

Detecting Cognitive Distortions from Patient-Therapist Interactions

Sagarika Shreevastava

Department of Computer Science, and
Department of Linguistics,
University of Colorado, Boulder
sagarika.shreevastava@colorado.edu

Peter W. Foltz

Institute of Cognitive Science
University of Colorado, Boulder
peter.foltz@colorado.edu

Abstract

An important part of Cognitive Behavioral Therapy (CBT) is to recognize and restructure certain negative thinking patterns that are also known as cognitive distortions. This project aims to detect these distortions using natural language processing. We compare and contrast different types of linguistic features as well as different classification algorithms and explore the limitations of applying these techniques on a small dataset. We find that pre-trained Sentence-BERT embeddings to train an SVM classifier yields the best results with an F1-score of 0.79. Lastly, we discuss how this work provides insights into the types of linguistic features that are inherent in cognitive distortions.

1 Introduction

Cognitive Behavioral Therapy (CBT) is one of the most common methods of psycho-therapeutic intervention to treat depression or anxiety. Due to the COVID-19 pandemic, mental health issues are on the rise. At the same time, more and more interactions are now held virtually. Furthermore, mental health issues are not limited to the one-hour-per-week window that patients usually get with their therapists. This has led to a growth in the demand for digitally accessible therapy sessions. As mental health care is often inaccessible to people, there is a need for innovative ways to make it more widely available and affordable (Holmlund et al., 2019).

One possible solution is to develop an automated system that could serve by performing some ancillary tasks more efficiently. Towards that, Natural Language Processing (NLP) and Machine learning (ML) algorithms are now gaining widespread popularity and are being implemented in many fields where language is used. While we are far from a chatbot replacing a therapist’s nuanced skillset, having easy access to an intelligent support system can help fill in these gaps.

One of the major aspects of CBT is to recognize and restructure certain types of negative thinking patterns. Some established negative thinking patterns are commonly observed in patients dealing with anxiety or depression. These cognitive distortions arise due to errors in reasoning (Beck, 1963). The aim of educating the patient about these distortions during CBT is to equip the patient with the right tools to detect errors in their own thought processes. Once the patient is aware of the error in their reasoning, they can start to work on restructuring how to perceive the same situations in a healthier way.

1.1 Cognitive Distortions

The concept of cognitive distortions was first introduced by Beck (1963). There is no definitive number of types of distortions, and the number varies widely in existing literature depending on the level of detail in reasoning considered by the author. For example, the Cognitive Distortion Scale developed by Briere (2000) consists of only five types. In this work, we consider a total of ten types of cognitive distortions that are described below:

1. **Emotional Reasoning:** Believing “I feel that way, so it must be true”
2. **Overgeneralization:** Drawing conclusions with limited and often un negative experience.
3. **Mental Filter:** Focusing only on limited negative aspects and not the excessive positive ones.
4. **Should Statements:** Expecting things or personal behavior should be a certain way.
5. **All or Nothing:** Binary thought pattern. Considering anything short of perfection as a failure.
6. **Mind Reading:** Concluding that others are reacting negatively to you, without any basis in fact.
7. **Fortune Telling:** Predicting that an event will always result in the worst possible outcome.

8. **Magnification:** Exaggerating or Catastrophizing the outcome of certain events or behavior.
9. **Personalization:** Holding oneself personally responsible for events beyond one’s control.
10. **Labeling:** Attaching labels to oneself or others (ex: “loser”, “perfect”).

These distortions are based on the 10 types of cognitive distortion defined by [Burns and Beck \(1999\)](#). Some of these distortions are either combined into a super-category, or further divided into sub-categories, and hence the varying number of types of distortions. For example, mind reading and fortune telling are sometimes grouped and considered as a single distortion called Jumping to conclusions.

1.2 Problem statement

The first goal of this research project is to detect cognitive distortions from natural language text. This can be done by implementing and comparing different methodologies for binary classification of annotated data, obtained from mental health patients, into Distorted and Non-Distorted thinking. The second goal is to analyze the linguistic implications of classification tasks of different types of distortions.

