UNIVERSITY OF TARTU
Faculty of Social Science
School of Economics and Business administration
Innovation and Technology management

Stanislav Sochynskyi

Automated cognitive distortion detection and classification of Reddit posts using machine learning

Master's Thesis (20 ECTS)

Supervisor: Kairit Sirts, PhD,
Chair of Natural Language Processing, Research Fellow in Language Technology

Tartu 2021

Automated cognitive distortion detection and classification of Reddit posts using machine learning

Abstract:

A vicious circle of exaggerated thinking patterns, also known as cognitive distortions, can lead a person to anxiety and major depression. Automatic detection and classification of cognitive distortions can be beneficial for the initial mental health screening, the better use of counselling time, and improve accessibility of mental healthcare services. In this work, we apply logistic regression, Support Vector Machines (SVM), and fasttext classifiers to identify cognitive distortions in the real-world data from Reddit. For binary classification, the best F-score of 0.71 with the fasttext classifier. For multiclass classification task, the best F-score of 0.23 was achieved with Support Vector Machine (SVM) using tf-idf vectorisation. However, the metrics of some classes do not exceed the random chance baseline. A possible explanation is that the created dataset is sufficient to build a binary classifier, but more accurate models require more data to distinguish a larger number of classes. Additionally, we experimented with unsupervised clustering and topic modelling algorithms and did not find evidence that unsupervised methods could extract the patterns of cognitive distortions from a text. We developed an annotation guideline for manual annotation of cognitive distortions and applied it to annotate 2021 Reddit posts. We achieved kappa's score of 0.569 for binary case and 0.424 for multiclass case annotation, meaning moderate agreement between annotators. A higher number of classes leads to poorer consistency in annotation agreement, mainly due to overlapping definitions of cognitive distortions. Consequently, any automated methods cannot be expected to show high results in cognitive distortion classification.

Keywords:

Machine learning, mental health, natural language processing, cognitive distortions, data annotation

CERCS: P176 Artificial Intelligence, S260 Psychology.

Automaatne kognitiivsete kallete tuvastamine ja klassifitseerimine Redditi postitustest kasutades masinõpet

Lühikokkuvõte:

Liialdatud mõttemustrite nõiaring, mida nimetatakse ka kognitiivseteks kalleteks, võib inimese viia ärevuse ja depressioonini. Kognitiivsete kallete automaatne avastamine ja liigitamine võib olla kasulik esmaseks vaimse tervise sõeluuringuks, nõustamisaja paremaks kasutamiseks ning vaimse tervise teenuste kättesaadavuse parandamiseks. Selles töös kasutame logistilise regressiooni, tugivektormasina (SVM) ning fasttexti klassifikaatoreid, et tuvastada kognitiivseid kaldeid Redditist pärit reaalmaailma andmetest. Binaarsel klassifitseerimisel saadi parim F-skoor 0,71 fasttext klassifikaatoriga. Erinevate kognitiivsete kallete eristamise klassifikatsiooniülesande puhul saavutati parim F-skoor 0,23 tugivektormasinaga, kasutades tf-idf vektorisatsiooni. Siiski ei ületa siin mõne klassi mõõdikud juhusliku klassifitseerija piiri. Võimalik seletus on see, et loodud andmestik on binaarse klassifikaatori treenimiseks piisav, kuid täpsemad mudelid nõuavad suurema arvu klasside eristamiseks rohkem andmeid. Lisaks eksperimenteerisime juhendamata klasterdamise ja teemamudelite algoritmidega ning ei leidnud tõendeid selle kohta, et juhendamata meetodid suudaksid tekstist tuvastada kognitiivsete kallete mustreid. Töötasime välja annoteerimisjuhised kognitiivsete kallete käsitsi märgendamiseks ning rakendasime seda 2021 Redditi postituse annoteerimiseks. Annoteerijatevahelise kooskõla kappa skoor oli 0,569 binaarsel juhul ja 0,424 erinevate kaldetüüpide annotatsioonide puhul, mis tähendab

mõõdukat kooskõla annoteerijate vahel. Suurem arv klasse toob kaasa kehvema kooskõla annotatsioonides, mis on tingitud peamiselt sellest, et mõnede kognitiivsete kallete määratlusted kattuvad. Seega ei saa automatiseeritud meetoditelt eeldada kõrgeid tulemusi erinevate kognitiivsete kallete eristamisel.

Võtmesõnad:

Masinõppimine, vaimne tervis, loomuliku keele töötlus, kognitiivne moonutus, andmed annotatsioon

CERCS: P176 Tehisintellekt, S260 Psühholoogia

Table of Contents

1 Introduction

World Health Organisation claimed depression to be a world problem impacting 4% of the population [1]. Depression impacts the person itself and affects relationships with friends, family members, and anyone who is in deep care for the depressed person and are willing to help. On top of that, according to Sobocki, et al. [2], the estimated cost of depression for EU economics is 136 billion EUR yearly. Cognitive Behavioural Therapy (CBT) shows promising results in dealing with mental health issues [3]. CBT helps identify irrational thinking patterns, also called cognitive distortions [4], which related to depression[5], anxiety [6], and suicide ideation [7]. Cognitive distortions often are expressed in an individual's speech and writing. For example, after not arriving first in a marathon race, a person may say, *"I'm such a limp duck. I'll never win a race."* This sentence contains 2 out of 15 common cognitive distortions [8]: *"I'm such a loser"* - labelling; *"I'll never feel good again"* - overgeneralisation. As distorted perception put a person at significant risk of major depression development, it raises the importance of detecting cognitive distortion in the early stages [9].

Natural language processing is a technology that enables understanding of human language, both speech and writing, and extracts meaning from it. With the growth of the machine learning field, CBT online therapy and tools based on NLP techniques positively enhanced mental health diagnosis and treatment, making mental health check-ups more affordable [10][11]. Various studies aimed to classify persons' status (i.e., depressive, suicidal, anxious) from social media such as Twitter and Facebook. For example, Kramer et al. [12] studied emotional contagion by manipulating Facebook Newsfeed, which led to an increase or decrease of positive or negative posts produced by the person. Other studies present results on depression [13] and suicidal ideation [14] detection from Twitter posts or trained model to provide support for moderators of youth mental health online forums [15].

To our knowledge, there are a few studies that focused on cognitive distortion detection from texts. In 2012, Moris et al. [16] selected five cognitive distortions and asked 73 participants to manually annotate as distorted and not distorted 32 short texts taken from literature and online resources. The authors reported high accuracy for manual detection, but this study does not include a training description used to teach annotators how to detect cognitive distortions. The first attempt for automated detection and classification was conducted by Shickel et al. [17]. In the article, authors expanded the number of distortions up to 15 and used three different datasets: Amazon Mechanical Turk[1] workers were asked to provide personal life examples to fit a distortion description. The other two real-world datasets were shared by TAO Connect[2] and annotated with the help of four students from the psychology faculty. However, papers lack discussion on the data annotation process and do not include used annotation guideline. As the results of their work, Shickel et al. formed three newly datasets, nevertheless, authors highlight that there is still a substantial shortage of annotated datasets in this domain. They also mention that Mechanical Turk dataset could be biased as platform workers were framed with a cognitive distortion definition. Shickel et al. do not make any claims about detection and classification models baseline achieved with TAO Connect datasets, however, they consider real-world data more promising considering application of future results.

[1] https://www.mturk.com/
[2] https://www.taoconnect.org/

Reddit[3] is the 5th most visited online discussion platform in the US, with over 130K communities organised around readers' interests. The website offers an opportunity for every user to create subforums titled "subreddits" for everyday topics from sports and friendships and more complicated topics such as bullying at school, addictions, depression, and other issues related to mental health. It appears to be a suitable source of real-world data as users there are not prompted and imposed with any limitations in terms of lexicon, size of the posts, and structure that they should follow.

In this thesis, different unsupervised and supervised techniques will be examined to identify cognitive distortions from real-world posts published on Reddit. As for the supervised methods, the goal is to set up a baseline for detection and classification tasks as all previous works on the same topic have their problem. We will develop an annotation guideline and use it to annotate data so that it will be possible to build models using logistics regression, SVM and fasttext algorithms. Built models will be evaluated against a standard set of classification metrics, and results will be discussed. We will also analyse the agreement score between two annotators on a subset of annotated posts and highlight challenges that occurred throughout the annotation process. The aim of studying unsupervised methods such as K-means, BIRCH and topic modelling to analyse how well they can capture cognitive distortions from real data. It whether there is a potential for application of an unsupervised learning approach in mental health.

The main contributions of this thesis can be summarised as follows:

- To our knowledge, this is the first work that develops and shares annotation guideline for manual detection and classification of multiple cognitive distortions based on a given text.
- We establish a baseline for cognitive distortions detection and classification using Reddit data annotated according to the developed annotation guideline.

The rest of the thesis is organised as follows: **Chapter 2** introduces works on the intersection of NLP and psychology in general and cognitive distortions in particular. In addition, we examine papers that discuss the data annotation for mental health datasets. **Chapter 3** explains technical methods applied in Chapter 5 and Chapter 6. Chapter 4 includes information about dataset and describes annotation guideline. In **Chapter 5** we report the experiment results for clustering and topic modelling, and **Chapter 6** includes experiment outcome of a supervised classification experiment. Finally, **Chapter 7** summarises thesis findings and concludes it with potential improvements.

[3] https://www.reddit.com/

2 Related works

This subchapter presents an overview of works that were completed on the intersection of NLP and mental health, analyse works that have been done on cognitive distortion detection and classification, explores articles related to mental health data annotation and outlines research goals for this thesis.

2.1 Mental health and NLP

The application of NLP techniques for the mental health domain is a growing subfield of data science. There has been quite much work in this area involving different goals—depression detection, identification of PTSD and suicidal identical, and focuses on already developed disorders rather than early signs.

Pestian et al. [18] demonstrated results on emotion detection from anonymised suicide notes. De Choudhury et al. [19] worked on predicting clinically depressed individuals using Twitter posts. O'Dea et al. [20] focused on forming a quick response to assist moderators who work with suicidal ideation. Coppersmith et al. [21] used Twitter publications to identify depression and PTSD on Twitter.

Moreno et al. [22] examined status updates on Facebook and revealed that it is possible to identify symptoms of major depressive episodes. Similarly, Kotikalapudi et al., [23] study investigate college students' web activity to detect signals of depression. Perušic et al. [24] built a binary classifier to predict early signs of depression. They evaluated the importance and identified the most informative semantic categories of features such as sentence length, positive motions. In 2016 Guntuku et al. [25] conducted a large comparison metastudy on detecting depression from four social media sources of data. It concludes that screening based on social media can add value to a mental health screening strategy.

2.2 Cognitive distortion detection and classification

Cognitive distortions are an early signal of developing depression and anxiety. There are a few research papers done on distorted thinking patterns identification and classification. The detection (or binary classification) task aims to identify the presence or absence of cognitive distortion in a given document. The classification (multiclass classification) task aims to label distorted documents with cognitive distortions from the pre-defined set of labels/cognitive distortions. The most relevant research was published in 2020 by Shickel et al. [17]. Authors worked with supervised (e.g., logistic regression, XGBoost, RNN) methods for the automatic detection and classification 15 [26] the most common cognitive distortions. They also experimented with unsupervised (hierarchical clustering) and statistical methods (Topic Modeling) to detect natural and hierarchical groupings of cognitive distortions that share common traits. For the detection and classification tasks, researchers collected three datasets: CrowdDist, MH-C and MH-D. Median passage length for all sets varies from 42 to 47 words (~2-3 sentences).

To form a CrowdDist dataset Amazon Mechanical Turk (MTurk) workers were presented with definitions of cognitive distortions. They were asked to describe events from their life that fall into the definition scope of the presented cognitive distortions. The final dataset contains 7,666 responses from 1,788 unique individuals with 511 documents per distortion on average. Authors admit the CrowdDist dataset limitation: presented description of cognitive distortion could have narrowed and prompted MTurk workers to fit their memories into the cognitive distortion description. This creates a barrier to predict cognitive distortions from the real data. People might use different wording to describe the same event when this event happened compared to the words and expressions that will be used a few

days after the event happened. The results of topic modelling reported in work show that people used wording that is expected to be seen in a specific cognitive distortion, which indirectly proves our concerns about unbiasedness of CrowdDist dataset.

The MH-C and MH-D datasets consist of journal entries provided by an online mental health therapy service for students TAO Connect. The size of MH-C is 1,164 entries with 15 label types, and MH-D consists of 1,799 entries with binary labels. The dataset is imbalanced and cognitive distortions are not evenly represented. Only MH-D contains 194 entries labelled as *"Not Distorted"*. All entries were annotated by four undergraduate students from the psychology department. Annotators included only those entries when the majority of annotators agreed on the label. There is no information on whether any annotation guidelines were created for the annotation purposes and the training procedure. The issue with the dataset comes from the service's target audience: TAO Connect aims to help college students who suffer from anxiety and depression. Unfortunately, cognitive distortions not only manifest in college and occur at work, in family relationship, and in day-to-day life situations that might not be related to studies at all. This brings risks for the model not to be able to detect and classify cognitive distortions from a non-college perspective. Being focused only on students' entries limits the capabilities of models to detect and classify cognitive distortions in different environments. Also, customers of mental health online services might be more educated about different concepts in psychology, or a service can offer the specific pattern "how-to" to make entries. This might narrow the vocabulary to limited scope compared to what a person would write, giving them no limits and no "how-to" guidance.

Shickel et al. reported results only for logistic regression as the best working method for detection and classification tasks. They achieved a high F-score in the detection task for predicting positive label (F-score = 0.95), but this is expected as the data was heavily imbalanced towards texts with a distorted label. For multiclass classification task authors reported F-score equal to 0.45, but trained model is not able to predict 7 out of 15 distortions: "Being Right", "Blaming", "Fallacy of Change", "Fallacy of Fairness", "Global Labeling", "Heaven's Reward Fallacy" and "Personalisation".

However, Shickel et al. is not the first attempt to detect cognitive distortions. Morris et al. [16] proposed a tool that uses crowdsourcing collective intelligence to assist individuals to alter the emotional response to stressful situations and life events. As a part of the outlined framework for empathetic responses, workers analyse input texts to detect cognitive distortion. In case cognitive distortion was identified, it is then classified between one of the five given labels: "Overgeneralisation", "Catastrophizing", "All-or-Nothing", "Fortune Telling", "Reverse fortune-telling". After that, the worker generates empathetic answers using cognitive restructuring, which is easy to teach, as the author highlights.

As a part of the study, Morris et. al trained 73 participants from Amazon's Mechanical Turk and asked each of them to annotate the set of 32 input statements as "Distorted" or "Not Distorted". The length of input texts was limited to 3 sentences to ease the annotation. With minimal instructions and provided framework, authors report 89% accuracy for cognitive distortion detection tasks. Authors claim that the relatively easy concept of cognitive distortions, and high accuracy of non-expert annotation, gives the confidence that it is possible to build automated tools for detecting and classifying cognitive distortions.

