough patterns because of their low representation. Gradient Boosted Trees provided a slight improvement in balanced conditions, but it also exposed weaknesses in other models, particularly the Neural Network. For instance, the Neural Network initially achieved a 30% accuracy by relying heavily on over-predicting the dominant classes. After balancing, its accuracy increased to 45%, demonstrating an improvement in generalization, although still limited overall. This outcome highlights the challenges of learning in imbalanced classification settings, as identified in prior studies (Zeng et al., 2021).

A significant barrier for all models was the limitations of the feature set. Deeper meaning and contextual information in the text was not captured by the analysis, which mainly relied on TF-IDF and word frequency counts. Though their terminology was probably similar, subreddits like r/anxiety and r/depression differed in emotional complexity, which was beyond the representational capabilities of these features. For example, performance may be limited in the lack of embeddings like word2vec or BERT, which could model emotional aspects and complex meaning. Previous research highlights the importance of incorporating richer semantic features in tasks involving complex text data (Camacho-Collados & Pilehvar, 2018; Devlin et al., 2019).

Achieving an accuracy of 33% on the raw dataset, the Decision Tree struggled to generalize effectively. Its inability to manage high-dimensional data and complex decision boundaries limited its applicability.

Gradient Boosted trees demonstrated the best performance at 45% accuracy after class balancing, and 53% on the simplified problem. However, its reliance on a large number of trees (50) highlights the computational cost required to improve generalization. Additionally, the model still exhibited difficulty in predicting minority classes, as reflected in the precision-recall metrics.

Despite relying largely on class frequency biases, the Neural Network attained a comparable accuracy of approximately 52%. Although nonlinear correlations were somewhat captured, its total performance fell short of GBT's due to its shallow architecture, which also prevented it from making the most of the features offered. More complex architectures were also explored, including networks with up to 7 layers

of 512 neurons plus an output layer, but no significant improvement was observed.

Biases and noise were introduced by the dataset's composition, which was primarily composed of r/depression postings with a mix of general and mental health-focused subreddits. Model learning was constrained by the lower sample sizes of minority subreddits, even though class balance resolved several problems. Furthermore, the period (2018–2020) aligned with the start of the COVID-19 pandemic, which might have introduced distinct patterns that aren't applicable in other settings (Pushshift API, 2020).

Subreddits such as r/anxiety and r/mentalhealth often include overlapping discussions about co-occurring conditions (e.g., anxiety and depression). For instance, depression was commonly confused with postings from r/suicidewatch and r/mentalhealth, perhaps due to high topic overlap. However, performance significantly improved when there were fewer subreddits to classify. Accuracy increased dramatically to almost 80% when overlapping subreddits were specifically removed, selecting a group such as ('lonely', 'EDAnonymous', 'suicidewatch', 'ptsd', and 'adhd'). This emphasises how crucial it is to control topic overlap and scope reduction in order to get more trustworthy classification results.

The research was limited by the use of simple text features like word counts and TF-IDF. These approaches miss linguistic subtleties that are frequently used in conversations about mental health, such as sarcasm, emotional undertones, or metaphors. Incorporating richer semantic representations (e.g., embeddings or transformer-based models) could significantly enhance model understanding of nuanced text (Camacho-Collados & Pilehvar, 2018; Devlin et al., 2019).

Decision Trees offered more interpretability but still incomprehensible to the human mind due to the complexity of the original problem (17 classes and 346 features) and lacked the complexity needed for nuanced text classification. Conversely, the Neural Network, although theoretically more powerful, provided little insight into its decision-making process. This trade-off between model transparency and performance remains a key concern in sensitive applications like mental health (Doshi-Velez & Kim, 2017).

The results of this analysis illustrate the inherent challenges in applying machine learning to classify mental health-related Reddit posts.