In particular, this research aims to answer the following questions:

1. Which type of NLP features is more suitable for cognitive distortion detection: semantic or syntactic? Simply put, to compare *what* is said and *how* is it said in the context of this task. And, how important is word order in this context?
2. How well do these NLP features and ML classification algorithms perform this task with a limited-sized dataset?

1.3 Related work

Previous work done in this field includes the Stanford Woebot, which is a therapy chatbot ([Fitzpatrick et al., 2017](#)). The dialogue decision in Woebot is primarily implemented using decision trees. It functions on concepts based on CBT including the concept of cognitive distortions. However, it only outlines several types of distortions for the user and leaves the user to identify which one applies to their case.

Another study established a mental health ontology based on the principles of CBT using a gated-

CNN mechanism ([Rojas-Barahona et al., 2018](#)). The model associated certain thinking errors (cognitive distortions) with specific emotions and situations. Their study uses a dataset consisting of about 500k posts taken from a platform that is used for peer-to-peer therapy. The distribution of types of distortion is very similar to our results. These tasks come with annotator agreement issues - their inter-annotator agreement rate was 61%. One possible reason for the low agreement rate given by the authors is the presence of multiple distortions in a single data point.

As there is a lack of publicly available structured data that was curated specifically for the detection of cognitive distortions, datasets from other domains, such as social media data or personal blogs are used instead. One such study was conducted on Tumblr data collected by using selected keywords ([Simms et al., 2017](#)). By using the LIWC features (Section 3.3) to train a Decision Tree model to detect the presence of cognitive distortions, they were able to lower the false positive rate to 24% and the false-negative rate to 30.4%.

A similar study was conducted by [Shickel et al. \(2020\)](#) on a crowdsourced dataset and some mental health therapy logs. Their approach was to divide the task into two sub-tasks - first to detect if an entry has a distortion (F1-score of 0.88) and second to classify the type of distortion (F1-score of 0.68). For this study, 15 different classes are considered for the types of distortion. For both of the tasks - logistic regression outperformed more complex deep learning algorithms such as Bi-LSTMs or GRUs. On applying this model to smaller counseling datasets, however, the F1-score dropped down to 0.45.

2 Methods and Dataset

One of the most common roadblocks in using Artificial Intelligence for Clinical Psychology is the lack of available data. Most of the datasets that have patients interacting with licensed professionals are confidential and therefore not publicly available.

Here, we use a dataset, named Therapist Q&A, obtained from the crowd-sourced data science repository, Kaggle¹. The dataset follows a Question and Answer format and the identity of each patient is anonymized, to maintain their privacy.

Each patient entry usually consists of a brief description of their circumstance, symptoms, and

¹<https://www.kaggle.com/arnmaud/therapist-qa>

their thoughts. Each of these concerns is then answered by a licensed therapist addressing their issues followed by a suggestion. Since the patient entry is not just a vague request and it provides some insight into the situation as well as their reaction to it, it can be used to detect if they were engaging in any negative thinking patterns.

2.1 Annotation of dataset

For the annotation task, we have just focused on the patient’s input. One of the key factors in detecting cognitive distortions is context. While the data does give some insight into the situation a patient is in, it should be noted that the description itself is given by the patient themselves. As a result, their version of the situation itself may be distorted.

In this task, we focus on detecting cues in language that would indicate any type of distortion and there was no way to verify the veracity of their statements. Thus each entry is perceived as a viable candidate for cognitive distortion and given one out of 11 labels (‘No distortion’ and 10 different types of distortions as listed in section 1.1). It is noted that an entry can have multiple types of distortions. However for this project, the annotators were asked to determine a dominant distortion for each of the entries, and an optional secondary distortion if it is too hard to determine a dominant distortion. The decision between dominant or secondary distortion was made based on the severity of each distortion. Since the project aims to detect the presence of these distortions, the severity of distortions was not marked by any quantitative value. They were also asked to flag the sentences that led them to conclude that the reasoning was distorted.