One more paper by Xuejiao et al. [27] proposed a CNN-based system to detect cognitive distortions based on daily mood logs and automatic patient thoughts. The justification for using automatic thoughts adds to what discussion about the CrowDist dataset: very often, people cannot detect cognitive distortions by themselves, and therapists can not follow pa-

tients 24/7. Although authors were the first who considered using word embeddings(word2vec) for text vectorisation in the system, they did not run any experiments, and it is rather a theoretical architecture.

2.3 Annotation Studies

All the work on cognitive distortion on detection and classification generally describes the annotation process and highlights the absence of text datasets. Shortage of data and annotation guidelines brings challenges in terms of the reproducibility of studies. It sets high barriers to start new ones: instead of developing new methods and approaches, authors first had to create annotation guidelines annotate data.

The subfield of annotation guidelines for mental health is not well developed with very few solid works, including an annotation guideline. It describes the annotation process's challenges for social media data. For example, Mowery et al. [28] developed an annotation guideline for major depressive disorder based on DSM-5[29] depression criteria and psycho-social stressors. They ran an experiment and annotated Twitter data with created guidelines and included annotator agreement score results. The authors identified some depressive symptoms and psycho-social stressors; however, they highlighted considerable challenges in the annotation process of mental health symptoms. Earlier, Homan et al. [30] created a 4-value distress scale for rating tweets, with annotations performed by novice and expert annotators. Milne et al. [15] complete a shared task on triaging content in mental health peer-support forum. In this work, the authors implemented a tool that supports forum moderators and provides information on which forum participants require urgent help as they in the risk zone of committing suicide. They split messages from forum participants into four categories: Green, Amber, Red and crisis. One out of four different colours were assigned to texts concerning the urgency of the inquiry. To complete the classification task, they labelled 1227 observations in total. In order to annotate posts correctly, the authors wrote a set of questions and created an annotation decision tree that helped them make decisions about what class what post to assign.

2.4 Research goals

In our work, we see the potential to study (1) the application of supervised methods to detect and classify cognitive distortions on the real-world ("wild ") data that were not prompted and attempt to set up a baseline for these tasks. We also want to (2) continue experiments with unsupervised clustering algorithms and topic modelling to see whether these methods can capture the reflection of cognitive distortions in real-world data. This work is based on data extracted from the Reddit discussion board as there are no restrictions imposed in terms of vocabulary and how-to word expression of feelings. The real-world dataset also might broaden the number of environments where distorted thinking patterns can occur, e.g. workplace, home, and day-to-day life. This is the first work (3) that will contain data annotation guideline, describes the process of annotation and examines agreement score difference between 2 independent annotators. Based on the reported results by Shickel et al., we also see the potential to decrease the number of cognitive distortions for classification task to avoid problem zero value metrics. Annotated dataset is also enhanced with the negative labels, and we expect to achieve a balance between the representation of each cognitive distortion.

3 Technical background

This chapter aims to introduce various techniques and algorithms for work with textual data that were further applied.

3.1 Text vectorisation

The main object of NLP research is text. For example, mail messages, logs, messages in chats, speech, and even images. A document (text) is an object consisting of words that are organised in a specific order. Words are usually called tokens, and a corpus or dataset is a set/collection of documents. In order to work with texts, it is necessary to turn them into a mathematical representation – from each text, form a vector with which a computer can work. The process of transformation of text into its numerical presentation (vector) called vectorisation [31].

3.1.1 Bag of Words (BoW) term frequency

To describe various vectorisation techniques, let us consider the following corpus of 4 sentences (1) as an example:
1. A fluffy kitten purred.
2. A fluffy cat purred and meowed.
3. A fluffy kitten meowed.
4. A loud fluffy poodle ran and barked.

One of the methods is to count occurrences of a particular token (word) in the dataset. This is the simplest representation which is called a bag of words [32]. To build up this representation, the frequency of all the unique tokens (words) that appeared in all documents is calculated. As a result, initial texts are transformed into a term-document matrix. Columns of this matrix represent sentences (documents in our case), and rows contain unique tokens that form a dictionary of the dataset (see Table 1).

Table 1 Term-document matrix[4]

	Sentence #1	Sentence #2	Sentence #3	Sentence #4
kitten	1	0	1	0
fluffy	1	1	1	1
cat	0	1	0	0
meowed	0	1	1	0

However, there are a few drawbacks to this approach to consider. The word order is lost, and it is possible to shuffle any word in the sentences, and representation will remain the same [33]. These vectors have no semantic meaning. For example, there is a vector for a cat and a kitten - in the BoW approach, they are not related in any way. Another problem is scalability. In real tasks and challenges, the size of the dictionary is usually very large - it

[4] only 3 words were used to reduce the amount of occupied space

can be hundreds of thousands of words, or even several million. This results in very long sparse vectors and blocks usage of non-linear methods since the vectors are large.

3.1.2 Term Frequency-inverse document frequency (tf-idf)

To overcome the problem of sparse vectors, some tokens can be removed on the occurrence frequency in documents of the corpus. For example, prepositions, articles are very frequent but do not add extra value to discriminate texts as they only exist for grammatical structure. On the other side of the scale are low-frequency tokens such as typos or rare words that are not seen in other documents. This can cause a model to learn dependencies that are actually not there.

To select medium frequency term frequency-inverse document frequency (tf-idf) algorithm used. This method assigns a higher value if the term has a high frequency in the given document and low term frequency in the whole collection of documents [33][34]. The term frequency (tf) is be calculated as follows [34]:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ is the number of times the term appeared in document d and $\sum_{t' \in d} f_{t',d}$ - total number of terms in a document.

The inverse document frequency (idf) calculates the frequency of the word in the entire corpus. It decrease the weights for frequent words and increases the weight of rare words, and can be calculated as follows:

$$idf(t,D) = log \frac{|D|}{|\{d \in D : t \in d\}|},$$

where |D| is a total number of documents in the corpus; $|\{d \in D : t \in D\}|$is the number of documents where the token t appears.

The tf-idf term value is calculated as follow:

$$tf\ idf(t,d,D) = tf(t,d) \cdot idf(t,D).$$

Although the vocabulary can be optimised with TF-ID, it still does not capture position terms order, semantics, and terms co-occurrences in a different context compared to word embeddings.

3.1.3 Word embedding

Word Embeddings is a concept presented by Mikilov, and Le is another vectorisation technique that generates distributed representation vectors out of terms and helps to carry semantic meaning [35]. Doc2vec [36] [37] is a document embedding technique that enables figuring out various relationships between words. Besides the words, doc2vec also uses document ID in training. In that case, documents are treated in the same way as if they are words: there is ID in a collection of documents and based on that, the embedding will be built for a document.

There are two architectures of doc2vec: Distribute Memory (DM) and Distributed Bag of Words (DBOW). The main difference is that DM attempts to predict a focus word given context words and paragraph ID. At the same time, DBOW stands for providing probability to provide the context given the document [38]:

$$DM = \rho(w_i|w_{i-h}, \dots, w_{i+h}|\boldsymbol{d}),$$

$$DBOW = \rho(w_{i-h}, \dots, w_{i+h}|\boldsymbol{d}),$$

For example, taking corpus (1) and attempt to predict kitty, cat, poodle using DM.

Context words for kitty: [fluffy, meowed, purred], (ID - 1,3)

Context words for cat: [fluffy, purred], (ID - 2)

Context words for poodle: [fluffy, ran], (ID - 4)

Table 2 Paragraph Matrix

	fluffy	purred	barked	meowed	loud	ran
kitten	2	1	0	1	0	0
cat	1	1	0	0	0	0
poodle	1	0	0	0	0	1

The matrix (Table 2) is passed to the neural network with one layer. As a result, doc2vec generates a vector for the word and the document.

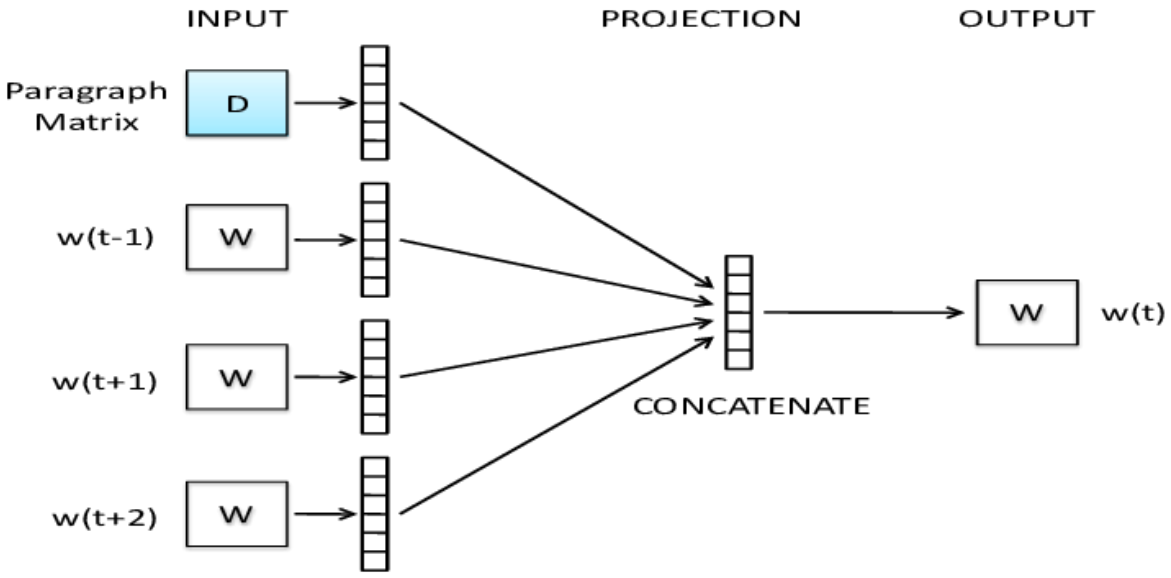

Figure 3.1 doc2vec Distribute memory architecture[5]

In doc2vec method, multiple neurons capture a single concept (e.g. word meaning, part of speech), and also, a single neuron contributes to multiple concepts to form distributed representation of a document. These concepts are not pre-defined and are learnt through work of algorithm, and that why they can be used in different contexts.

[5] https://www.researchgate.net/figure/Doc2vec-model-with-a-distributed-memory-method_fig4_320653820

3.2 Clustering

Text clustering belongs to a class of unsupervised machine learning task. Clustering algorithms group a collection of unlabelled documents so that texts in the same cluster are more similar to each other than those in other clusters. Text clustering algorithms process data and determine if natural clusters (groups) exist in the data.

3.2.1 K-means

K-means is the most known vector-based clustering algorithm that requires a pre-defined number of clusters as input [39]. It returns a label from a pre-defined number of clusters for each data point. K-means starts from random initialisation of K cluster centroids in the vector space (**Error! Reference source not found.**, b) [40].

$$\mu_1, \mu_2, \dots, \mu_k \in R^n$$

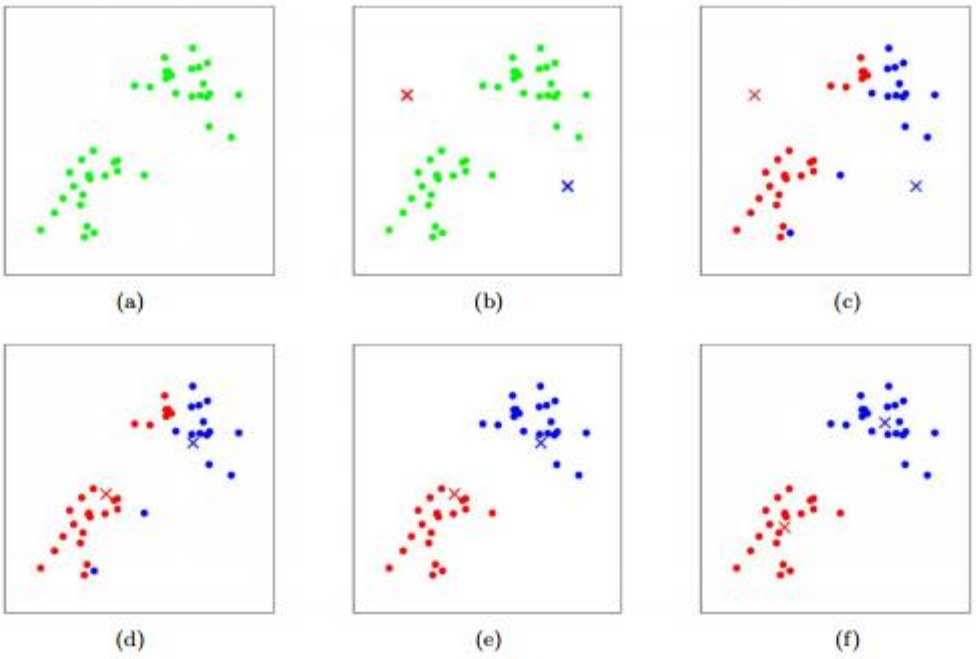

Figure 3.2 K-means algorithm[6]

After that algorithm computes the distance metric between every observation and all clusters' centres, observations are assigned to the closest cluster based on the lowest distance (**Error! Reference source not found.**, c).

$$c^{(i)} = \underset{j}{\operatorname{argmin}} \left\| x^{(i)} - \mu_j \right\|^2,$$

where $c^{(i)}$ – observation label. After all, observations are assigned to new clusters, centres of clusters are recalculated based on values of newel added observations within a given cluster (**Error! Reference source not found.**, d).

[6] https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

Steps (c) and (d) are repeated until the convergence is achieved so that the distance between 2 distinctive clusters is maximum, but the distance between observations within one cluster is minimum. K-means is the easiest algorithm to understand and interpret. It has a low execution time, comparing to other methods.

3.2.2 Birch algorithm

Balanced Iterative Reducing and Clustering using Hierarchies **(**BIRCH) [41] is a hierarchical algorithm that capable of dealing with problem of scaling dataset size by generating a height-balanced tree Clustering Feature Tree (CFT) with a collection called Characteristic Feature Nodes (CFN). Every CFN has Characteristic Feature Subclusters (CFS) that contain compact summaries about data. Every CF contains several samples in subcluster, linear sum, squared sum, centroids and squared norm of centroids. BIRCH performs clustering using the data summaries instead of the original dataset. This approach allows optimising memory needed for computation. BIRCH does not work with categorical data and can only deal with metric attributes represented by explicit coordinates in a Euclidean space. Also, the quality of Birch performance decreases on a high dimensional dataset.

Zhang et al. [42] outlined the work of BIRCH clustering algorithm in four phases. It starts from building an initial CF tree based on the scanned data within the memory given and results in a summary of the data. The second phase is optimal preprocessing to improve performance speed and quality. At this stage, the algorithm scans smaller CF tree to dense group subclusters into large ones and remove outliers. In the third phase, all leaf entries are clustered using existing clustering algorithms (e.g. K-means). As a result, the collection of clusters the fall into main distribution pattern is generated. During 4th phase, BIECH uses centroids calculated in phase 3 as seeds, and reassign data point to the closest seed to ensures that all copies of a single data point belongs to the same cluster and move them to one cluster if they are not.

