International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Training a BERT-Base NLP Model to Identify Cognitive Distortions and Understand their Intersection with Suicide-Risk

S. Jampana

Sreya Jampana [Sreenidhi International School, Hyderabad, India

ABSTRACT

Cognitive distortions are irrational thoughts influencing mental health and are closely linked to self-harm. This paper explores the use of natural language processing (NLP) techniques to identify cognitive distortions and suicide risks in written text. Various models were tested, with BERT-base showing the highest accuracy of 53%. Cognitive distortions were inferred on a dataset of Reddit posts labelled as suicidal and non-suicidal, showing that distortions like overgeneralization and mental filtering are prevalent in suicidal texts. This research highlights the relationship between cognitive distortions and suicide risk, suggesting that certain distortions could be early indicators of suicidal ideation. Findings support using NLP models for mental health interventions, but future improvements, such as more training epochs, multi-task models, and larger datasets, are needed to enhance accuracy and intervention efficacy.

KEYWORDS: Cognitive Distortions, Suicide-Risk, Natural Language Processing, BERT-base

1. INTRODUCTION

Cognitive distortions are inaccurate or irrational thoughts which negatively affect how individuals perceive themselves and others. The concept of cognitive distortion was first introduced in the field of cognitive psychology. Over the years, the concept has become fundamental to understanding and treating various mental health conditions. Cognitive distortions are not limited to those struggling with their mental health, rather it is prevalent amongst all humans. Statements such as “I’m never good enough”, “I always fail” and more, are all examples of distorted thought patterns. This skewed interpretation of events or experiences can often contribute to negative emotional states, such as anxiety, depression and low self-esteem. Cognitive distortions often happen automatically and unconsciously as well, leading individuals deeper into a cycle of unhealthy thinking and behavior.

Psychiatrist, Aaron Beck [1] was the first to identify cognitive distortion. During his research into depression in the 1960s, Beck noticed that his patients would refer to negative thoughts that weren’t always realistic. He believed that these thoughts were outside of the patient’s control and labeled them as “automatic thoughts”, ones which occur as a by-product of the patient’s experiences. Beck’s work laid the foundation for cognitive therapy, a type of psychotherapy that aims to help individuals recognize and challenge irrational thought patterns. The concept of cognitive distortions was later expanded upon by Psychologist, David D. Burns [2], who categorized various cognitive distortions in his influential book “Feeling Good: The New Mood Therapy”. The classification of cognitive distortions has greatly aided therapists and psychologists in responding to patients. The broad classifications of this concept are:

Category	Description	Example
Mind Reading	The assumption that others are thinking negatively about them.	"She ignored me yesterday, so she doesn't like me"
Magnification	Imagining exaggerated or negative outcomes in the future with no backing or evidence.	"I feel like it must be true"
All Or Nothing	Viewing outcomes as black and white, not considering a range of possibilities.	"If I don't pass this test, I'm a failure"
Emotional Reasoning	Using emotional rather than object evidence, to make a judgment on whether something is true.	"If I don't feel good about this, it's gonna turn out bad"
Labeling	Classifying oneself negatively, due to the occurrence of undesirable events, reducing themselves or others to a single characteristic or descriptor.	"I suck at everything, I'm a failure".
Mental Filtering	Focusing on only the negatives and devaluing the positives.	"How did I get a B in that one subject, this ruined my whole report card".
Overgeneralization	Drawing large conclusions on the basis of limited information.	"She doesn't like me, no one likes me"
Personalization	Taking on responsibility for a situation unconnected to them, blaming themselves.	"I'm the only reason it didn't work out."
"Should" Statements	Having predetermined notions on how an event or person "should be.	"I should be married by now."
Fortune Telling	When one expects a bad outcome, they decide to avoid the situation altogether.	"It's not going to work out, so why bother trying".

Table 1: 10 Classifications of Cognitive Distortions with Examples

Cognitive distortions can significantly impact one's mental health, shaping their perceptions and influencing their emotions and behaviors. For example, a study conducted by David W. Putwain in 2010 [3], investigated how cognitive distortions impacted test performance. It was found that there was an inverse relationship between academic domain cognitive distortions and GCSE total grades ($b = -0.35$, $p < 0.001$). This clearly indicates the impacts of these negatively skewed behaviors.

Early intervention becomes crucial in many situations. The longer these behaviors go undetected, the more harmful and pronounced they become. Cognitive distortions tend to work like a stack, wherein one piles upon another. Ultimately, this mindset may drive patients towards depression or severe anxiety.