The annotators coded 3000 samples out of which, 39.2% were marked as not distorted, while the remaining were identified to have some type of distortion. The highly subjective nature of this task makes it very hard to achieve a high agreement rate between the annotators. On comparing the dominant distortion of about 730 data points encoded by two annotators, the Inter-Annotator Agreement (IAA) for specific *type* of distortion was 33.7%. Considering the secondary distortion labels as well and computing a more relaxed agreement rate bumped the agreement to ~ 40%. On the other hand, the agreement rate increased to 61% when we focus on distorted versus non-distorted thinking only. The IAA metric used here is the Joint Probability of Agreement. These disagreements

were resolved by enabling the annotators to discuss their reasoning and come to a consensus. The types of distortion were found to be evenly distributed across the 10 classes of distortions mentioned earlier (figure 1). The annotated dataset will be made available to the public to encourage similar work in this domain.

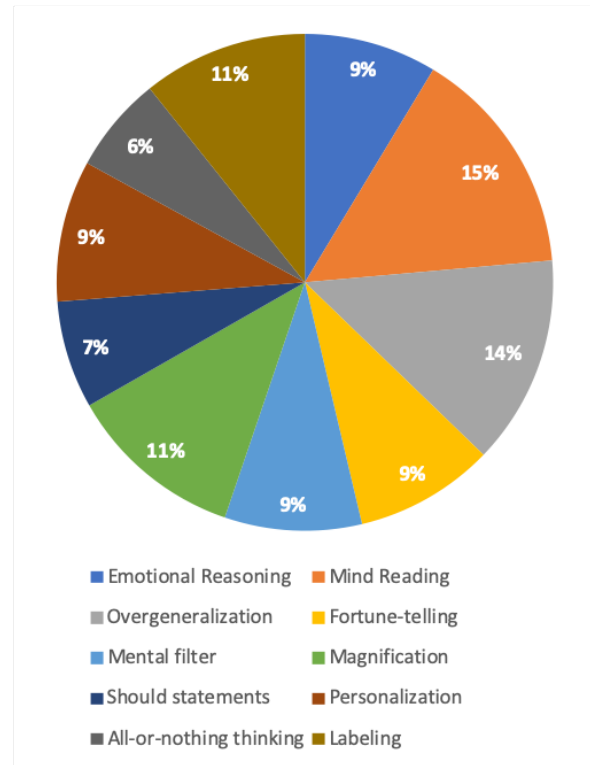


Figure 1: Distribution of the types of Cognitive Distortions in the Kaggle dataset

2.2 Experiments

Due to the limited size of the annotated dataset, several machine learning algorithms such as complex deep learning methods were eliminated from the experiments. Finally, the four types of features (Table 1) were tested using the following classification algorithms:

1. Logistic regression
2. Support vector machines
3. Decision trees
4. K- Nearest Neighbors (k = 15)
5. Multi-Layer Perceptron (with a single hidden layer having 100 units)

All of these classification algorithms were implemented with the default hyper-parameter settings using the python package commonly used for ML

algorithms, scikit-learn ².

3 Feature Selection

To address the different aspects of language, feature selection was divided into two categories - Semantic and Syntactic features. Two different training approaches were implemented for each of these categories. A brief description of each training method is given below.

	Bag-of-words approach	Sequential approach
Semantic	SIF	S-BERT
Syntactic	LIWC	POS

Table 1: Types of linguistic features. Note that LIWC features are not limited to the Syntactic category.

3.1 Smooth Inverse Frequency (SIF)

There are multiple ways of encoding Sentence embeddings where the word order does not matter. One of the most common methods is simply using the mean value of all the word embeddings.