3.2.3 Optimal number of clusters

As K-mean and Birch require a fixed number of clusters as an input, a very common task for clustering is to find out the optimal number of clusters as the number of ground truth labels is unknown. For that purposes, different techniques are used to calculate scores and

Elbow method

One of the most common approaches to finding an optimal number of clusters is applying the elbow method [43]. It runs a clustering algorithm multiple times, with an increasing number of cluster choice. The smaller the inertia, the denser the cluster. The score is calculated as the sum of square distances from each point to the assigned cluster. After that, each cluster score is plotted as a function, and a sharp downshift of a line (and forming elbow) should show the optimal number of clusters [44].

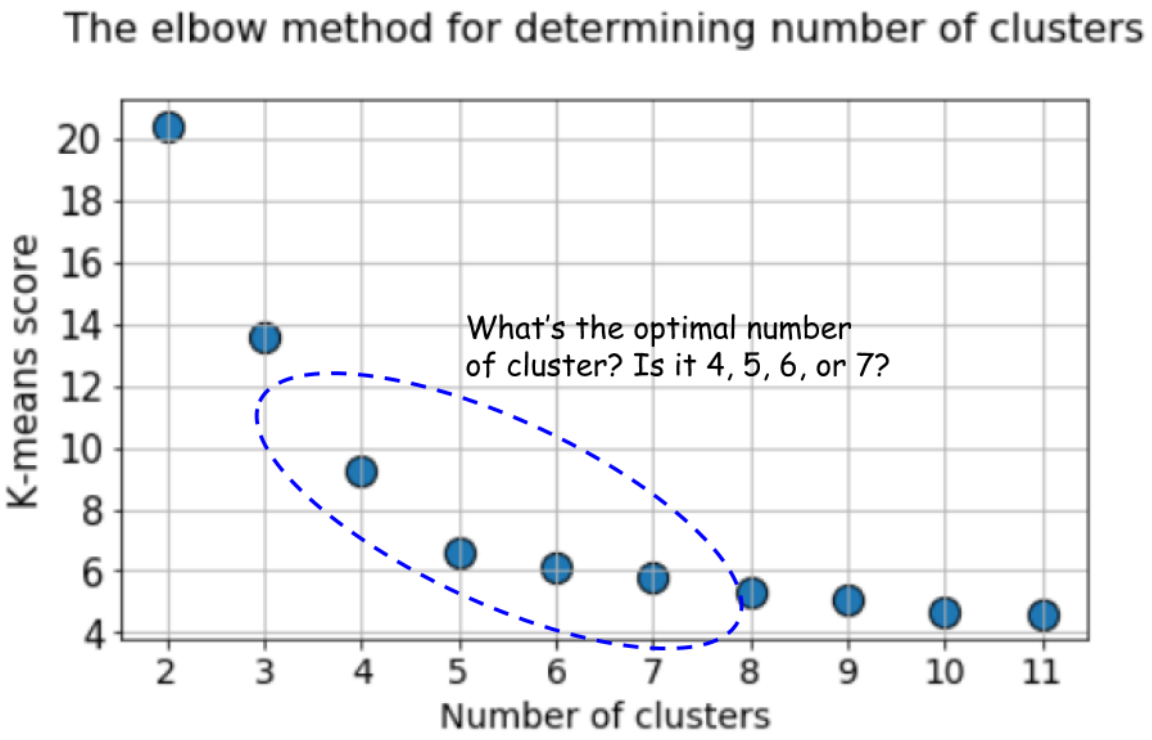

Figure 3.3 Output example of elbow method [43]

Figure 3.3 presents an example of issues that might occur with Elbow method. The graph line is smoothing without forming a clear elbow, and it is not evident what number of clusters to choose. Silhouette score (Figure 3.4) and Davies-Bouldin Index can help to overcome this problem.

Silhouette score

Silhouette score [45] is used to measure how each point in one cluster is similar to the other cluster's points and the neighbouring clusters. The score is ranged from -1 to 1, where zero value indicates overlapping clusters, negative values indicate that those samples might have been assigned to the wrong cluster. A higher score indicates more distinct and dense clusters, i.e. sample is far from the neighbouring clusters. The silhouette coefficient [46] is calculated as follows:

$$s = \frac{b - a}{\max(a, b)},$$

where a is the average distance between a given point and all others that belongs to this cluster; b is the average distance between an observation and all other points.

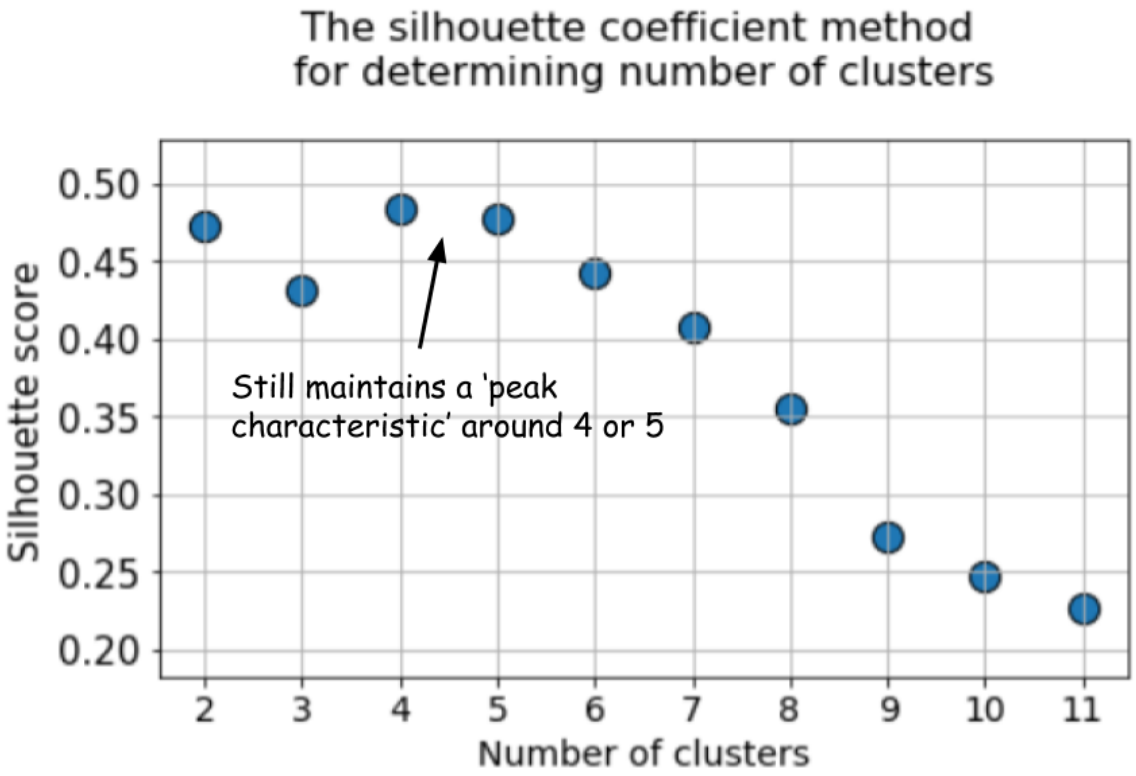

Figure 3.4 Example of silhouette coefficient score

Davies Bouldin index

Just like Silhouette score, Davies-Bouldin index [47] signifies the average 'similarity' between clusters without ground truth labels. The similarity is the ratio of within-cluster distances to between-cluster distances. Less dispersed clusters and larger distance between each cluster results in a better score. However, a method is limited to usage of Euclidean distance function since it calculates the distance between clusters' centroids

The Davies-Boulding score is calculated as an average similarity between each cluster C_i and similar C_j[48]. The mathematical formulation is a follows:

$$db = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j}\left(\frac{s_i + s_j}{d_{ij}}\right),$$

where $s_{i,j}$ −the mean distance between each point of cluster i and cluster centroid (cluster diameter), d_{ij} − distance between cluster centroids i and j.

3.2.4 Clustering visualisation

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that captures global structure of the data [49][50] and helps to visualise clustering results. The major drawback of PCA is that it will fail to maintain the local structures of the dataset. t-distribution Stochastic Neighborhood Embedding (t-SNE) [51] [52] is a non-linear visualisation method that is better at capturing relations between neighbours and maintaining local structures. The distance between two points in the visual space is embedded using the probability distribution of pairwise similarities in the higher dimensionality; thus, t-SNE

shows clusters of similar documents and the relationships between groups of documents as a scatter plot. t-SNE is computationally expensive, so typically, a simpler decomposition method such as PCA first.

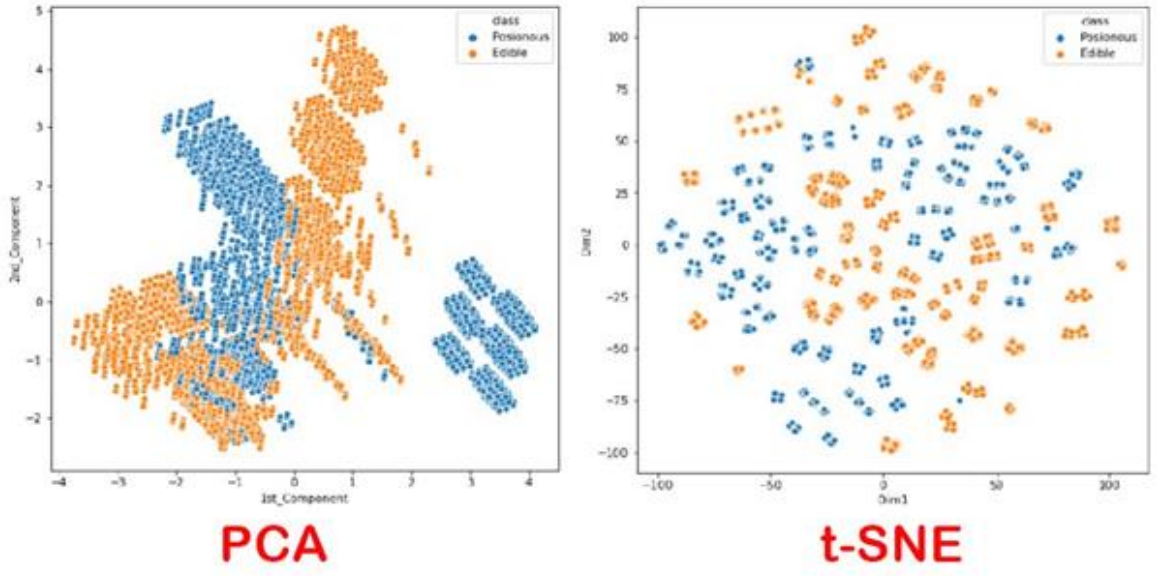

Figure 3.5 PCA and t-SNE visualisation examples[7]

3.3 Topic modelling

Topic is a group of tokens or keywords in a corpus that frequently occur together in the document and have similar tf-idf intervals. Topic modelling is an unsupervised text mining approach that automatically discovers similarities in a text corpus and helps discover its underlying topic structures.

Latent Dirichlet Allocation (LDA) [53] is a generative probabilistic model of text corpus that considers each document as a collection of topics with specific groups of keywords. The Dirichlet model assumes that before saying something, a person rolls a dice twice:

- Roll #1. To decide what topic is going to be discussed? Topics in the document are modelled as a Dirichlet probability distribution.
- Roll #2: To decide what word will be used. The words in the topic are modelled as a different Dirichlet probability distribution

Every topic in the LDA model is treated as a distribution over words existing in the documents collection. Algorithm find sets of topic-keywords distribution that are more likely to be in one collection by changes the order of the topics distribution within the documents and keywords distribution within the topics [54]. Figure 3.6 shows general LDA model representation.

[7] https://medium.com/swlh/everything-about-t-sne-dde964f0a8c1

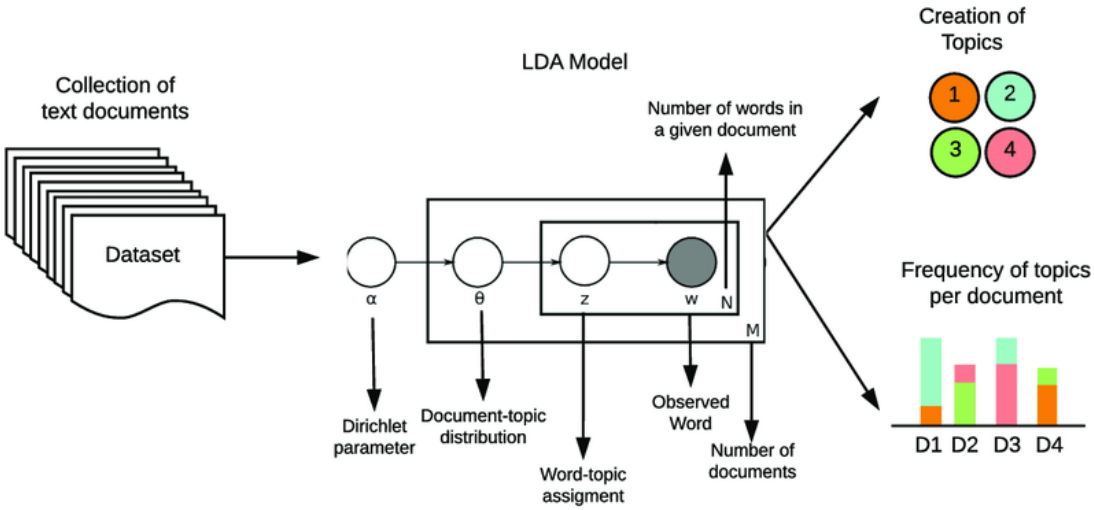

Figure 3.6 Graphical model representation of LDA [55]

The large box represents a corpus, while the smaller box represents the repeated choice of topics and words within a document. α is a corpus-level parameter, and supposed to be sampled once during corpus generation. The variable θ is a document-level variable and sampled once per document. Finally, the variables z and w are word-level variables and are sampled once for each word in each document.

3.4 Classification

Text classification is a basic supervised machine learning task aiming to classify text into different categories based on a labelled dataset. This task is based on "training" and "validate" approach. Firstly, the algorithm learns on training dataset, and during "validate" process trained model is tested on train and validation datasets. The quality of a built model is estimated based using various classification metrics.

3.4.1 Logistic regression

Logistics regression, also called the sigmoid function, is a non-linear machine learning algorithm that classifies observations by estimating the probability that observation belongs to a particular category [56]. It can take any numeric or categorical value and transform it into a value between 0 and 1, but not exceeding those limits. In general case, the formula of Logistic Regression can be written as following [57]::

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots,$$

where P is probability of predicted event, $b_0; b_1; b_2$ ... −regression coefficients, $X_1; X_2$ ... − independent variables. Exponentiate both sides of the equation we get:

$$\frac{P}{1-P} = e^{b_0+b_1X_1+b_2X_2+\cdots}.$$

The ratio $\frac{P}{1-P}$ is called odds. Odds are defined as the ratio of the probability of any event and the probability of an alternative event.