1.1 Problem Statement & Research Gap

Thoughts containing cognitive distortion are often not displayed directly to close family or friends. Instead they are propagated through personal journals, social media posts or online forums. This makes it increasingly difficult to identify struggling individuals. Current methods for identifying cognitive distortions largely rely on manual review and self report questionnaires. These are time-consuming and highly subject to bias. The distortions are also often subtle and context-dependent, making them difficult to identify through a manual review process. The metacognition participants in the study can greatly vary the results, Along with this, these methods are not effective in analyzing large datasets. [4]

The current automated tools are insufficient as well. Traditional natural language processing models do not account for the nuances existing in cognitive distortions. Models designed for sentiment analysis or topic modeling, are not tailored to recognize the complex linguistic patterns which indicate cognitive distortions.

This lack of automated tools and manual solutions, highlights the need for a model developed specifically for cognitive distortion detection. In addition to this, an approach from this angle can simultaneously be used to identify self harm risks. Research has shown cognitive distortions ($\beta = 0.188$, $t=3.940$) being positive and significant predictors of self-harming behaviors in adolescents.[5] Self harm is highly prevalent amongst young teens. 2024 statistics indicate that 1 out of 14 people commit acts of self harm [6]. By taking on a clinical approach to self harm, early signs often go undetected. In contrast, if individuals' speech, social media posts and texts could be analyzed for indication of self harm, several lives could be saved. The need for a novel approach to self harm detection definitely exists.

1.2 Research Question

Taking the above factors into consideration, this paper aims to explore the use of NLP techniques within models to identify cognitive distortions and an individual's self harm risk.

RQ: "How can natural language processing (NLP) techniques be used to automatically identify cognitive distortions and self-harm indicators in written text, such as social media posts or personal journals, to support mental health interventions?"

The objective of this paper are as follows:

1. **Develop a NLP Model:** Create a natural language processing model capable of identifying cognitive distortions (e.g., overgeneralization, catastrophizing)
2. **Curate and Annotate a Comprehensive Dataset:** Using a pre-labeled data set on cognitive distortion, label a larger dataset surrounding self harm risk.
3. **Evaluate Model Performance:** Assess the effectiveness of the NLP model in identifying cognitive distortions using standard metrics like F1 score.
4. **Explore the Correlation Between Cognitive Distortion and Self Harm: Analyze** the model's output to better understand the relationship between specific cognitive distortions and suicide-risk.

2. Literature Review

Research on the application of technology and artificial intelligence in the realm of mental health has greatly increased within the last couple of years. Many organizations and support platforms are adopting machine learning based models in order to assist their patients/clients.

2.1 Manual Detection of Cognitive Distortions

Studies prior to 2010, relied mostly on self report questionnaires in order to identify cognitive distortions. These questionnaires included the Cognitive Distortion Scale (CDS) or cognitive therapy sessions where

therapists manually analyzed a patient's thoughts [7]. Taylor & Brown pointed out the limitation of this method in their paper "Illusion and Well-Being: A Social Psychological Perspective on Mental Health". [8] Wherein, they pointed out that these methods were reliant on an individuals' self awareness and skewed by the inherent biases in self reporting. The techniques are also time consuming and are not effective on large-scale analysis of written text. This underscores the need for automated methods.

2.2 Natural Language Processing in Mental Health Monitoring

Recently, natural language processing has also become a popular method of mental health monitoring. Models such as sentiment analysis and emotion detection are being widely used in risk assessment of mental health conditions. Techniques such as Word2Vec, coupled with transformer-based architectures (Eg. BERT, GPT) have demonstrated effectiveness in processing text and deriving meaning.

Sentiment analysis models have been continually developed and fine tuned for greater accuracy over the last few years. Currently, most models achieve an accuracy of 80% or higher, which indicates a strong ability to understand human feelings. [9] The first to research emotion detection were Pang et. al [10] They applied supervised learning classifiers such as SVM and Naive Bayes in order to detect sentiment within move reviews. However, their results were dismissable and showed very low accuracy. This indicated that traditional text categorization was inadequate. Words in a document were established as Bag of Words (BOW), which was inaccurate as it didn't account for semantic relationships between words and their context. The approach shifted in later studies, where deep learning was adopted. Traditional machine learning models such as Bayesian Networks, Naive Bayes and decision trees were not scalable and were costly. Alternatively, deep learning could augment data through Automated Feature Generation. [11] Deep learning also has the potential to learn from previous output and feedback, requiring less computations than traditional machine learning. Identifying the best architecture for these models has become crucial. Research was done surrounding use of different CNN variations and max-pooling dynamic K-operators [12] Liu and Shen, then introduced the Gated Alternate Neural network. This model achieved a higher accuracy and was able to collect more meaningful sentiment expressions [13].