Another common approach is to treat these sentences as documents and use TF-IDF (Term Frequency - Inverse Document Frequency) vectors. However, the issue with treating sentences as documents is that sentences usually do not have multiple words repeated.

To address this, smooth inverse frequency (SIF) can be used instead. The SIF method for sentence embeddings improves the performance for textual similarity tasks, beating sequential deep learning models such as RNNs or LSTM (Arora et al., 2016).

Here, the sentence embeddings are generated using the SIF method on pre-trained GloVe embeddings (Pennington et al., 2014) for each word in the sentence.

3.2 Sentence-BERT (Bidirectional Encoder Representations from Transformers)

For the sequential semantic representation of these entries, a pre-trained sentence-BERT model was used (Reimers and Gurevych, 2019). To ensure that in this vector space, semantically similar sentences are closer, the authors have used Triplet Objective Function as the loss function. This triplet objective function minimizes the distance between the

²<https://scikit-learn.org>

anchor sentence and a positive sample while maximizing the distance between the anchor sentence and a negative sample.

3.3 Linguistic Inquiry and Word Count (LIWC) Features

The linguistic inquiry and word count (LIWC) is a tool used to analyze textual data (Pennebaker et al., 2001). The LIWC program generates about 80 features based on the words used in the text. While we categorize the LIWC features as syntactic in table 1, these features reflect the percentage of words in different categories. A lot of these features are syntactic, such as the count of pronouns, proper nouns, etc. Other categories are psychological, linguistic, cognitive, or other (Tausczik and Pennebaker, 2010).

LIWC features are widely used for conducting linguistic analysis in almost any domain. Specific to mental illness, these features were used to detect the linguistic indicators of Schizophrenia (Zomick et al., 2019), Depression (Jones et al., 2020) and even Cognitive Distortions (Simms et al., 2017).

3.4 Parts of Speech (POS) tag embeddings

The main motivation behind using Parts of speech tags was to eliminate any specific Noun or Verb from heavily dominating the classification process. Two entries having the same context can have different distortions. Using POS tags as features have proved to be useful for similar applications, such as detecting depression from text (Morales and Levitan, 2016).

Syntactic features generally do not consider word order as an important aspect. To maintain the impact of word order each word is replaced with its Part-Of-Speech (POS) tag ³ using the pre-trained Spacy language model ⁴.

These POS tags are then converted to embeddings by similarly training them as word embeddings using Skip-gram word2vec model (Mikolov et al., 2013). This is done to encode POS tag-order in the embeddings. Once each tag has an embedding, these vectors are padded with zeros for normalization.

³<https://universaldependencies.org/docs/u/pos>

⁴<https://spacy.io/usage/linguistic-featurespos-tagging>

4 Results and Discussion

4.1 Detecting Cognitive Distortion

The task of detecting cognitive distortions is treated as a binary classification problem here. From the F1 scores given in Table 2, we can see that the SVM outperforms all the other candidate algorithms. All types of features were found to be performing best with Support vector machines.

	SIF	BERT	LIWC	POS	BERT + LIWC
Log. reg.	0.75	0.74	0.77	0.73	0.74
SVM	0.77	0.79	0.78	0.77	0.76
Decision Tree	0.65	0.67	0.67	0.66	0.64
k-NN	0.74	0.75	0.76	0.75	0.75
MLP	0.73	0.70	0.77	0.72	0.74

Table 2: The F1 scores on testing each type of features mentioned above on a 80-20 training test data split.

SIF embeddings perform very similarly to the sentence BERT embeddings. This indicates that the word order might not give much insight for this task when it comes to the semantic features.

The LIWC features, while comparable, always perform slightly better than the POS tags as features. As the POS tag embeddings have the word order encoded it, whereas LIWC features (be it semantic or syntactic) do not, this reinforces our conclusion that the word order does not contribute much to the classification task.