The formula of probability *P:*

$$P = \frac{e^{b_0+b_1X_1+b_2X_2+\cdots}}{1+e^{b_0+b_1X_1+b_2X_2+\cdots}} = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\cdots)}}$$

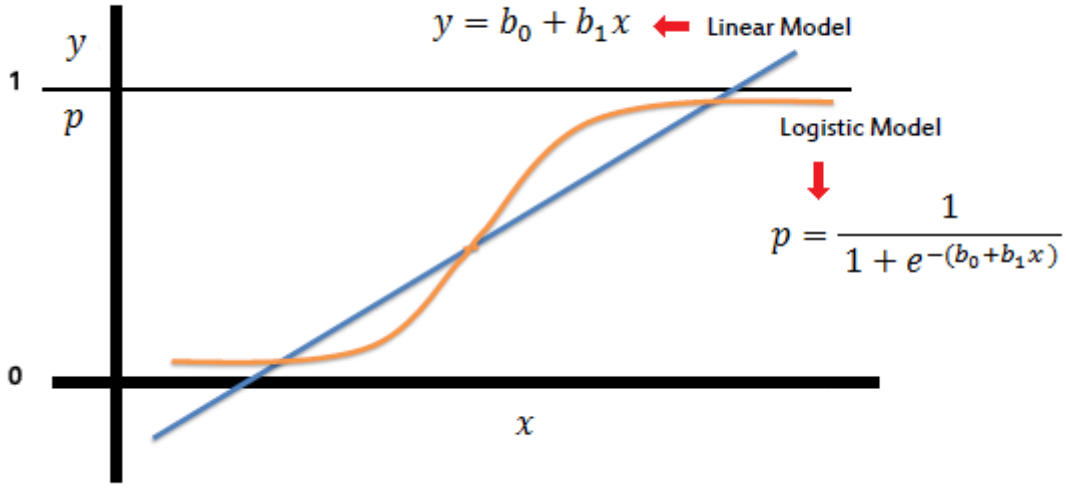

Figure 3.7 Logistic regression function [57]

3.4.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning method used for classification and regression tasks. SVM method finds an optimal separating hyperplane to determine which category a new data point belongs to by maximising the distance between two linearly separable data subsets [58]. SVM method is applied in different fields, such as classification of text or images and recognising handwritten characters.

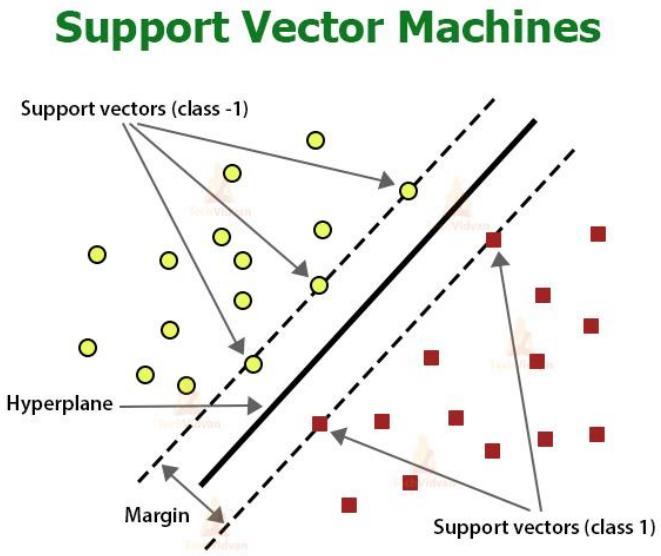

Figure 3.8 SVM representation [59]

For example, for the two-dimensional example illustrated in Figure 3.7, SVM finds an optimal line that lies as far from the nearest class data points as possible. The dashed lines are called support vectors. If the input data is not linearly separable, SVM uses a kernel function to map the data into a higher-dimensional space to make it linearly separable [60]. The hyperplane is optimal if it gives the largest minimum distance to the observations of the training dataset.

3.4.3 fasttext

FastText is an open-source, free library by Facebook AI Research for text classification and word vector representation that morphological structure of a word carries important information about the meaning of the word [61]. Comparing to doc2vec, in fasttext, similarity is not just the co-occurrence of words. The general idea of fasttext is to enrich embeddings with subword information. fasttext goes one layer deeper and split the word in a set of n-grams treat, learn embeddings from n-grams and calculate final embedding as the sum of all n-grams [61][62].

3.5 Hyperparameter tuning

The tuning parameters are the process of searching for an ideal set of parameters for model architecture. Hyperparameters are parameters that can be adjusted and fine-tuned to improve the quality of the machine learning model. Hyperrapameters often control the complexity of the model that affects any variance-base trade-off that can be made. [63]

Grid search is arguably the most basic hyperparameter tuning method. This technique simply builds a model for each possible combination of all of the hyperparameter values provided. Then algorithm evaluates each model and selects the architecture which produces the best results based on the outlined metrics, e.g. accuracy, recall. [64]

3.6 Classification evaluation metrics

There are various measures for evaluating a classification model, and they can all be driven from the confusion matrix [65].

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. A confusion matrix is built up based on the following terms:

- True positive (TP): shows how many of those items that were predicted as positive are indeed positive.
- True negative (TN): shows how many of those items that were predicted as negative are truly positive.
- False positive (FP or Type I error): shows how many of those items that were predicted as positive are actually negative.
- False negative (FN or Type II error): shows how many of those items that were predicted as negative are actually positive.

The first standard evaluation measure is accuracy. It is the proportion of actual results among the total number of cases examined. Accuracy is a good choice of evaluation for classification problems which not skewed or has low or no class imbalance. In the opposite case, this metric will be biased towards the majority class. Mathematical formulation of accuracy is:

$$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)}$$

A model can be reasonably accurate but not valuable for cases that rarely happen. Thus, if the aim is to assess class-related classification performance, typically precision, recall and F1-score are used. Precision estimates how many positively identified samples were correct, while recall estimates the correctly identified proportion of positive samples.

$$Recall = \frac{TP}{(TP + FN)}$$

$$Precesion = \frac{TP}{(TP + FP)}$$

For an imbalanced dataset, a weighted F-score is a better measure. F value is the harmonic mean of precision, and recall maintains a balance between the precision and recall for a built classifier. If precision is low, the F1 is low, and if the recall is low, the F score is also lower. The mathematical formulation is as follows:

$$F1 = 2 * \frac{precesion * recall}{precesion + recall}$$

The main problem with the F1 score is that it gives equal weight to precision and recall [66]. F-score could also be dependent on support, i.e. the number of true instances for each label when the weighted average is calculated.

4 Annotation and data

This chapter introduces data and annotation task, discussion of data pre-processing, annotation process, examining the agreement score between two annotators and addressing the ethics of studying mental health related data.

4.1 Source of data

Dataset used in this work originates from the "Depression" subreddit. It consists of posts and comments to them that were written between 01.01.2014 – 31.12.2017[8]. Post structure example is presented in Figure 4.1.

```
Thread ID:              %% 7Nov96
Thread ID:              %% 7Nov96
Thread ID:              %% 7Nov96
Thread title:           %% Being late for the dinner is such a shame
Reddit username:        %% randomnperson
Timestamp:              %% 1111123234
User's karma score:     %% 3

Post text:              "We were late to the dinner party and caused everyone to have a terrible time.
                        If I had only pushed my husband to leave on time, this wouldn't have happened."

Nest ID:                %% da1ceekl
Comment ID:             %% C4_7nov96
Thread ID:              %% 7Nov96
Thread title:           %% Being late for the dinner is such a shame
Reddit username:        %% AnnaDovlatova
Timestamp:              %% 1111129452
User's karma score:     %% 1

Comment text:           "No way! Dnt you dare to take all responsibility for this slow snail on you, Queen!"
```

Figure 4.1 Simulated example of post structure from a dataset[9]

For each post, the raw text file contained information about whether the post started a thread or whether it is a comment to a post, thread title, the Reddit username of the author and the timestamp, and the post itself and karma score.

It was decided to extract only posts that started a thread as they might contain the most valuable information that describes a personal situation and problems. Comments most likely are either supporting or asking to follow up questions. All the empty posts, duplicates, and other unnecessary fields (e.g. karma score) were filtered out using regular expressions, resulting in 172141 posts that started a thread. Source code can be found at GitHub[10]

4.2 Dataset

Exploratory analysis of posts length (Figure 4.2) and quartiles calculations revealed that for 172142 texts mean length is 1101 characters, 50% of texts have between 305 and 1381 characters, and tail continues up 39865 characters. A small sample of texts within the 305 and 1381 range was manually examined.

[8] This dataset was aquired from Kairit Sirts

[9] The example of the text for Figure 1 was taken from the „15 Common cognitive distortions" blog post. Link: https://psychcentral.com/lib/15-common-cognitive-distortions/

[10]https://github.com/SochynskyiStas/Master-s-Thesis-code/tree/master/Data%20preprocessing%20and%20agreement%20score

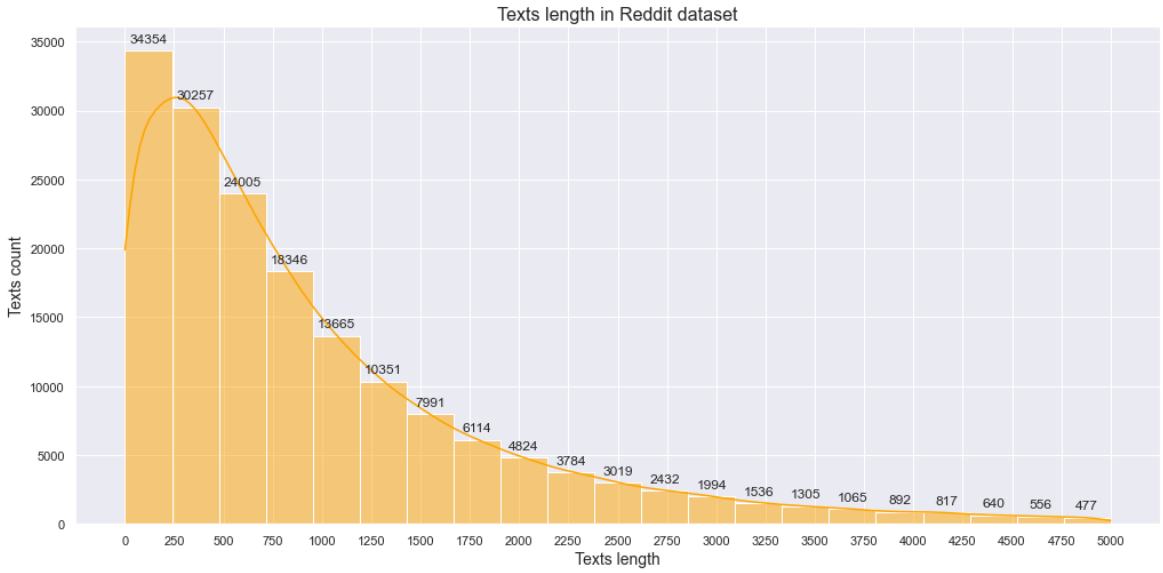

Figure 4.2 Texts length distribution chart

Texts below 200 characters might not contain sufficient information to decide about the presence or absence of certain distortions. For example, manually observed Reddit posts mostly contained swearing words or too few sentences to understand the contexts of the problem. On the other side, texts over 1500 contain a large amount of information and might include multiple cognitive distortions, potentially confusing an annotator on what label should be assigned. Also, longer texts require more time to be read and analysed. It was decided to filter out posts with size in a range starting from 200 characters up to 1500 characters so that annotation time of one publication would not exceed 120 seconds. It estimated that performance of an annotator would be 30 posts per hour. It also expected that texts of resulted dataset would not contain more than one cognitive distortion.

Using a simple comparison statement, the number of texts were reduced to 104491. Twenty thousand posts were put in lowercase and formed dataset (1) for further work. Source code can be found at GitHub.

4.3 Data annotation

4.3.1 Label set

The initial label set for annotation guideline consisted of the top ten frequent cognitive distortions[11] and an extra label for not distorted posts (Table 4.1).

[11] https://www.healthline.com/health/cognitive-distortions

Table 4.1 List of labels

Cognitive distortion	Definition	Final cognitive distortion
Not distorted	Does not contain any cognitive distortion. **Example:** expresses feelings, asking questions about medication	Not distorted
Black and White thinking	A person only sees life events in extreme categories: either perfect or a total fail. **Example:** when a student does not all As in the semester and consider this as a fail.	Black and White thinking
Overgeneralisation	A person perceives a single (or a few) negative events as a never-ending pattern of defeat. **Example:** *"The interview went great, but they did not call me back. I'll never get a job!"*	Overgeneralisation
Mental filtering	A person picks out single negative detail and dwells on it exclusively so that their perception becomes distorted. **Example**: a person focuses on a negative comment to their work and ignores any positive or neutral comments.	Disqualifying positive
Disqualifying positive	A person rejects positive experiences by insisting they *"don't count"* for some reason or another, despite that it may contradict their everyday experiences. **Example:** *"I've only cut back from smoking 40 cigarettes per day to 10. It does not count because I've not fully given up yet."*	Disqualifying positive
Jumping to conclusions	A person makes a negative interpretation even though there are no definite facts to support their conclusion. Includes mind reading and fortune teller cognitive distortion. **Example:** *"I know she hates me and does not want to be friends with me."*	Jumping to conclusions (mind reading, fortune telling error)
Emotional reasoning	A person assumes that their negative emotions\thinking necessarily reflect the way things are. **Example:** *"I feel terrified about going on aeroplanes. It must be very dangerous to fly."*	Emotional reasoning

Should thinking	Persons attempts to motivate themselves with *"should"* and *"shouldn't"*. "Musts" and "oughts" are also issues. **Example:**" I should have got the painting done this weekend"	Should thinking
Catastrophizing	Person exaggerates the importance of situations OR person inappropriately shrinks until they appear tiny. **Example:** *"I ruined my presentation by mispronouncing a couple words"*	Catastrophizing (exaggeration and minimisation)
Labelling	Instead of describing an error, a person prefers to attach a negative label to oneself or people. **Example:** *"I'm a loser "*	Labelling
Personalisation	A person is considered to be the cause of a negative external event (that already happened) for which you were not primarily responsible. Another case is the opposite: a person blames others for negative situations that happened in their life. **Example:** *"My mom is always upset. She would be fine if I did more to help her."*	Personalisation

After a brief analysis of definitions, it was decided to join *"Mental filtering"* and *"Disqualifying the positive"* in one category based on their similarity: In both cases, a person ignores positive or neutral events, but with a different justification. By definition, a few cognitive distortions consist of multiple cases of distorted thinking. For example, *"Jumping to a conclusion"* contains 2 cases: *mind reading, fortune telling error*. "*Personalisation*", *"Labelling"* each contains 2 cases: one is when a person either have distorted thinking about themselves *(e.g. "I'm dumb ")* or about others *("He's dumb")*. "*Catastrophizing*" category also contains 2 cases: in one case person tend to make an elephant out of a fly (e.g. *"I could not reply to a few questions "*out of 100). In the opposite case, person minimise their problems or the consequences of their inaction (e.g. *"I have plenty of days to start and finish my master's thesis ")*. Having two use cases for *"Jumping to a conclusion"*, "*Personalisation*", *"Labelling"* and "*Catastrophizing*" labels can result in a larger number of observations in these categories compared to others. It might lead to an imbalanced dataset and overfit the classification models for one or another class. This should be considered during results interpretation of the final model.