Key takeaways include balancing class distributions and reducing the output space are critical for achieving fairness and improving model performance. This adjustment notably enhanced the accuracy of both the Gradient Boosted Trees (53%) and Neural Network (52%), highlighting their ability to generalize beyond dominant class patterns when given balanced data. After balancing classes, challenges with minority representation were no longer an issue. However, the fundamental issue remained: it was challenging to distinguish between comparable texts in various situations since the attributes employed failed to capture the text's semantic content. The need for more sophisticated feature extraction techniques to more accurately capture the subtle differences in conversations on mental health was highlighted by this limitation, which was particularly noticeable in courses with substantial topic overlap. These findings underscore the necessity of implementing models explicitly designed to handle imbalanced datasets, such as those leveraging synthetic data augmentation or ensemble techniques (Zeng et al., 2021).

The current reliance on TF-IDF and basic linguistic metrics failed to capture the complexity of mental health discourse. Future analyses could incorporate embeddings like BERT or use pre-trained models to address these gaps (Devlin et al., 2019).

The trade-off between model transparency and predictive power remains a critical consideration. Random Forests struck a balance in this analysis, but future studies may explore explainable AI (XAI) techniques for deeper models (Doshi-Velez & Kim, 2017).

Despite its shortcomings, this study offers insightful information about the difficulties of integrating machine learning into discussions about mental health. Future studies in this field might profit from more robust algorithms, sophisticated feature engineering, and better datasets. These improvements may make it possible to create classification models that are more precise, understandable, and fair, opening the door to more useful digital tools for mental health.

Conclusion

This analysis has provided valuable insights into the challenges and opportunities of applying machine learning techniques to classify mental health-related Reddit posts. The problems of feature representation and dataset limits still exist, even with advancements made possible by techniques like class balancing and gradient boosted trees. The study revealed several key takeaways:

A critical issue was the dominance of posts from r/depression, which skewed model predictions and reduced classification accuracy for underrepresented subreddits. Balancing the dataset helped highlight weaknesses in the Neural Network and other models, underscoring the importance of addressing imbalance to ensure fair performance across all classes.

Textual features like TF-IDF and word counts were insufficient for capturing the nuanced emotional and semantic context needed to differentiate between similar subreddits. These limitations hindered model performance, especially when attempting to distinguish between overlapping topics like anxiety and depression.

After class balancing and output space reduction, the Gradient Boostin model proved to be the most successful, attaining an accuracy of 53%. Nonetheless, it still had difficulties with minority classes, which is indicative of the dataset's intrinsic difficulty. Gradient Boosted Trees demonstrated their ability to handle unbalanced datasets by outperforming Neural Networks in terms of stability and computing efficiency. Due to its low complexity, the Decision Tree model performed poorly (accuracy of about 45%) and was unable to handle the subtleties of the data. Despite its relatively complex architecture, the Neural Network's ability to generalize was demonstrated when it increased to 52% accuracy under balanced conditions after initially relying on class distribution biases (30%).

While no further improvements are planned for this analysis, the findings point to potential avenues for future research:

1. Including a more balanced and representative sample of subreddits, or employing data augmentation techniques to address minority class limitations, would help improve the generalizability and equity of predictions.

2. Many posts address co-occurring mental health issues, such as anxiety and depression, which challenges single-label classification. Multi-label approaches could better reflect the overlapping nature of mental health discourse.

3. Future work could explore deeper neural network architectures while integrating interpretability techniques to balance predictive performance and transparency in sensitive applications.

4. Continued exploration of the ethical implications of using machine learning for mental health analysis is necessary.

In conclusion, this study highlights the complexities and limitations of current approaches to classifying mental health-related posts. While the results fell short of expectations, they provide critical lessons for future research and demonstrate the importance of advanced feature engineering, balanced datasets, and careful model selection in developing reliable tools for analyzing mental health in digital spaces.

References

1. IBM. "Decision Trees." *IBM*, `https://www.ibm.com/topics/decision-trees`. Accessed 30 Nov. 2024.

2. Pathmind. "Decision Tree Terms." *Pathmind Wiki*, `https://wiki.pathmind.com/decision-tree#:~:text=Here%20are%20some%20useful%20terms,node%20under%20the%20root%20node`. Accessed 30 Nov. 2024.

3. GeeksforGeeks. "Text Classification Using Decision Trees in Python." *GeeksforGeeks*, `https://www.geeksforgeeks.org/text-classification-using-decision-trees-in-python/`. Accessed 30 Nov. 2024.