To get the best of semantic as well as syntactic insights, we tried a hybrid model that combines these features. This method yielded strikingly similar results to the other tests. For example, the combination of best performing semantic as well as syntactic features, i.e. S-BERT with LIWC features, still yields the highest F1-score of 0.76 by using SVM. This result may be because the combined model tends to overgeneralize in training, which in turn results in a slight decrease in performance on the test set.

4.2 Detecting the Type of Cognitive Distortion

While the aforementioned results show good performance in detecting the presence of cognitive distortions, detection of the type of distortion fails to yield good results. None of the algorithms mentioned above got a weighted F1-score more than

0.30. This could also be attributed to a poor IAA rate of $\sim 34\%$ which creates an upper bound for the performance in this task. Despite the discouraging classification results, we can draw some meaningful conclusions based on these experiments.

One way to test if semantically similar sentences tend to have the same type of distortion was to use k-Nearest Neighbors (k-NN) on the semantic embeddings using cosine similarity. When applied to the sentence-BERT embeddings, using k-NN on the multi-class classification problem yields 24% accuracy in the best case scenario as shown in figure 2.

Here, ‘the count-based k-NN’ would simply count the most redundant class of the ‘k’ nearest neighbors of a new data point, and then classify it as the same distortion. Whereas, the ‘probability based’ model applies more weight to the entries which were semantically closer to the data point in question. Both of these models perform best at $k \geq 15$. At lower values, however, the count-based model performs slightly better than the probability-based model. So, we can conclude that semantically similar sentences do not have the same cognitive distortions.

Focusing on the syntactic features, if we analyze the behavior of these distortions based on their POS tags we can draw some conclusions about the type of language used for these distortions (figure 3). For example, the distortion “labeling” had a higher probability of having Adjectives, interjections, and Punctuations. The distortion “mind-reading” has a higher probability of having Pronouns, more specifically 3rd-person pronouns. Both of these examples are in accordance with the definition of the respective distortions.

On the other hand, some findings are more unexpected. The expectation with the “should statement”

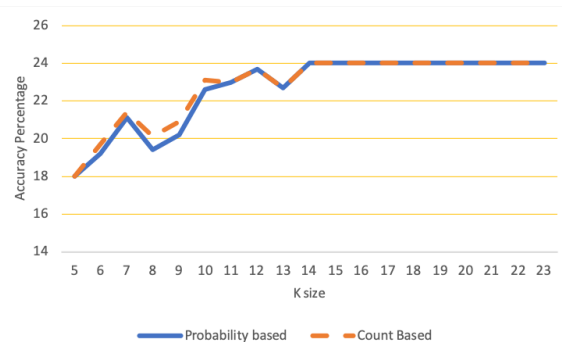


Figure 2: Performance of k-Nearest Neighbors as a multi-class classifier for Cognitive Distortions

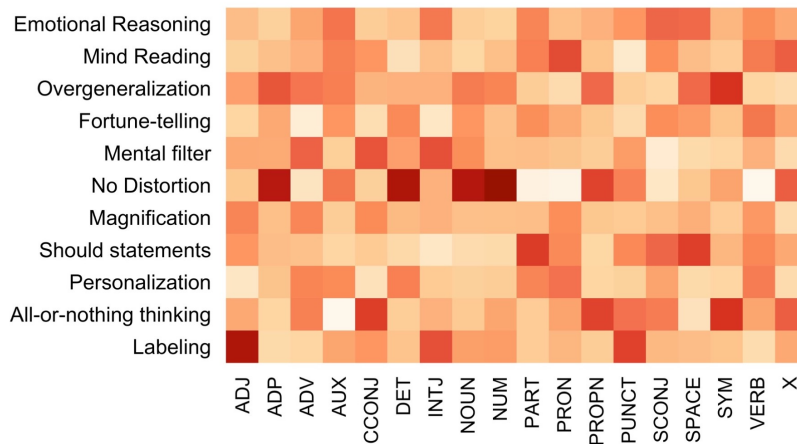


Figure 3: Heatmap showing normalized frequency of each POS tag used in different types of distortion. The darker colors indicates higher than normal frequency and the lighter colors indicate lower than normal frequency for a particular tag in the corresponding distortion.

was to have a higher probability of having auxiliary verbs such as ‘should’, ‘must’, ‘ought to’ etc. However, the results show that should statement have a lower than average probability of having auxiliary verbs. An example of this distortion without using any of the words listed above could be “While others my age are busy with their jobs and life I am just wasting my time”.