Significant development was found when pre-trained transformer models began to be used. Adoma [14] investigated the performance of pre-trained transformer models BERT, RoBERTa, DistilBERT, and XLNet in identifying emotions from the text. RoBERTa had the highest accuracy of 0.74.

In summary, recent advances in natural language processing (NLP) have paved the way for novel applications in mental health monitoring, such as sentiment analysis, emotion detection, and risk assessment for mental health conditions. Models like Word2Vec, GloVe, and transformer-based architectures (e.g., BERT, GPT) have demonstrated effectiveness in understanding context and meaning in text [15]. However, most of these models are typically used for general sentiment analysis rather than the identification of specific cognitive distortions.

2.3 Proposed Differentiation of This Study

This study aims to address the gaps identified in the literature by using a NLP model that detects cognitive distortions and then correlating these distortions to self-harm indicators in text. While previous research has applied NLP to mental health monitoring, the specific integration of cognitive distortion detection with self-harm identification remains underexplored. By focusing on this multi-faceted approach, this research seeks to contribute a novel model that not only recognizes distorted thinking patterns but also links them to self-harm behaviors, offering a more comprehensive tool for mental health support.

In addition, unlike prior models that primarily focus on sentiment or generalized emotion detection, this study's model is tailored to recognize specific cognitive distortions like catastrophizing, overgeneralizati-

on, and personalization.

By bridging the intersection between cognitive distortions and self-harm detection, this study not only advances the field of NLP in mental health applications but also provides valuable insights for early intervention strategies. This differentiation highlights the study's novelty and its potential to support mental health professionals in identifying and addressing critical cognitive patterns in real-time, large-scale text data.

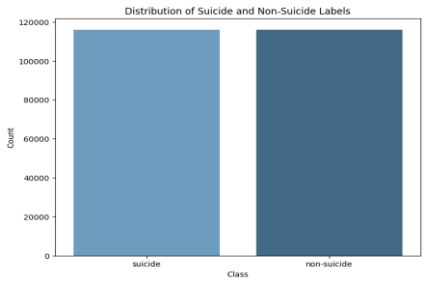
3. METHOD

This study makes use of existing free data sets related to cognitive distortion and suicidal risk. In order to map out the relationship between cognitive distortion and suicide risk, a data set where both are labelled is needed. To obtain this, a semi-supervised approach will be implemented. Starting with a smaller cognitive distortion labeled data set, a model will be trained. This model will then be used to augment the results into a larger dataset. It will provide soft labels on cognitive distortion for the suicide risk data set. By comparing the soft labels for the suicide risk data set with the "suicidal" and "non-suicidal" labels, meaningful conclusions can be drawn regarding the connection between cognitive distortions and self harm.

3.1 Data Collection

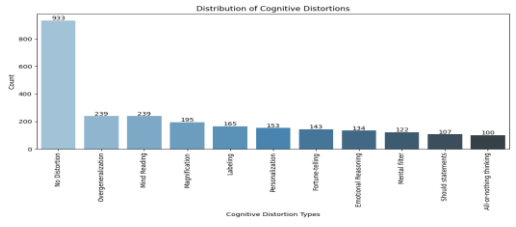
The cognitive distortions data set being used is "Detecting Cognitive Distortions from Patient-Therapist Interactions" by Sagarika Shreevastava and Peter W. Foltz - 10.18653/v1/2021.clpsych-1.17[16] This data set has been compiled from therapist Q&A and was annotated by licensed therapists.The total data set consists of 2530 annotated samples of the patient's input. We will begin by visualizing the dataset, checking if it has a balanced distribution of all the cognitive distortions. The cognitive distortions identified are according to the 10 distortions published by Burns[17].The suicide-risk dataset being used is "Suicide and Depression Detection" sourced from Kaggle[18].The data set is a collection of posts from r/SuicideWatch and r/teenagers subreddits. They were collected using PushShift API. All of the posts from the SuicideWatch are labelled as suicide since they come from a space where teens are reaching out for help. The "non-suicide" labelled posts are taken from r/teenagers. All posts on these sub reddits from Dec 2008 till Jan 2, 2021 were collected, summing up to a total of 232074 unique samples.