Annotation guideline

The final label set is reflected in the annotation guideline (see Annotation guideline) used to annotate the dataset for detection and classification tasks of this work. It includes ten labels: 9 for cognitive distortions and one for not distorted thinking. Each of the cognitive distortion descriptions contains an example of distorted thinking and questions that help validate cognitive distortion.

4.3.2 Annotation tool

For annotation purposes, it was decided to use to open-source annotation solution Label studio[12]. Right out of the box, this tool includes pre-setup templates to work with a variety of documents: texts, Images, Audios and tasks such as classification, tagging, or object detection. LabelStudio allows adjusting configurations[13] (Figure 4.3) for the specific task person. It also stores all the label data on the computer that enables better control over the storage of the dataset.

After configuration set up of annotation interfaces, LabelStudio requests to import data in multiple formats: .json, .csv, .tsv, and archives consisting of those. Label studio convertor[14] takes internal Label Studio .json based format and export output as JSON, CSV, TSV.

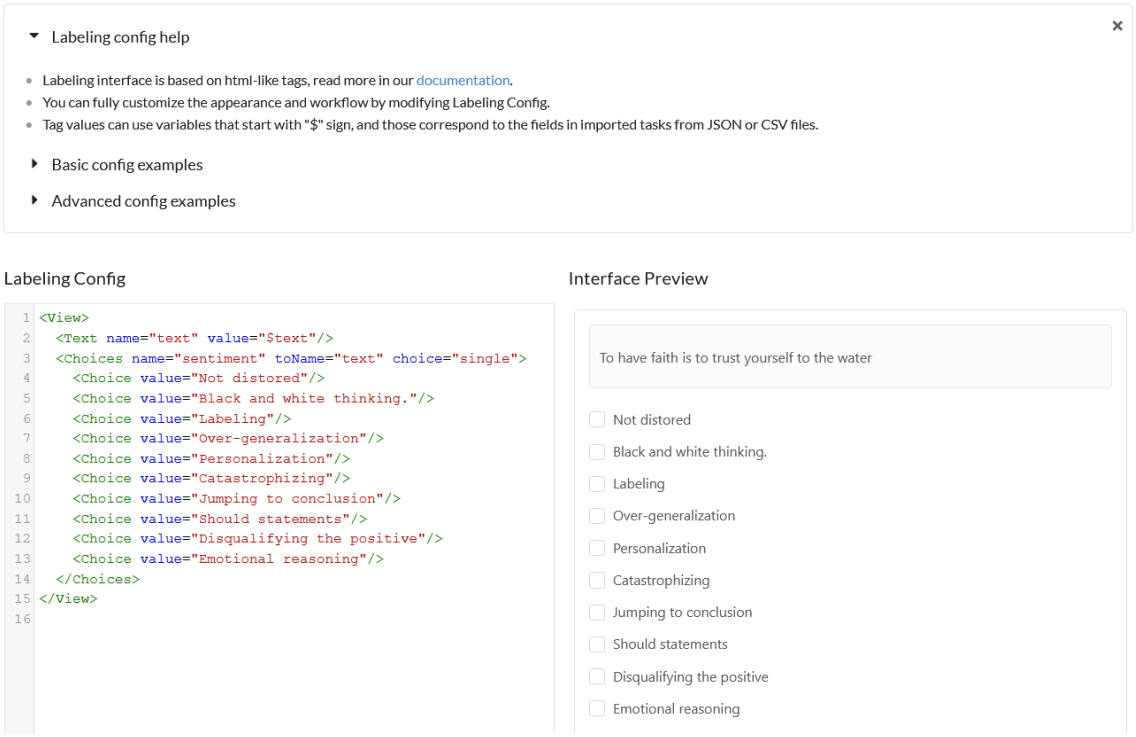

Figure 4.3 Label Studio configuration view

4.3.3 Annotation process

20000 (1) posts, ranged between 200 and 1500 characters, were randomly selected and imported to LabelStudio. LabelStudio presents posts in a random order to the annotator too. The annotation was completed by the author of this thesis, an informed non-professional individual.

There are multiple ways to annotate text data depending on the task: documents can be split into separate sentences and annotated sentences by sentence. A label can also be assigned to a whole text or only to selected words (named entity tagging). In the context of the real-world dataset, a single cognitive distortion can be expanded over a few sentences, and missing the context might result in the wrong annotation. For instance, there might be a sentence:

[12] https://labelstud.io/
[13] https://towardsdatascience.com/introducing-label-studio-a-swiss-army-knife-of-data-labeling-140c1be92881
[14] https://github.com/heartexlabs/label-studio-converter

"They hate me." According to the Annotation , it should be annotated as *"Jumping to conclusion"* as there are no facts to prove this point. However, preceding sentences of the original post can contain the story of conflict at work or parents violence against a person. In that case, *"They hate me."* conclusion is annotated as "Not distorted". For that reason, it was decided to annotate whole texts with a single label.

Initially, it was assumed that one post annotation from (1) dataset would not take more than 120 seconds. The actual annotation time for some post grew to up 300 seconds per post for a few reasons: mainly because assigning the most accurate label requires proofreading the definition and examples of different cognitive distortions multiple times. Also, transition time between texts in LabelStudio was not taken into account, taking around 5-10s. If there were difficulties in assigning a label to post within 5 minutes, it was skipped.

4.3.4 Annotation discussion

Data annotation is by no means a straightforward process, and there are multiple challenges: part of them related to the content that's been annotated, another to the annotation process itself.

In general, the topics of annotated posts can be grouped in the following categories: asking for help or advice, questioning about medication treatment impact, cheering up others by sharing personal challenges, and depressing posts. The last category of posts is within the scope of this work. However, not all of them contain distorted thinking patterns.

Here comes the first challenge: an annotator expected to differentiate distorted thinking and post where people express their feelings. Annotator must question a post, analyse the context, pay attention to words used and annotate using definitions, questions, and example from Annotation guideline. For example, a person might harm themselves to cope with life problems – this is probably some escaping-based coping mechanism, but not distorted thinking. It becomes cognitive distortion when a person mentions that this is the only way to feel loved and says that they will never be able to control desire for self-harm.

Another challenge comes with the large post (~over 900 characters). Texts were containing more than one phrase that could be labelled as different cognitive distortions. In such cases, it was decided to annotate with cognitive distortion that is the most expressed in the text. For example, the publication can describe a story about first love break up a person might jump to the conclusion that *"no one will love me again"*. However, post context allows to assign *"*Overgeneralisation*"* as it was the first time break up happened, but the person already thinks that there is some pattern that they will not defeat.

The next challenge was related to overlaping and interconnection of different cognitive distortions definitions. For example, a person may state: *'We broke up recently. He never truly loved me.'* A person sees this situation in a "Black & White thinking" manner. However, if a sentence contains *"We had a nice trip to Paris where we had romantic night walks. But we broke up recently. He never truly loved me",* it can be labelled as "Mental filtering": a person does not want to cherish all the positive moments and focuses only on a sad part. It also can be emotional reasoning as a recent break-up hurt them.

As with "Black & White thinking" and "Mental filtering", a similar case occurred with "Black and White thinking" and "Overgeneralisation" that sometimes posts included "Labeling" structures. For example, there were posts where students were afraid of losing scholarship because not all of their grades are As, and they labelled themselves as 'stupid' or 'failures'. On one side, students experience a "Black and White" thinking pattern: non A grades are equal to poor performance. On the other side, they explicitly mentioned the label. Unfortunately, the literature does not provide clear advice on how such cases should be

mitigated to stay consistent during an annotation. In our work, large in scale labels absorbed minor labels like "Labeling". Such cases bring inconsistency in the annotation process, cause an imbalance of label representation, and potentially worsen model performance.

In fewer cases, sentences interpretation also cause a delay in the annotation. For example, in one of the texts, a person wrote, *'this times cannot come back to me'*. One way of interpretation is straightforward: obviously, events from the past stay in the past and can not be part of the present. On the other side, a phrase could be interpreted as a person's overgeneralisation and conclusion that nothing good will ever happen to this person again.

It is also essential to come up with an appropriate label title. Label naming can cause annotator biasness and might bring psychological discomfort during annotation. For example, in this work, catastrophising contained two opposite cases: either exaggerating the importance of the event happen or shrinking its importance and impact. In a few cases, people described domestic violence, sexual abuse, which were psychologically difficult to annotate as "Catastrophising". For this label, it is recommended to split it into "Magnification" and "Minimisation".

Another important factor is the personality of the annotator. Coming back to the self-injury example. It is essential to suspend personal judgment about people's coping mechanism. During annotation of cognitive distortions, the point is not to estimate whether the selected coping mechanism good or bad. The goal is to understand how a person percept it.

In conclusion, we propose some action items on how the annotation process can be improved to mitigate discussed challenges:

- To overcome some definition closness, it might be reasonable to annotate cognitive distortions using span annotation approach. In that case, the whole document is available to provide context, and different sentences and parts can be labeledwhich potentially
- As for post length, posts should not be too short not to contain any meaningful information and not too long to contain several distortions. When the whole post is labelled, it is recommended to reduce its length to range between 300 and 800 characters, which may help avoid cases with multiple cognitive distortions and stay within the annotation time boundary.
- Annotation guidelines can be adjusted after the annotation of the first batch of posts to change the label set and enhance annotation guidelines with extra information. Additional collaboration with industry expert might be beneficial to clarify the difference between different cognitive distortions.

4.3.5 Statistics of the annotated dataset

In total, 2021 random Reddit texts from a dataset (1) were observed using Label Studio: 1187 posts were labelled as including distorted thinking 744 as not distorted based on the Annotation guideline. Out of 2021 posts, 90 were skipped as decision time for annotation went over 5 minutes. The distribution of posts among labels reported in Table 4.2.

Table 4.2 Annotated dataset statistics

Cognitive distortion	Number of posts
Black and White thinking	171
Overgeneralisation	135
Disqualifying positive	112
Jumping to conclusions	190
Emotional reasoning	179
Should thinking	84
Catastrophizing	98
Labelling	116
Personalisation	102
Distorted (TOTAL)	**1187**
Not distorted	**744**

The number of posts for each category varies from 84 to 190. "Jumping to conclusions" category holds the most of distorted observations – 190. As it was assumed in the discussion about the annotation guideline, two subcategories allowed more posts to be allocated to this category. However, it is not the same for catastrophising, labelling, personalisation: although these labels also have two subcases, they did not cause any imbalance as was expected. The next, the most popular label is "Emotional reasoning", with 179 observations. By definition, "Emotional reasoning" and "Jumping to conclusions" are quite close, and the only thing that differentiates them is that "Emotional reasoning" relies on feelings and nudges a person to make conclusions based on them. The least number of observations are allocated for the "Should thinking" *(84)* and "Catastrophizing" *(98)* labels. The reasons for a lower number of observations can also be related to the fact that they were expressed within large categories. For instance, "Overgeneralisation" or "Black and White thinking" was more expressed in a larger number of sentences. However, annotator's error also could have impacted the annotation results.

4.4 Annotation agreement score

This subchapter presents the analysis on annotators subjectivity and examines how different annotators labelled texts using the same Annotation guideline. In order to evaluate the agreement score, 300 posts were annotated by another informed non-professional individual.[15] Two cases results were compared: for the binary case (distorted, not distorted) and one for a multilabel case. Cohen's kappa coefficient is used to evaluate reliability of two annotators taking into account probability of chance agreement.

[15] Kairit Sirts, PhD

Kappa is calculated as:

$$k = \frac{p_0 - p_e}{1 - p_e}$$

where p_0 – relative observed agreement among raters, p_e – hypothetical probability of chance agreement.[16]

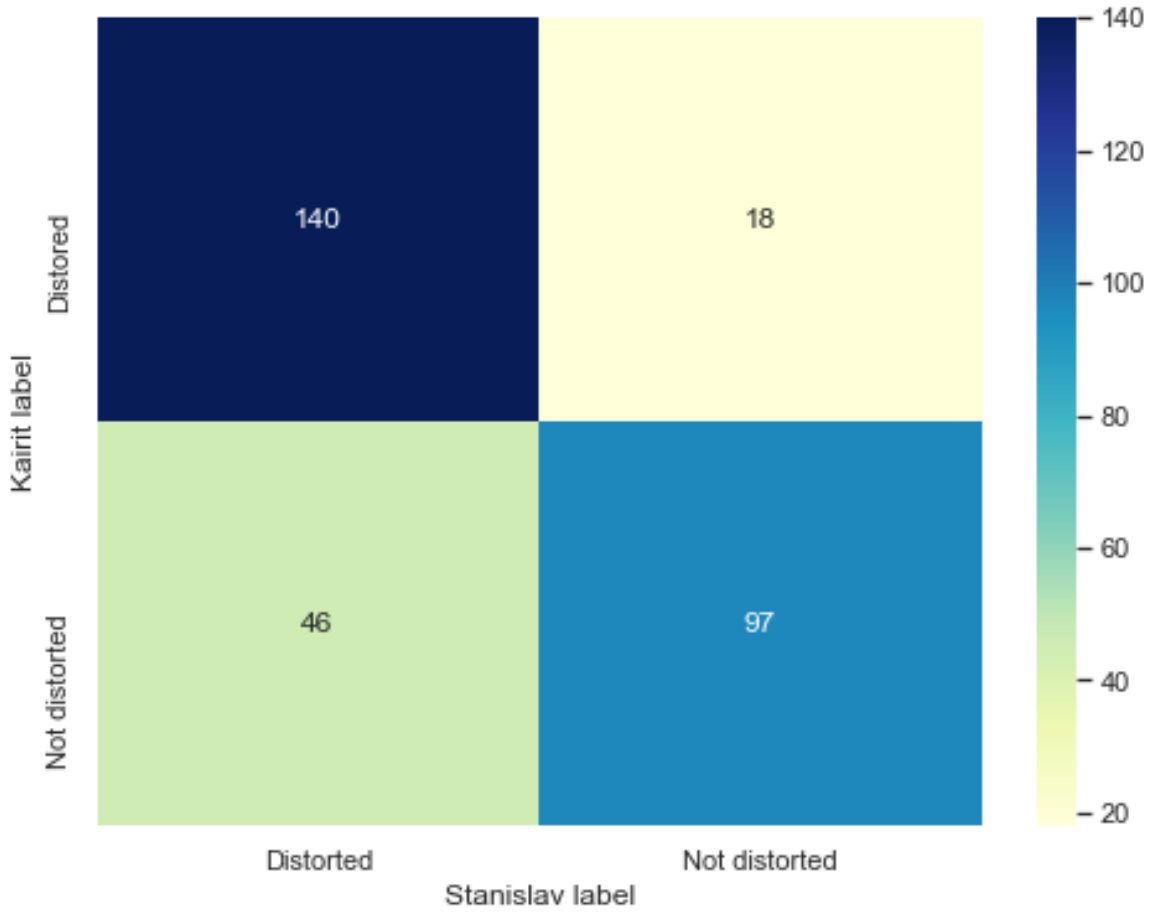

Figure 4.4 Confusion matrix for binary case

The agreement ratio for a binary case is 78.74%. Cohen's kappa value is 0.569, which falls within the moderate agreement and implies an average agreement between two annotators. Figure 4.4 presents a confusion matrix's visualisation and shows that the thesis author observes cognitive distortions more frequently while the second annotator is restrained.