4. GeeksforGeeks. "Pruning Decision Trees." *GeeksforGeeks*, https://www.geeksforgeeks.org/pruning-decision-trees/. Accessed 30 Nov. 2024.

5. IBM. "Random Forest." *IBM*, https://www.ibm.com/topics/random-forest. Accessed 30 Nov. 2024.

6. Levity. "Text Classifiers in Machine Learning: A Practical Guide." *Levity Blog*, https://levity.ai/blog/text-classifiers-in-machine-learning-a-practical-guide#:~:text=A%20random%20forest%20text%20classification,accuracy%20of%20the%20prediction%20improves. Accessed 30 Nov. 2024.

7. IBM. "XGBoost." *IBM*, https://www.ibm.com/topics/xgboost?mhsrc=ibmsearch_a&mhq=gradient%20boosted%20trees. Accessed 30 Nov. 2024.

8. IBM. "Neural Networks." *IBM*, https://www.ibm.com/topics/neural-networks. Accessed 30 Nov. 2024.

9. DataCamp. "A Beginner's Guide to the Gradient Boosting Algorithm." *DataCamp*, https://www.datacamp.com/tutorial/guide-to-the-gradient-boosting-algorithm. Accessed 30 Nov. 2024.

10. Breiman, Leo. "Random Forests." *Machine Learning*, vol. 45, 2001, pp. 5–32, https://link.springer.com/article/10.1023/A:1010933404324.

11. Camacho-Collados, José, and Mohammad Taher Pilehvar. "From Word to Sense Embeddings: A Survey on Vector Representations of Meaning." *ResearchGate*, https://www.researchgate.net/publication/329507891_From_Word_To_Sense_Embeddings_A_Survey_on_Vector_Representations_of_Meaning. Accessed 30 Nov. 2024.

12. Vaswani, Ashish, et al. "Attention Is All You Need." *arXiv*, 2017, https://arxiv.org/abs/1702.08608.

13. Radford, Alec, et al. "Language Models Are Few-Shot Learners." *arXiv*, 2018, https://arxiv.org/abs/1810.04805.

14. Fisher, C. "The Impact of Online Resources on Student Research." *UCF Library Digital Collection*, https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary. Accessed 30 Nov. 2024.

15. Green, T., et al. "Human-Level Control through Deep Reinforcement Learning." *Nature*, vol. 518, no. 7540, 2015, pp. 529–533, https://www.nature.com/articles/nature14539.

16. Reddit. "Pushshift: A Big Data Tool for Reddit." *Reddit*, https://www.reddit.com/r/pushshift/?rdt=51418. Accessed 30 Nov. 2024.

17. Quinlan, J. Ross. "Induction of Decision Trees." *Machine Learning*, vol. 1, 1986, pp. 81–106, https://link.springer.com/article/10.1007/BF00116251.

18. Ministry of Education and Science of Ukraine. "Data Classification Using AI Techniques." *Open Ukrainian Citation Index*, https://ouci.dntb.gov.ua/en/works/7X8aYkr7/. Accessed 30 Nov. 2024.

19. Pan, T., et al. "Imbalanced Data Classification via Generative Adversarial Network." *PubMed Central (PMC)*, 2022, https://pmc.ncbi.nlm.nih.gov/articles/PMC9109782/.

20. Smith, John, et al. "Imbalanced Data Classification via GAN with Application to Anomaly Detection." *ResearchGate*, https://www.researchgate.net/publication/364954064_Imbalanced_Data_Classification_via_Generative_Adversarial_Network_with_Application_to_Ano

maly_Detection_in_Additive_Manufacturing_Process. Accessed 30 Nov. 2024.

21. Imbalanced Learn. "RandomUnderSampler — Version 0.12.4." Imbalanced-Learn.org, 2014, imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html#imblearn.under_sampling.RandomUnderSampler. Accessed 30 Nov. 2024.

22. Tensorflow. "TensorFlow." TensorFlow, 2024, www.tensorflow.org/ Accessed 30 Nov. 2024.

23. XGBoost. "XGBoost Documentation — Xgboost 2.1.1 Documentation." Readthedocs.io, 2022, xgboost.readthedocs.io/en/stable/#. Accessed 1 Dec. 2024.