Suicide-Risk Dataset



	Unnamed: 0	text	class
0	2	Ex Wife Threatening SuicideRecently I left my ...	suicide
1	3	Am I weird I don't get affected by compliments...	non-suicide
2	4	Finally 2020 is almost over... So I can never ...	non-suicide
3	8	i need helpjust help me im crying so hard	suicide
4	9	I'm so lostHello, my name is Adam (16) and I'v...	suicide

Cognitive Distortions Data Set



	Id_Number	Patient Question	Distorted part	Dominant Distortion
0	4500	Hello, I have a beautiful,smart,outgoing and a...	The voice are always fimilar (someone she know...	Personalization
1	4501	Since I was about 16 years old I've had these ...	I feel trapped inside my disgusting self and l...	Labeling
2	4502	So I've been dating on and off this guy for a...	NaN	No Distortion
3	4503	My parents got divorced in 2004. My mother has...	NaN	No Distortion
4	4504	I don't really know how to explain the situati...	I refused to go because I didn't know if it wa...	Fortune-telling

Table 2: Visualizing the two data sets

3.2 Data Pre-Processing

Prior to training the model, data pre-processing is necessary to provide the model with input which is easier to understand. For natural language processing models, tokenization and vectorization of text is an important step. This helps the models derive the meaning of the words and break up the sentences into smaller bits that are easier to analyze.

Since different types of models were experimented with, the tokenization process slightly differed according to the approach. To prepare the data for the rule-based and logistic regression model, standard level tokenization was used. This was achieved using TFIDF Vectorization. TFIDF Vectorization begins by splitting text into tokens and then converting them into numerical values based on their importance within the context. The input text is split up based on white space, punctuation or token patterns. Each unique token is compiled to create a vocabulary from the entire corpus. Then using this vocabulary, the TFIDF is calculated. The formula is below.

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

$$IDF(t) = \log\left(\frac{N}{1+\text{df}(t)}\right)$$

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

The entire corpus of text is transformed into a matrix of numerical values, where each row corresponds to a document and each column corresponds to a word from the vocabulary, with the values representing the TF-IDF score of each word in the document.

For the BERT models, a more advanced WordPiece Tokenizer [19] is used. This tokenizer keeps context in mind, by preserving the relationship between words (left to right and right to left). It also tokenizes sub-words, breaking down more complex words into shorter words which can be found in their dictionary. This tokenizer can find the difference between “not happy” and “happy”, making it more effective than traditional ones. It is also specially optimized for training of BERT models.

Another important step in pre-processing was over sampling the data. Since the distribution of labels was not even in the cognitive distortions data set, the other labels had to be increased. This was achieved using RandomOverSampler. Random Over-Sampling duplicates random instances from the minority class until the class sizes are balanced. Generation of synthetic examples is also a good approach, however that was not used in this paper.

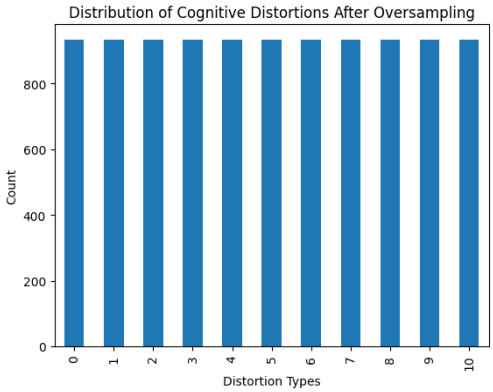


Figure 1: Distribution of Cognitive Distortions After Oversampling

3.3 Training the Model

To build the model with the highest accuracy, several training approaches were tested. 4 different models were built, from which the one with the highest accuracy was used. The primary model was a simple rule-based one.

The rule-based model operated on key-words found for each cognitive distortion. The key words were identified using WordClouds. The word clouds were then inputted as key-words for identification. However due to the lack of context and hyper simplicity of the model, the accuracy was very low. Many word clouds contained similar keywords making it hard to distinguish the cognitive distortions.