As for multilabel case, the number of observed agreements equal to 167 (55.48% of the observations). Number of agreements expected by chance: 68.4 (22.73% of the observations). Eight additional labels resulted in a decrease Cohen's kappa value to 0.424 comparing to binary case. It still falls within moderate agreement and implies an average agreement between two annotators. For the binary case, it was already established that there is a high agreement for not distorted cases. The "Not distorted" class makes the confusion matrix less readable as its frequency overrides everything else. To report confusion matrix, "Not distorted" cases were excluded from confusion matrix reported in Figure 4.5.

[16] https://idostatistics.com/cohen-kappa-free-calculator/#risultati

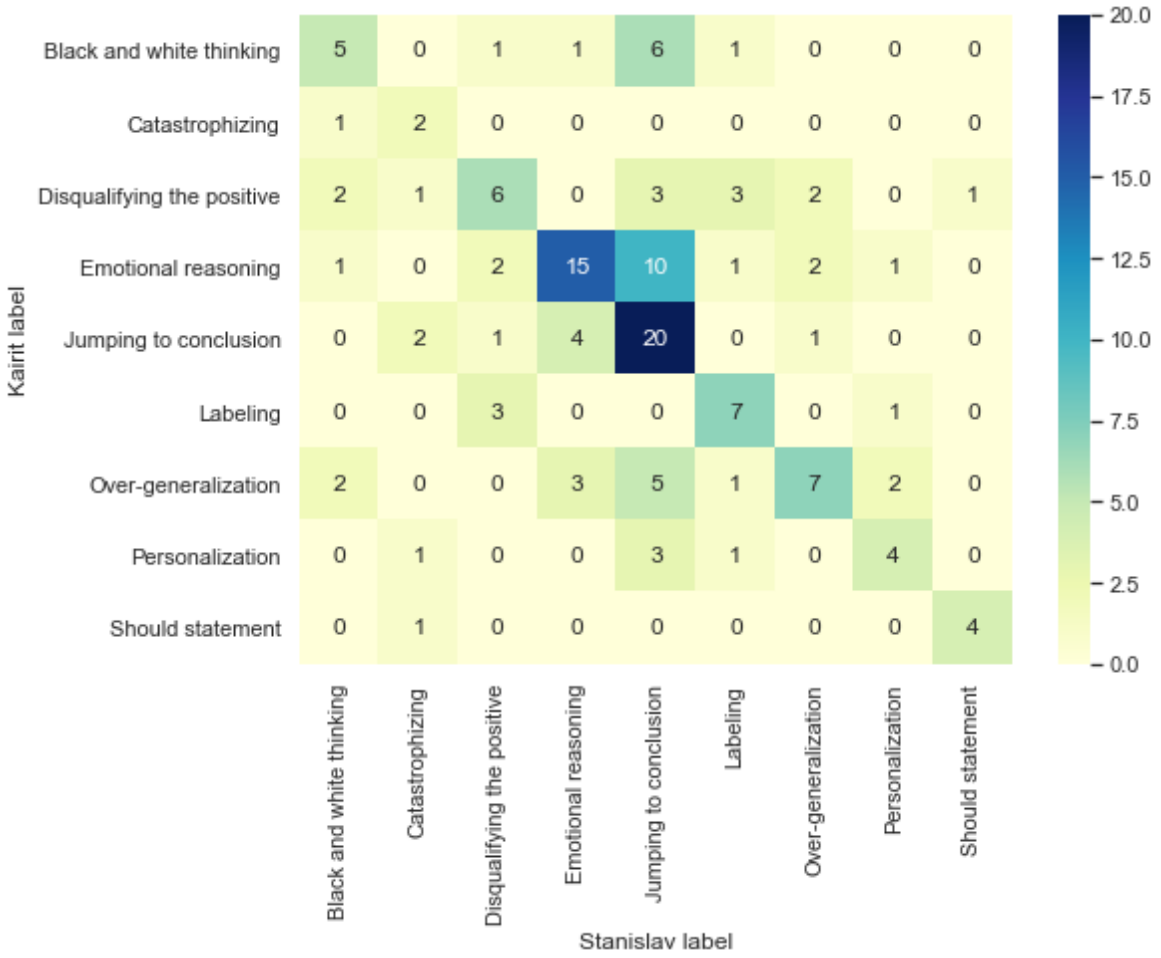

Figure 4.5 Confusion matrix for multilabel case (excluding not distorted labels)

The analysis revealed that the most confusing is "Emotional reasoning" and "Jumping to conclusion". It can be related to the fact that the definitions of these categories are quite similar and include making conclusions that are not based on facts. "Jumping to conclusions", "Black and white thinking", and "Overgeneralisation" also have overlapping observations and the reasons discussed in the Annotation discussion part.

4.5 Ethical consideration

Extracted post from Reddit contains personal and sensitive information. The obtained dataset includes thousands of individual posts. They reduce the risk of model over-fitting to any individuals writing style. Ideally, any research involving a similar data type should be done with their full knowledge and consent. Consequently, obtaining consent individually for each participant would require an impractical investment of time, and valuable data would have been lost if we required consent from each participant. In that case, the only means of contact would be via Reddit, where many participants are only active for a short period. Moreover, data was extracted from 2014 to 2017. Most of the account could have been deleted, closed or forgotten by their owners.

Part 4 of the Reddit User Agreement[17] ("Your content") explicitly highlights that publishing their content user grants transferable permission to Reddit, and their publications can be copied, modified, analysed in all media by third parties.

[17] https://www.redditinc.com/policies/user-agreement-october-15-2020

Despite granted consent, data still needs to be treated with care and respect because it involves sensitive subject matter. There are two groups of participants to whom this might cause harm the researchers who annotated the data (1) and the people who authored the content (2). For group (1), it was essential to have stable emotional wellbeing for the annotation work. Reading depressing posts might trigger depression in annotator or recall the unpleasant personal experience.

There are a few steps that were made to minimise potential harm for a group (2). Neither original nor processed dataset will be available online. As a second step, none of the posts was used for example purposes in a thesis as the exact Google search request might find publication.

5 Unsupervised methods

This subchapter presents experiments and findings of unsupervised techniques applied to (1) dataset. It was hypothesised that it would be possible to identify signs of cognitive distortions in the real-world data and see whether any interdependency between cognitive will be detected.

We ran experiments using K-means, BIRCH clustering algorithms. For clustering, dataset (1) was vectorised using doc2vec vectorisation technique with little pre-processing: lowering texts, removing punctuation, stop words (from *nltk* library), expanding contractions, forming bigrams and 3-grams. During experiment, we also applied Elbow, Silhouette, and Davies-Bouldin methods to find an optimal number of clusters. The results were visualised using PCA and t-SNE methods and included in the respective section.

We also applied topic modelling using LDA, and followed the similar pre-processing as for clustering dataset. The results were visualised using pyLDA library and available on the GitHub (see Code) as .html file. Source code can be found at GitHub[18]

5.1 Clustering

Our initial hypothesis is that clustering algorithms will split Reddit posts from (1) dataset into 10 distinct clusters, with respect to a list of 10 labels for supervised task (see Label set 4.3.1). K-means and BIRCH were selected as these algorithms split data into a pre-defined number of clusters K = 10. As for the vectorisation techniques, it was decided to apply doc2vec distributed bag of words (DBoW) and distributed memory (DM) methods from *gensim* library. DBoW and DM doc2vec models were built with the following parameters: neural model capacity in terms of document representation was set to 200, the maximum distance between the current and predicted word within a sentence was 5, and a number of epochs were set 50.

5.1.1 Kmeans

Clustering results are visualised in Figure 5.1. It was obtained using dimensionality reduction methods PCA to decrease number dimensions to 2. Figure 5.3, Figure 5.4 were also produce using PCA and t-SNE. Different datapoints are coloured with respect to the cluster number they were allocated to.

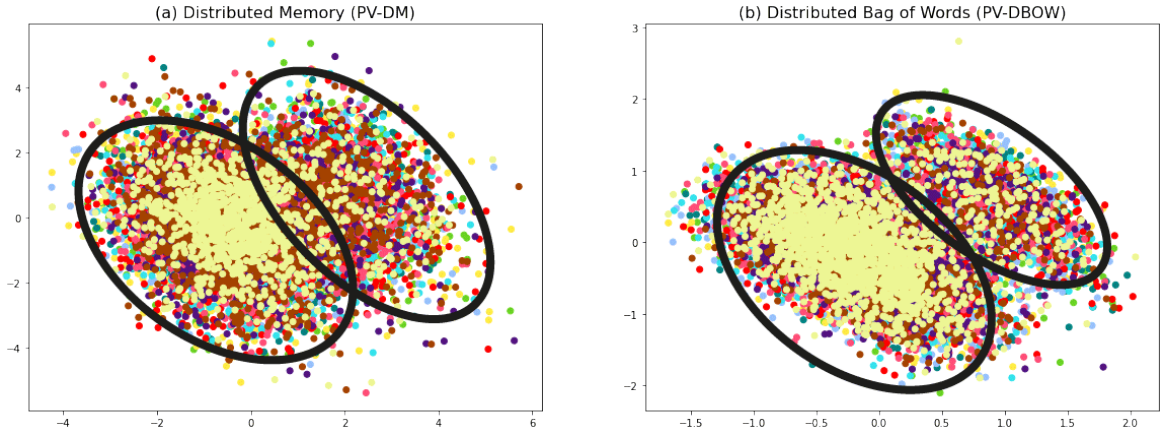

Figure 5.1 K-means clustering results visualisation for K = 10 using PCA

[18] Code: https://github.com/SochynskyiStas/Master-s-Thesis-code/tree/master/Clustering

The output clusters are interconnected and chaotic, and we decided to examine 100 posts manually. We did not find any explicit similarity between posts within one cluster. Indeed, in clustering algorithms, the underlying logic is not always possible to interpret. The data noisiness probably causes this case as data did go through ruthless preprocessing. Although in **Error! Reference source not found.**, all the 10 clusters are mixed up, it is still possible to distinguish two major clusters that are highlighted with dark circles. It was hypothesised that those clusters might contain positive and negative posts. We used Elbow method, Silhouette score, and the David-Bouldin index (Figure 5.2: (a.1, a.2),(b), (c) respectively) to estimate the optimal number of clusters and see whether it is possible to validate our hypothesis, and split a dataset into two interpretable clusters. Orange line represents DBOW doc2vec method, blue line – DM doc2vec method.

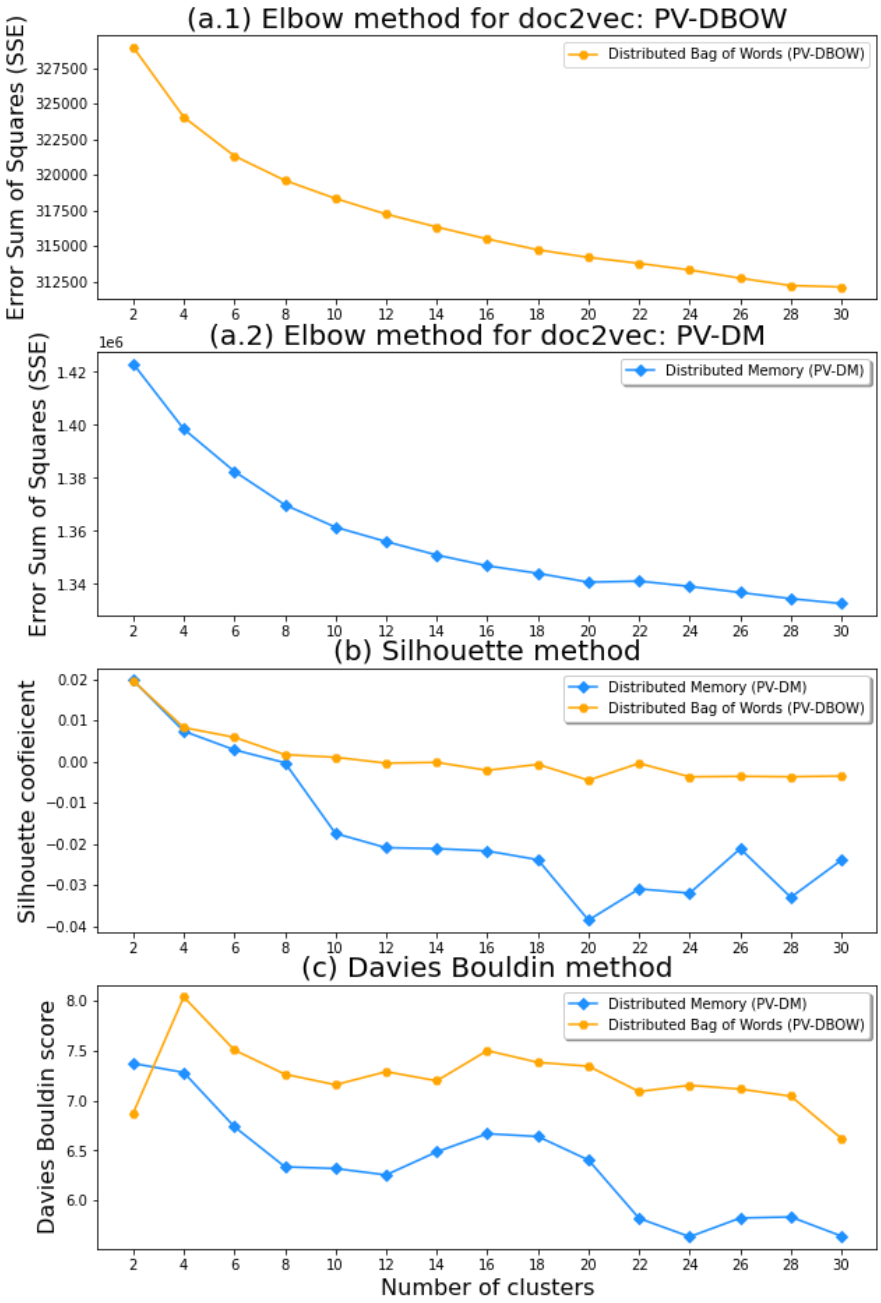

Figure 5.2 Optimal number of clusters for K-means

For both DM (a.1) and DBOW (a.2) approaches, Elbow method showed a constantly decreasing sum of squared error. There is no apparent optimal number of clusters. However, sum of squared error differs by a large extend for DM and DBOW. It means that clustering using DM method formed denser clusters, while clustering with DBOW results in a heavily sparse clusters. For DBOW and DM the Silhouette score starts from 0.2 for 2 clusters. The score decreases to 0 for 8 clusters. After that, DBOW slightly fluctuates, indicating overlapping nature of clusters that matches Figure 5.1. After K = 8, DM dropped to negative values range from -0.4 to -0.2 and concluded that using DM can worsen clustering performance. As for David-Bouldin index, values varies between 6 and 8 range. DM vectorisation method has a lower value and outperformed DBOW marginally. For DM vectorisation algorithm K-means with K between 8 and 12 and 22-30 shows the best clustering results. The best result for DBOW algorithm is only on the K=30; for K between 8 and 28, the results will be more or less the same.