Unsurprisingly, entries having no cognitive distortions usually behave very differently than the mean behavior of distorted data (and hence the high F1-scores for the binary classification task). This can also be supported by the analysis of the LIWC features, more than 50% of the features do not conform to the patterns exhibited by other distorted entries. In addition to having the lowest score on the LIWC features - ‘feel’, ‘perception’,

‘insight’, ‘negative emotion’, ‘risk’ and ‘reward’; The non-distorted entries also tend to have more Adpositions, Determiners, Nouns, and Numerals, which indicates low subjectivity (Sahu, 2016).

Conducting a similar analysis on the LIWC features, we can conclude that some types of distortions are easier to detect than others. While most of the features of the entries conform to a mean pattern, some of the distortions deflect in behavior for specific features. For example, Fortune-telling distortion has the highest score for ‘focus future’, emotional reasoning has the highest ‘feel’ score, and so on.

Figure 4 shows a visual representation of how difficult it is to classify a certain distortion. The x-axis shows the magnitude of deviation (normalized z-score) from the mean behavior. The higher deviation from mean behavior, the easier it would be to classify that label using the LIWC features. This was done by calculating the z-score for each feature to quantify how far is the data point from the mean. The mean behavior here represents the average LIWC features expected from a natural language entry by a patient in the context of this study. This analysis is consistent with the finding that it’s easier to detect ‘No distortion’ than any of the specific types of distortions since the ‘no distortion’ category shows maximum deviation from the mean behavior.

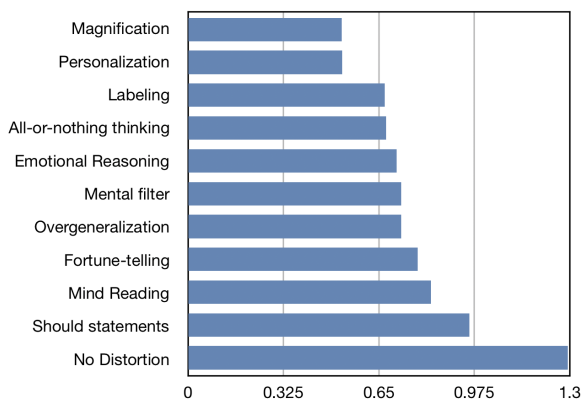


Figure 4: Normalized z-scores calculated for the types of cognitive distortion for each of 93 features of LIWC. Higher magnitude of z-score indicates higher deviation from the norm.

5 Conclusion

In this work, we compare and contrast the performance of five classification algorithms in detecting cognitive distortion.

We find that the task of determining whether or not an input indicates distorted thinking is computationally feasible, wherein semantic as well as syntactic features perform equally well. The order of words was found to have no impact on the results. Entries with cognitive distortions tend to be more subjective than non-distorted entries. The best classification results were obtained by SVM using pre-trained S-BERT embeddings with an F1 score of 0.79.

Regarding the task of identifying the type of distortion, we found that semantically similar entries do not always get categorized as the same distortion. Some of the distortions are easier to classify than others, e.g. ‘should statements’, ‘mind reading’, ‘fortune-telling’ etc. None of the implemented ML techniques obtained an F1 score higher than 0.30 on the classification of each type of cognitive distortion.

A challenging aspect of this research is getting a standardized inter-annotator agreement. One reason for that is a lack of clear distinction in psychology literature itself wherein some of these distortions are sometimes grouped as one. Another reason for this could be the presence of multiple distortions in a single patient entry (Rojas-Barahona et al., 2018).