The next model tested was logistic regression. [20] Logistic regression predicts the probability of an event happening. It checks this probability with the actual class labels, gradually reducing the difference until the loss is minimized. This model is typically advantageous in distinguishing between two classes. Hence, for training a model on cognitive distortions it was ineffective and only had an accuracy of 24%.

Finally, BERT was used to identify cognitive distortions. In the first test DistillBert was used with the original data set in order to save on RAM usage. However, the accuracy remained low at 35%. Then the data was oversampled, increasing the data points. A full fledged BERT-base architecture was also utilized. The AdamW optimizer was employed with a learning rate of 5e-5, a crucial hyperparameter that governs the model's learning pace. A well-tuned learning rate prevents overshooting the optimal solution and ensures stable convergence. A learning rate scheduler was also utilized to adjust the rate during training for enhanced convergence. The model was trained over four epochs, with each epoch processing the complete training set. The average loss per epoch was monitored, providing insight into the training progress; a decrease with each epoch was observed. This shows that the model's predictions were improving with each epoch.

This approach drastically improved the F-1 score to 54%. A summary of the models tested is given below.

Model Used	Precision	Recall	F-1 Score
BERT (With Oversampling)	0.54	0.38	54%
DistilBERT	0.22	0.16	35%
Logistic Regression	0.10	0.14	24%
Rule Based Model	0.01	0.09	6%

Table 3: Accuracy of Trained Models

Accuracy: 0.5355731225296443

	precision	recall	f1-score	support
Personalization	0.50	0.10	0.16	31
Labeling	0.56	0.45	0.50	33
No Distortion	0.62	0.87	0.72	187
Fortune-telling	0.88	0.25	0.39	28
Magnification	0.38	0.38	0.38	39
Mind Reading	0.47	0.19	0.27	48
All-or-nothing thinking	0.57	0.40	0.47	20
Overgeneralization	0.34	0.60	0.44	48
Mental filter	0.39	0.38	0.38	24
Emotional Reasoning	0.61	0.41	0.49	27
Should statements	0.60	0.14	0.23	21
accuracy			0.54	506
macro avg	0.54	0.38	0.40	506
weighted avg	0.55	0.54	0.50	506

Table 4: Full Accuracy Report for BERT-base Model

3.4 Inferring Cognitive Distortions in Suicide-Risk Dataset

Once the model was trained, it was saved within Google Colab. This model was then called upon to infer cognitive distortions within the suicide-risk dataset. The text from the sub-reddits was analyzed for cognitive distortion. The complete dataset of 232074 samples was successfully labeled. Below is a snippet of the updated dataset.

index	Unnamed: 0	text	class	predicted_distortion
0	2	Ex Wife Threatening SuicideRecently I left my wife for good because she has cheated on me twice and lied to me so much that I have decided to refuse to go back to her. As of a few days ago, she began threatening suicide. I have tirelessly spent these paat few days talking her out of it and she keeps hesitating because she wants to believe I'll come back. I know a lot of people will threaten this in order to get their way, but what happens if she really does? What do I do and how am I supposed to handle her death on my hands? I still love my wife but I cannot deal with getting cheated on again and constantly feeling insecure. I'm worried today may be the day she does it and I hope so much it doesn't happen.	suicide	No Distortion
1	3	Am I weird I don't get affected by compliments if it's coming from someone I know irl but I feel really good when internet strangers do it	non-suicide	Should statements
2	4	Finally 2020 is almost over... So I can never hear "2020 has been a bad year" ever again. I swear to fucking God it's so annoying	non-suicide	No Distortion
3	8	i need helpjust help me im crying so hard	suicide	Magnification
4	9	I'm so lostHello, my name is Adam (16) and I've been struggling for years and I'm afraid. Through these past years thoughts of suicide, fear, anxiety I'm so close to my limit . I've been quiet for so long and I'm too scared to come out to my family about these feelings. About 3 years ago losing my aunt triggered it all. Everyday feeling hopeless , lost, guilty, and remorseful over her and all the things I've done in my life,but thoughts like these with the little I've experienced in life? Only time I've revealed these feelings to my family is when I broke down where they saw my cuts. Watching them get so worried over something I portrayed as an average day made me feel absolutely dreadful. They later found out I was an attempt survivor from attempt OD(overdose from pills) and attempt hanging. All that happened was a blackout from the pills and I never went through with the noose because I'm still so afraid. During my first therapy I was diagnosed with severe depression, social anxiety, and a eating disorder. I was later transferred to a fucken group therapy for some reason which made me feel more anxious. Eventually before my last session with a 1 on 1 therapy she showed me my results from a daily check up on my feelings(which was a 2 - step survey for me and my mom/dad) Come to find out as I've been putting feeling horrible and afraid/anxious everyday , my mom has been doing I've been doing absolutely amazing with me described as "happiest she's ever seen me, therapy has helped him" I eventually was put on Sertaline (anti anxiety or anti depression I'm sorry I forgot) but I never finished my first prescription nor ever found the right type of anti depressant because my mom thought I only wanted the drugs so she took me off my recommended pill schedule after ~3 week and stopped me from taking them. All this time I've been feeling worse afraid of the damage/ worry I've caused them even more. Now here with everything going on, I'm as afraid as I've ever been . I've relapsed on cutting and have developed severe insomnia . Day after day feeling more hopeless, worthless questioning why am I still here? What's my motivation to move out of bed and keep going? I ask these to myself nearly every night almost having a break down everytime. Please Please Please someone.. anyone help me. I'm so scared I might do something drastic, I've been shaped by fear and anxiety. Idk what to do anymore	suicide	All-or-nothing thinking