Based on the positive values of the Silhouette score and the Davies-Bouldi index, it was decided to run a second round of experiment with K=2. Results are shown in Figures Figure 5.3, and Figure 5.4. Data points are coloured based on their presence in one of the clusters.

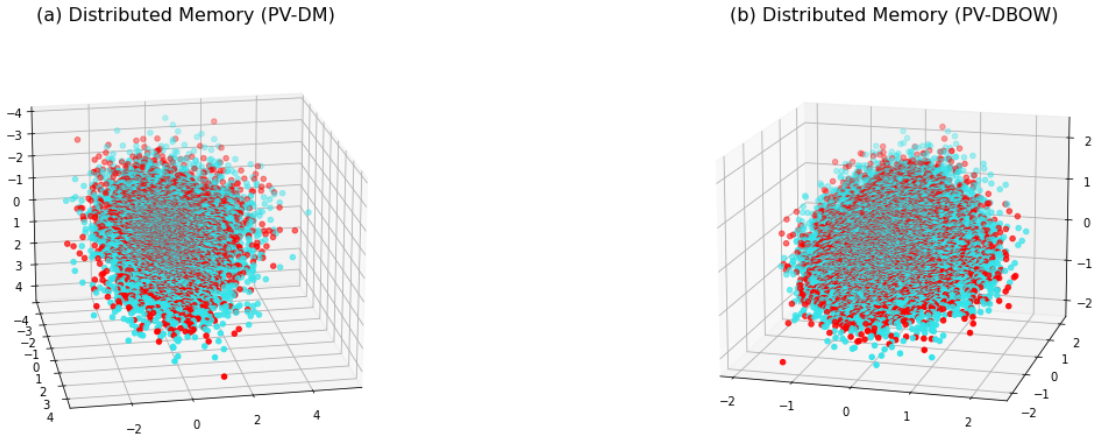

Figure 5.3 K-means clustering results visualisation for K = 2 using PCA

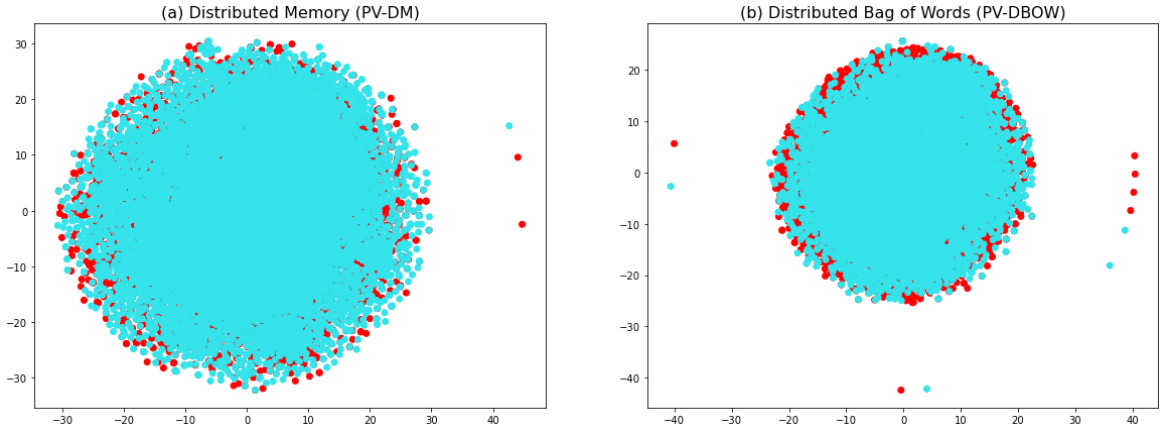

Figure 5.4 K-means clustering results visualisation for K = 10 using t-SNE

The figures presented above does not show two distinct clusters and rather repeats the results of Figure 5.1. Manual examination of 100 texts also did not reveal any explicit similarity or links between texts within one cluster and between distorted and not distorted thinking. Probably, data is very noisy and highly dimensional. K-means does not handle such situations very well. To conclude, with little data preprocessing, K-means cannot detect any signs of cognitive distortions in Reddit data.

5.1.2 Birch

Birch clustering results (Figure 5.5, Figure 5.7) are visualised using PCA. Data points are coloured concerning their cluster number.

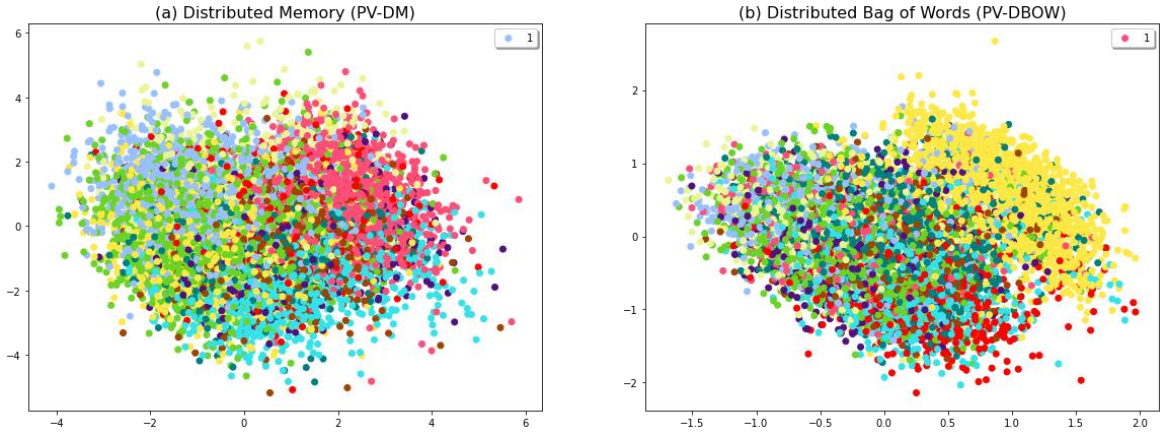

Figure 5.5 Birch clustering results visualisation for K = 10 using PCA

We start with a similar K-means experiment hypothesis about ten distinct labels in our dataset (1). Figure 5.5, both (a) and (b) show clustering results with Birch algorithm, and output clusters are still majorly overlapping. There are some signs of distinct clusters. For example, subplot (a) shows 4-5 highly scattered clusters, and subplot (b) represent two distinct clusters. There is no evidence for ten distinct clusters.

We decided to explore the optimal number of clusters with the Silhouette and David-Bouldin index. The results are visualised for K between 2 and 30 in Figure 5.6: (b) presents Silhouette score values, (c) - David-Bouldin index. Orange line - DBOW doc2vec method, blue line – DM doc2vec method.

The David-Bouldin index chart shows similar behaviour to Kmeans: DM vectorisation method shows lower index values than DBOW. The minimum value is achieved at K=2, and the difference in values is marginal: the lowest of DBOW is 7, and for DM is 8. Silhouette score (DM = 0.45; DBOW = 0.17 for K = 2) indicates that posts are well matched to their cluster and poorly matched to neighbouring clusters. It starts to decrease dramatically with the growth of clusters number repeating behaviour of Davis-Bouldin index. For the next step with Birch algorithm, the number of clusters was decreased to K = 2 and results are visualised in Figure 5.7 using PCA.

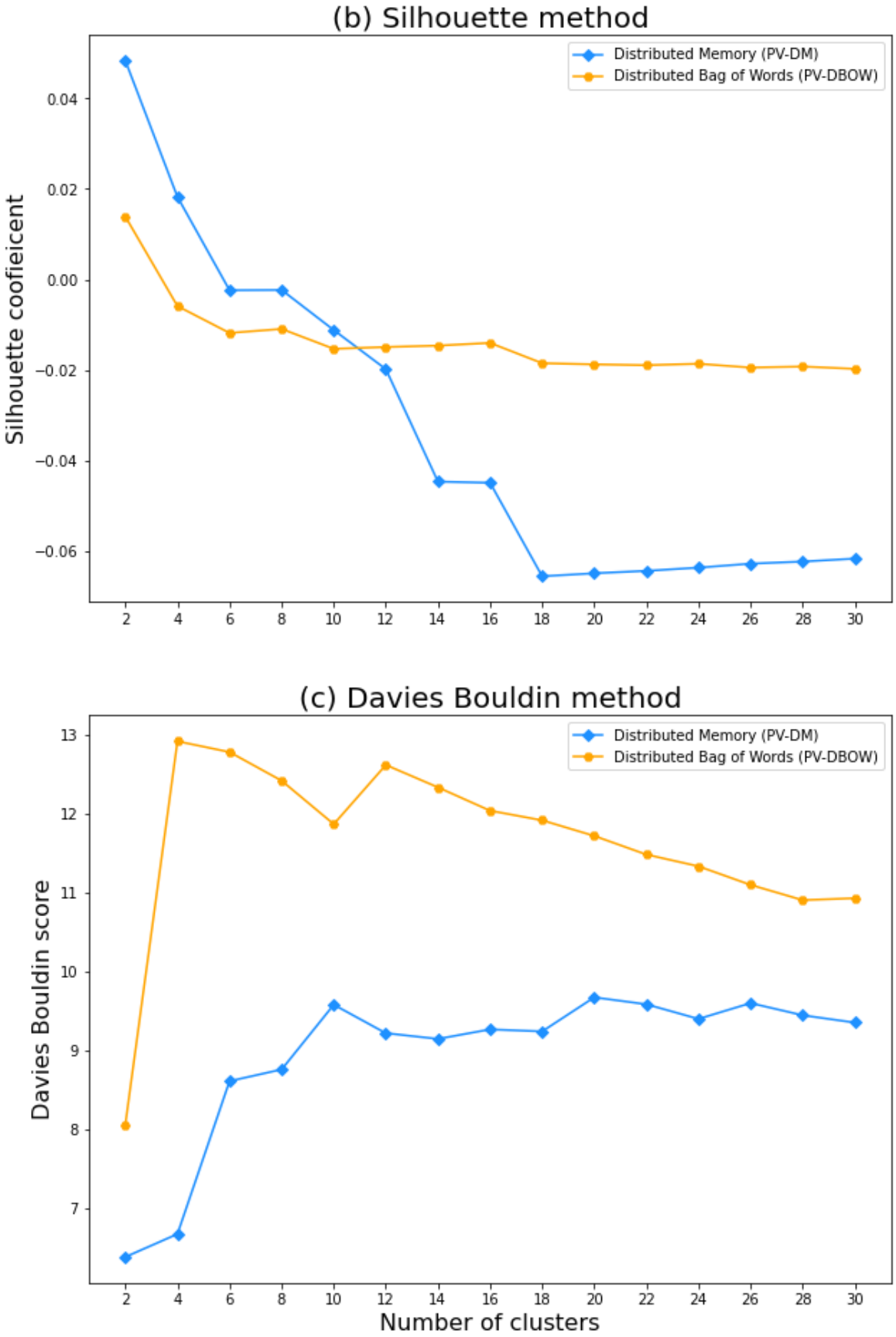

Figure 5.6 Optimal number of clusters for Brich algorithm

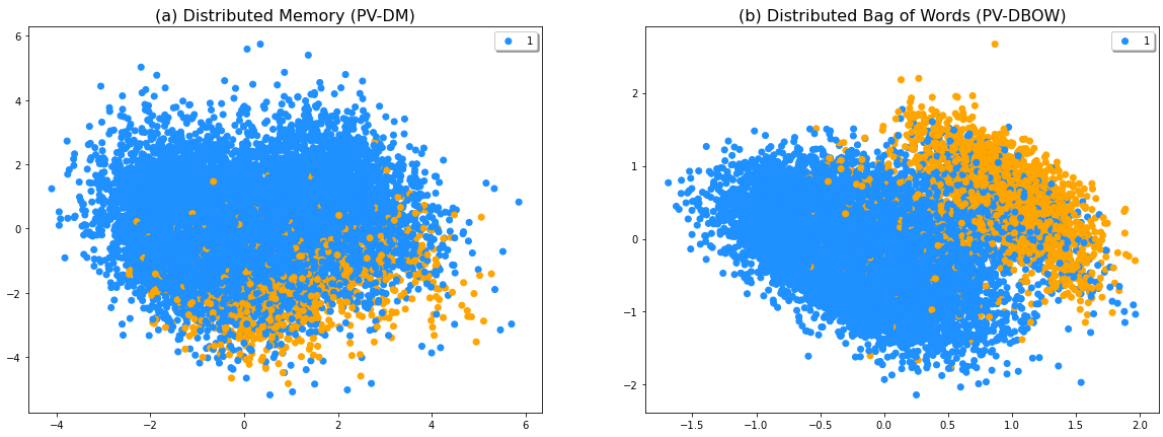

Figure 5.7 Birch clustering results visualisation for K = 2

Figure 5.7 shows that Birch algorithm with DBOW (b) vectorisation built more definite clusters compared to DM method (a). Manual examination of 100 posts from produced clusters did not show whether posts are connected to specific cognitive distortions.

Clustering experiment summary

To sum up, both K-means and Birch algorithms failed to detect cognitive distortions and did not manage to split Reddit posts into inpretebale distinct clusters. This potentially caused by data noisiness and little data preprocessing. It was decided to perform the LDA topic modelling to find support data for our hypothesis about the presense of cognitive distortions in the (1) dataset.

5.2 Topic modelling

In this experiment, we sought to find any indication of cognitive distortions as separate topic or as part of diffirent topics on the collection of Reddit posts (1). As well as, Shickel et al. (2020), we did not assume any number groupings of cognitive distortions and decided generate 25 topics using LDA from gensim library. The pre-processing included lowering texts, removing punctuation, stop words (nltk library stop word), expanding contractions, forming bigrams and 3-grams, excluding words with two and below characters.

The results were visualised using pyLDA library and presented in Figure 5.8 and can be found in the form of .html file on GitHub[19]. pyLDA visualisation includes Top 30 the most relevant terms for each topic. The area represents the topic prevalence: number 1 - the most popular topic, 10 – the least popular. Distance between two topics - represents the similarity of these topics. Lambda parameter can be adjusted to balance terms that are exclusive for this topic or choose more lament oriented to improve topic interpretation.