As with the clinical application of detection algorithms, there are some ethical risks to keep in mind. If the algorithm is implemented as an unregulated flagging system, the false negatives would go undiagnosed and the false positives would be put through an unnecessary position of second-guessing their cognitive capabilities. However, 100% accuracy of classification from a single interaction (as used for training here) may not be needed for such clinical applications. If this were to be implemented in a dialogue system, an ongoing conversation with the participant will serve to make the system more accurate and personalized. As the main goal is to develop effective feedback to help any participants, having less than perfect predictions is still valuable in informing the types of feedback that an automated clinical tool could provide to the participant.

Lastly, we discuss several applications of this work in the mental healthcare sector. It could be used to flag or screen people for referrals to mental health care providers. Likewise, it could also be used in tandem with the diagnosis to establish an estimate of the severity of the anxiety or de-

pression. This approach might also be useful in detecting delusions or paranoia as well as suicide risk in natural language. Lastly, the measure of a patient’s distorted thinking can be used as an indicator of remission which can be used to determine which therapy techniques (or therapists, from the perspective of insurance companies) are more effective. In conclusion, this tool can be adapted for applications in mental health screening, diagnosis, and tracking treatment effectiveness.

6 Future work

This is an ongoing project with the ultimate goal to implement feedback to support CBT through the detection of cognitive distortions. Our next step is to implement a multi-class classification framework to improve the type of distortion detection accuracy. Once this study is complete, the annotated dataset will be made available to the public to encourage similar work in this domain.

The annotators have also identified and flagged specific parts of sentences wherein the negative thinking patterns were most evident. We can then train a classification model by using algorithms such as IOB (inside-outside-between) type tagging which can pinpoint the errors in a patient’s reasoning that give rise to cognitive distortions.

Acknowledgements

We thank the annotators: Rebecca Lee, Josh Daniels, Changing Yang, and Beilei Xiang; and Prof. Martha Palmer for funding this work. We also acknowledge the support of the Computational Linguistics, Analytics, Search and Informatics (CLASIC) department at the University of Colorado, Boulder in creating an interdisciplinary environment that is critical for research of this nature. Lastly, the critical input from three anonymous is gratefully acknowledged in improving the quality of this paper.

References

- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.
- Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.
- John Briere. 2000. *Cognitive Distortions Scale (CDS) Professional manual*.

- David D Burns and Aaron T Beck. 1999. Feeling good: The new mood therapy.
- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.
- Terje B Holmlund, Peter W Foltz, Alex S Cohen, Håvard D Johansen, Randi Sigurdson, Pål Fugelli, Dagfinn Bergsager, Jian Cheng, Jared Bernstein, Elizabeth Rosenfeld, et al. 2019. Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological assessment*, 31(3):292.
- Lauren Stephanie Jones, Emma Anderson, Maria Loades, Rebecca Barnes, and Esther Crawley. 2020. Can linguistic analysis be used to identify whether adolescents with a chronic illness are depressed? *Clinical psychology & psychotherapy*, 27(2):179–192.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#).
- Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE spoken language technology workshop (SLT)*, pages 136–143. IEEE.
- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentencebert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Lina Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. *arXiv preprint arXiv:1809.00640*.
- Ishan Sahu. 2016. *A study on detecting fact vs non-fact in news articles*. Ph.D. thesis, Indian Statistical Institute, Kolkata.
- Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE.
- T Simms, C Ramstedt, M Rich, M Richards, T Martinez, and C Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE international conference on healthcare informatics (ICHI)*, pages 508–512. IEEE.
- Yla R. Tausczik and James W. Pennebaker. 2010. [The psychological meaning of words: Liwc and computerized text analysis methods](#). *Journal of Language and Social Psychology*, 29(1):24–54.
- Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. 2019. [Linguistic analysis of schizophrenia in Reddit posts](#). In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83, Minneapolis, Minnesota. Association for Computational Linguistics.