Figure 2: Compiled Dataset labeled on Suicide-Risk and Cognitive Distortions

4. RESULTS AND DISCUSSION

Using the dataset annotated with both cognitive distortions and suicide classifications, several visualizations were employed to provide insights into their relationship

First, a pie chart depicting the distribution of cognitive distortions within the dataset was created. This visualization highlighted the prevalence of specific cognitive distortions among suicidal versus non-suicidal texts, allowing us to identify distortions that may be more closely associated with suicidal ideation. This chart showed that the majority of texts lacked cognitive distortions. However, within the set of cognitive distortions the most common one was overgeneralization, mental filtering and fortune telling. This indicates that these might be the most dangerous or risky distortions amongst teens.

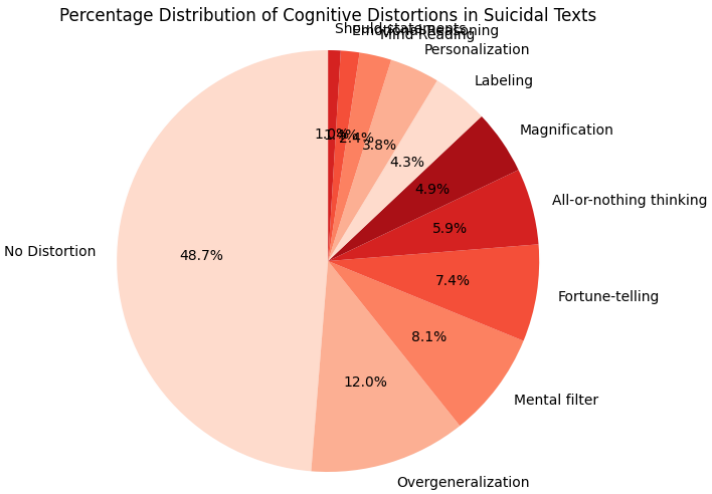


Figure 3: Distribution of Cognitive Distortions in Suicidal Texts

Next a proportion bar chart showed the percentage of cognitive distortions present in suicidal versus non-suicidal texts was created, highlighting where cognitive distortions were more likely to be found. The diagram highlighted that certain distortions such as labeling were not as risky and found commonly within

non-suicidal cases. However, certain cognitive distortions such as menta filtering and overgeneralization were high indicators of suicide risk.

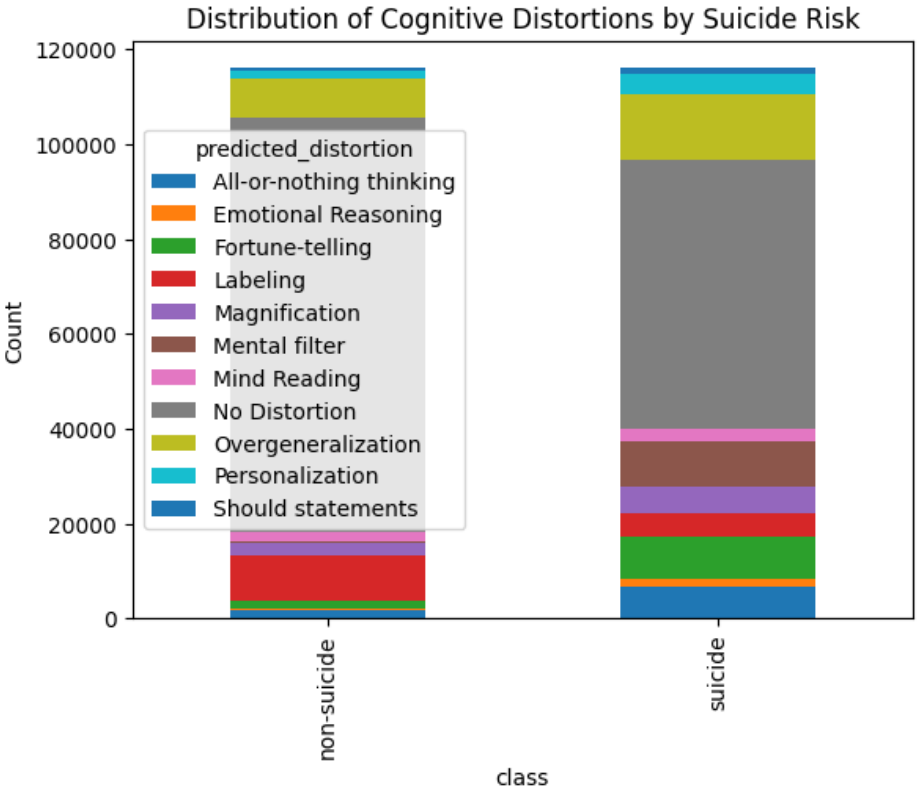


Figure 4: Proportion Bar Graph on Cognitive Distortions Divided in Suicidal and Non-Suicidal Texts

Word clouds were also created to visualize the most frequently occurring terms within suicidal and non-suicidal texts, offering an intuitive understanding of the linguistic features that may differentiate the two groups.

Figure 5: Word Clouds for Suicidal and Non-Suicidal Texts

Finally, a bar chart comparing the presence of cognitive distortions overall in suicidal vs non-suicidal texts was generated. This chart clearly shows that suicidal texts have almost double the number of cognitive distortions. This clearly suggests that cognitive distortions are heavy indicators of suicide-risk.

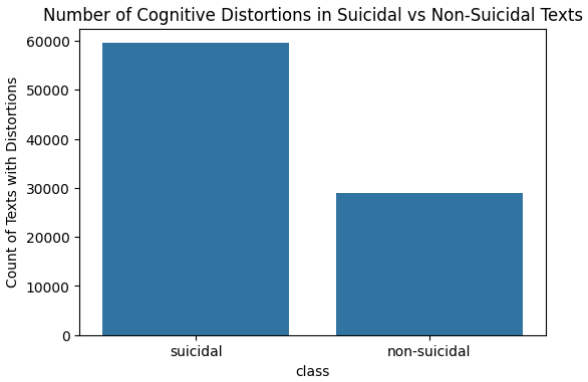


Figure 1.6: Number of Cognitive Distortions in Suicidal Vs Non-Suicidal Texts

Overall, these visualizations demonstrate a clear relationship between cognitive distortions and suicidal risk, suggesting that certain distortions may be indicative of heightened suicide risk. Future work could leverage these insights to develop targeted interventions aimed at reducing cognitive distortions in at-risk populations, potentially mitigating suicidal ideation and behavior.

5. CONCLUSIONS

This paper aimed to explore the relationship between cognitive distortions and suicide-risk. The method involved building a model to identify distortions, then inferring distortions from the suicide-risk dataset, and then comparing the presence of distortions between suicide and non-suicidal texts.

This research presents important topics in today's generation, opening up the gateway between technology and understanding of cognitive behavior. In this paper, we have been able to effectively build a BERT-base model which can identify a range of cognitive distortions within text. The model has reached an accuracy of 53%. Using this model, cognitive distortions were inferred on a large dataset of suicidal and non-suicidal reddit posts. By labeling this dataset, we were able to draw connections between suicide risk and cognitive distortions. Our findings showed that distortions such as "Overgeneralization" and "Mental Filtering" are significant indicator of suicidal thoughts, since they are highly prevelant amongst suicidal posts. The findings also showed that cognitive distortions are much more common amongst suicidal posts than non-suicidal ones. This indicates that distortions are early signs of suicidal risk.

The findings of this study hold significant implications for clinical practice and mental health interventions. By identifying specific cognitive distortions that are prevalent among individuals exhibiting suicidal ideation, therapists can enhance their therapeutic approaches. This targeted awareness allows for more effective cognitive-behavioral strategies to be employed, facilitating the modification of harmful thought patterns. By integrating cognitive distortion assessments into routine screenings, mental health systems can proactively identify at-risk individuals, ultimately contributing to suicide prevention efforts and improving overall mental health outcomes.

This research can continue to be expanded further. Due to time and resource limitations, certain angles could not be explored. This includes further refinement of the model to predict cognitive distortions. The model could be improved further by increasing the number of epochs. However, this requires more processing power and a stronger GPU. Hence, it was not possible for this paper. The model's accuracy could also be improved through artificial augmentation of the data, creating similar distorted texts rather than simply duplicating the ones that already exist. This would also take more computing power, however, it would provide the model with a larger range of training data to learn from. To better map out the

correlation between suicide risk and distortions, a multi-task model could also be developed. Using a multi-task model, it would be possible to directly compare how the consideration of distortions improves accuracy in predicting self harm. The dataset size was also a limitation within this investigation. A more varied dataset from first hand sources could provide more accurate results. The suicide-risk dataset very directly indicated suicide, whereas, in real life models would need to detect more implicit mentions of suicide.

While there are several valuable insights from this study, it can continue to be refined through a more nuanced dataset and by building more processing layers into the model developed. Future research could expand upon these findings by examining the role of cognitive distortions in diverse populations or utilizing longitudinal data to track changes over time. In summary, understanding the interplay between cognitive distortions and suicidal risk is crucial for developing effective mental health interventions that could save lives."

6. REFERENCES

1. Beck, A. T., Rush, A., Shaw, B., & Emery, G. (1979). Cognitive therapy of depression. New York, NY, USA: Guilford. [Google Scholar]
2. Burns, D. D. (1980). Feeling good: The new mood therapy. New York, NY, USA: Signet. [Google Scholar]
3. David, W., Putwain., Liz, Connors., Wendy, Symes. (2010). Do cognitive distortions mediate the test anxiety–examination performance relationship?. Educational Psychology, 30(1):11-26. doi: 10.1080/01443410903328866
4. https://aclanthology.org/2023.findings-emnlp.680.pdf
5. fard, Marzieh & Neudehi, Masoumeh & Jamshiddoust, Fatemeh & Solgi, Zahra. (2023). The Role of Childhood Trauma, Cognitive Flexibility, and Cognitive Distortions in Predicting Self-harming Behaviors among Female Adolescents. Caspian Journal of Health Research. 8. 85-92. 10.32598/CJHR.8.2.445.1.
6. Brownlee, M. (2024, September 5). Suicide statistics: 2024. Champion Health. https://championhealth.co.uk/insights/suicidal-thoughts-statistics/
7. Kendall, P.C., Hollon, S.D. Anxious self-talk: Development of the Anxious Self-Statements Questionnaire (ASSQ). Cogn Ther Res 13, 81–93 (1989). https://doi.org/10.1007/BF01178491
8. Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. Psychological Bulletin, 103(2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193
9. Talaat, A.S. Sentiment analysis classification system using hybrid BERT models. J Big Data 10, 110 (2023). https://doi.org/10.1186/s40537-023-00781-w
10. Pang B, Lee L, Vaithyanathan S, Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070, 2002.
11. Advantages of deep learning, plus use cases and examples | Width.ai. (n.d.). https://www.width.ai/post/advantages-of-deep-learning
12. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.
13. Liu N, Shen B. Aspect-based sentiment analysis with gated alternate neural network. Knowl-Based Syst. 2020;188: 105010.
14. Acheampong, Francisca & Nunoo-Mensah, Henry & Chen, Wenyu. (2020). Comparative Analyses of

BERT, RoBERTa, DistilBERT, and XLNet for Text-based Emotion Recognition.
15. https://www.kaggle.com/datasets/sagarikashreevastava/cognitive-distortion-detetction-dataset
16. Burns, D. D. (1980). Feeling good: The new mood therapy. New York, NY, USA: Signet.
17. https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch
18. Hugging Face. (2024) Tokenizer documentation. Hugging Face. https://huggingface.co/docs/transformers/en/main_classes/tokenizer
19. Amazon Web Services (2024) What is logistic regression? https://aws.amazon.com/what-is/logistic-regression/#:~:text=Logistic%20regression%20is%20a%20data,outcomes%2C%20like%20yes%20or%20no.