[19] https://github.com/SochynskyiStas/Master-s-Thesis-code/tree/master/Clustering

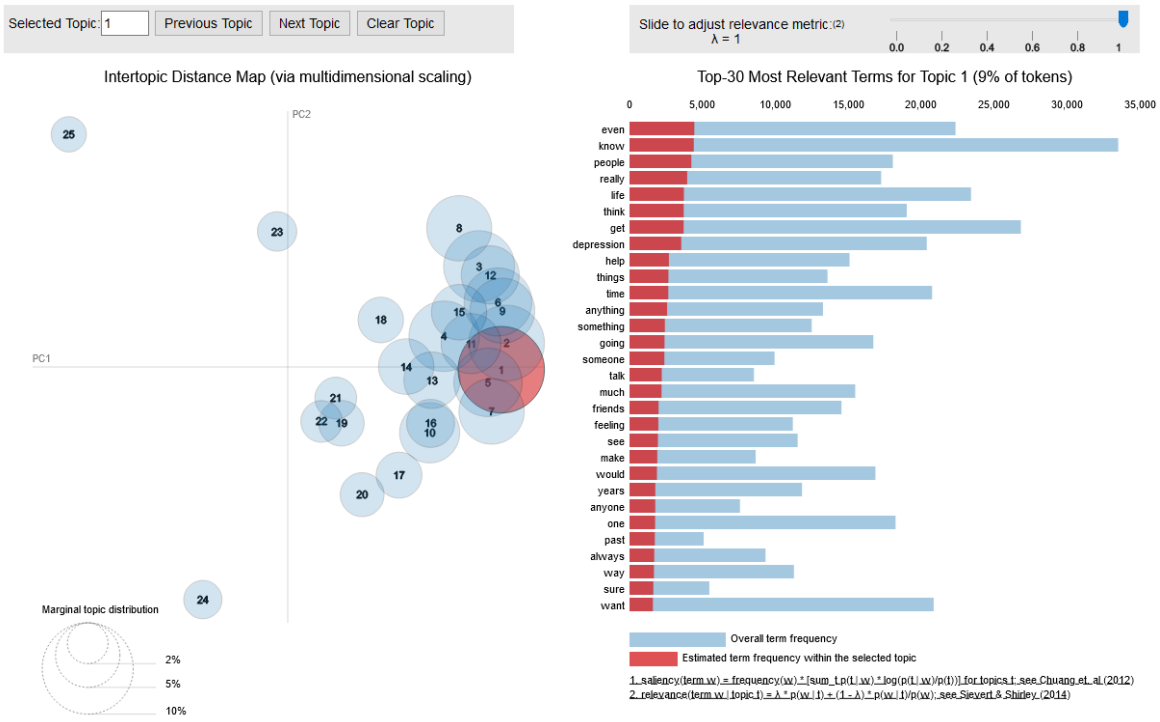

Figure 5.8 LDA visualisation results

In Figure 5.8, most of the topics are located in the middle-right part of graph. Some of these topics are related to medication (4, 20, 14, 25), work (22,8) or school (10, 3). Topics 1, 2, 5, 7,11, 9 are concentrated very close and related feelings about life and family and friends. The farthest topics are 23-25. Topic 23 is related to church thematic ("divorce", "best friends"), 25th topic seems to be related to search for help, and 24th can be summarised as containing phrases related to depression.

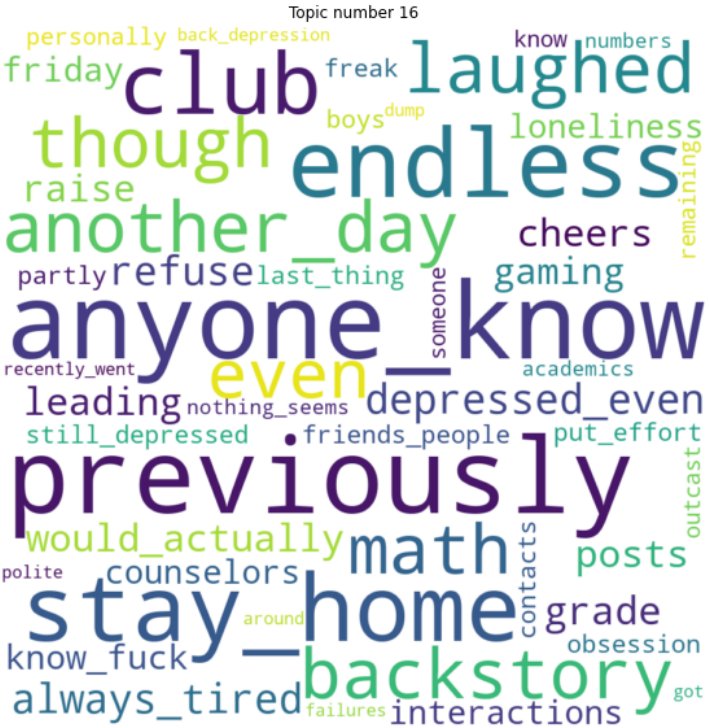

Figure 5.9 Word cloud for Topic 10

As for cognitive distortions, topic modelling captured phrases that indirectly hints at the presence of cognitive distortions in (1) dataset. For example, Topic 2, 7 and 5 holds phrases like “never”, “always”, “really, “know”, “nothing”. This are the most common words used in “Black and White thinking” distorted thinking. For example, 23rd and 13th topics contains "anyone could". This phrase in frequent for “Catastrophizing” label, minimisation subcase, when person tries undervalue their performance. Topic 10 and 9 contain "something always", "wish would" and "nothing_seems", and "endless", "hopelessness", "meaningless" – these phrases are expected to be seen emotional reasoning, black and white thinking, and overgeneralisation. Topic 23 has term "Unattractive" that can signal us about *"Labeling"* class. Phrase "care_anymore" from topic 16 and 9 can also can be treated as a sign of overgeneralisation label. Topic 23 also contain phrase "going_happen" which fits "Jumping to conclusion", fortune telling sub case.

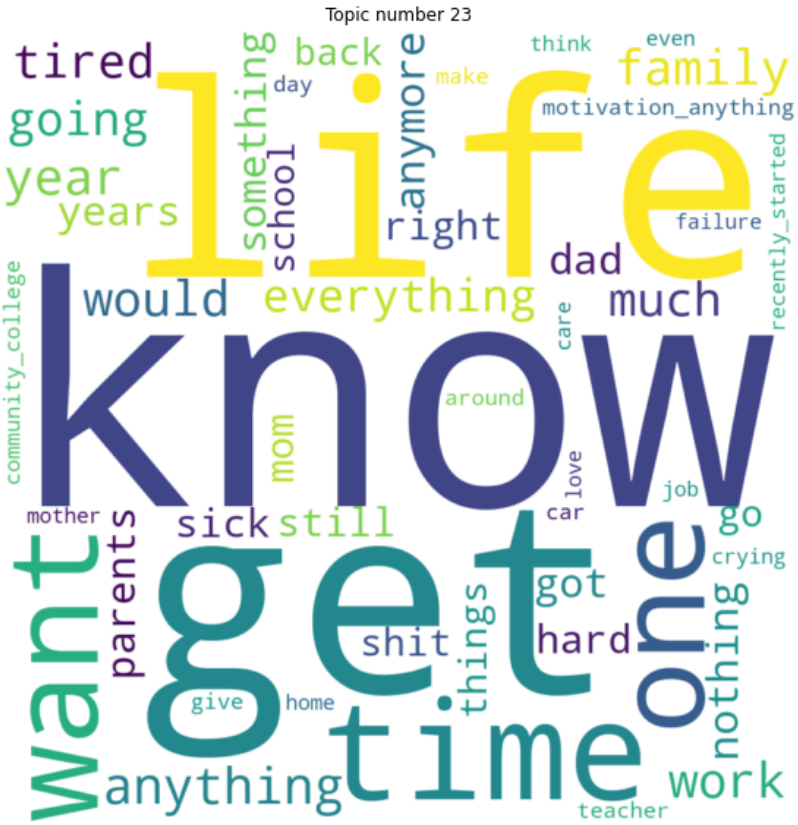

Figure 5.10 Word cloud for Topic 10

To conclude, despite data noisiness, LDA method showed results in forming topic related to depression and feelings. There are phrases that fall within a distorted thinking pattern, but it is difficult to draw distinct Potentially, by introducing lemmatisation and improving other parts of pre-processing it might be possible to achieve better results with LDA, and even identify the most common cognitive distortions for each domain: work, home, school, family relations etc.

6 Classification

For the classification task, a set of 1931 (excluding 90 skipped) posts were annotated. The dataset has an imbalanced distribution of target classes: 744 not distorted posts and 1187 posts distorted with assigned labels based on Annotation guideline. As tasks sought to set up a baseline for cognitive distortions detection and classification, the processing included minimum steps: the posts were lowered, stop words were removed using nltk library, and words with a frequency below three are excluded for BoW and tf-idf vectorisation methods. The code can be found on GitHub.[20]

6.1.1 Cognitive distortion detection

Detection of cognitive distortions is a binary classification task. All the distorted labels were replaced by label "1", and label "0" was assigned to not distorted. The data was split into a train (70%) and test (30%) dataset. To ensure that relative class frequencies are preserved in each subset (train and test), stratified sampling was applied.

In the first phase of the detection task, we applied *GridSearchCV* from the *sklearn* library to find the models' most optimal parameters. The param_grid parameter of *GridSearchCV* requires a list of parameters and the range of values for each parameter of the specified estimator. The list of estimators for logistic regression consisted of: penalty*(none, l1, l2, elasticnet)*, C value within 1000 and 0.00000000001 range and solver *(liblinear, saga, newton-cs, lbfgs)*. For SVM algorithm list contained the following estimators: kernel *(linear, poly, rbg, sigmoid)*, gamma and C value within the same range as logistics regression. Grid search analysis was not performed on fasttext algorithm.

To evaluate prediction on the test set, scoring was build based on weighted F-score metric. It was selected as it helps to find balance between Type 1 and Type 2 errors. In our context, Type 1 Error is bad because it might falsely tell a person that there is cognitive distortion. It might cause that person will start to think that something is wrong with them. At the same time, Type 2 error is the worst-case scenario, as the resulted outcome can depreciate person problem and worsen the condition. Grid search methods were run until two F-value and parameters in a row are equal, but up to 5 times. Classification algorithms were used from imported library *sklearn*. The results are presented in Table 6.1.

[20] https://github.com/SochynskyiStas/Master-s-Thesis-code/tree/master/Classification

Table 6.1 Hyperparameter tuning results

Methods	**Parameters**	**F-score**
Logistics regression BoW	C = 0.04, penalty = l2, solver = liblinear	0.7636
Logistics regression TF-IDF	C = 1, penalty = l2, solver = newton-cg	0.7734
Logistics regression doc2vec (DM)	C = 100, penalty = l2, solver = newton-cg	0.8227
SVM BoW	C = 1000, gamma = 0.0001, solver = rbf	0.7686
SVM TF-IDF	C = 1000, gamma = 0.0001, solver = rbf	0.7621
SVM doc2vec (DM)	C = 10, gamma = 0.00001, solver = linear	0.7934

Generally, F-scores values are very similar and fall within 0.76 – 0.83 range. Logistic regression using doc2vec (distributed memory) text vectorisation method holds the highest F-score of 0.82.

Table 6.2 presents trained model metrics for a testing dataset given hyperparameters value from Table 6.1. In case of binary classification, a micro average of accuracy is equal to recall and excluded from reported tables.

Table 6.3 Binary classification results table

Methods	Recall	Precision	F-score
Logistics regression BoW	0.7069	0.7008	0.6944
Logistics regression tf-idf	0.7069	0.7029	0.6896
Logistics regression doc2vec (DM)	0.6501	0.6419	0.6412
SVM BoW	0.6759	0.6659	0.6582
SVM tf-idf	0.7069	0.7060	0.6855
SVM doc2vec (DM)	0.6480	0.6394	0.6381
fasttext	0.711	0.7155	0.7119

According to Table 6.2, the weighted F-score varies within a narrow range between 0.6 and 0.71. The best performing method based on reported F-score, recall and precision is fasttext with weighted F-score = 0.7119. It is important to highlight that during optimisation of learning rate, it was observed that for values from 0.0 up to 0.4, recall, precision and f-score metrics were growing rapidly. In contrast, for values above 0.4, metrics stabilised and fluctuated almost at the same level. The optimal number of epoch was taken from the work on efficient text classification by Joilin et al. [67]. The second-best performing model is logistic

regression with BoW vectorisation method with F-score weighted = 0.6944. The results are differs marginally from tf-idf logistic regression (F-score = 0.6896).

Built models show 0.1 lower values comparing to hyperparameter tuning results. Especially, models with doc2vec vectorisation have contradictory results, as they were expected to show the best performance. The difference in results can signal about generalization issues, meaning that built models are less accurate in predicting labels for new input data. Another rationale for difference in the F-score with results in Table 6.1 is that grid search method can be biased as it creates subsamples and trains on 70% of data in each iteration. Moreover, the dataset is imbalanced, which also impact model quality.

Table 6.4 provides a granular look into per class metrics for two best-performing algorithms.

Table 6.4 Best performing algorithm metrics

Methods	Label	Recall	Precision	F-score
Logistics regression BoW *(C = 0.04, penalty = l2, solver = liblinear)*	Distorted	0.8515	0.6667	0.7815
	Not distorted	0.4753	0.7221	0.5550
fasttext *(lr = 0.69, epochs = 5)*	Distorted	0.8039	0.7513	0.7767
	Not distorted	0.5740	0.6465	0.6081

Generally, fasttext model performed well with a weighted F1 score of 0.71 across all posts from test dataset. Binary classifier showed solid results in identifying distorted text (precision = 0.75, recall = 0.8, F-score = 0.95), but its performance decreases for non-distorted posts (precision = 0.6465, recall = 0.5740, F-score = 0.6081). It can be explained by low sample representation of „Not distorted“ (38.52%) category in the dataset. Logistic regression results are more skewed for negative (precision = 0.7221, recall = 0.4753, F-score = 0.5550) and positive (precision = 0.6667, recall = 0.8515, F-score = 0.7815) labels. Comparing to fasttext results, Logistic regression F-score for „Distorted“ label is marginaly higher, while F-score for „Not distorted“ is 0.05 lower.

6.1.2 Multiclass classification

In the multiclass classification task, we aim to classify detected presence of cognitive distortions automatically. All the negative labels were excluded for multiclass classification, and 1187 annotated posts were used to form a dataset. Train and test subsets were generated on the same 70%/30% ratio with stratified sampling. The hyperparameters for multiclassification models for logistics regression and SVM are shown in Table 6.5.

Table 6.5 Multiclass classification hyperparameters values

Methods	**Parameters**	**F-score**
Logistic regression BoW	C = 0.01, penalty = l2, solver = liblinear	0.2399
Logistic regression TF-IDF	C = 1, penalty = l2, solver = newton-cg	0.2536
Logistic regression doc2vec (DM)	C = 1, penalty = l2, solver = liblinear	0.7534
SVM TF-IDF	C = 100, gamma = 0.01, solver = rbf	0.2458
SVM doc2vec (DM)	C = 100, gamma = 0.006, solver = rbf	0.6942

Models that are based on doc2vec method shows suspiciously impressive F-scores. Probably, not only the way grid algorithm works, but also data size and lack of label representation of different classes escalated problem and grid search reports skewed values. The results of cognitive distortion classification experiments are given in Table 6.6. Table includes weighted F-scores for six models that trained to classify nine distinct cognitive distortions.

Table 6.6 Multiclass classification results

Methods	Logistic regression			SVM		
Labels	BoW	TF-IDF	doc2vec (DM)	TF-IDF	doc2vec (DM)	fasttext
Black and White thinking	0.26	0.20	0.2198	0.2222	0.2151	0.2222
Overgeneralisation	0.15	0.15	0.1017	0.1111	0.0364	0
Disqualifying positive	0.24	0.21	0.1404	0.1785	0.0968	0
Jumping to conclusions	0.29	0.28	0.2342	0.2406	0.2000	0.2762
Emotional reasoning	0.29	0.19	0.2785	0.3582	0.2683	0.3464
Should thinking	0.33	0.25	0.4500	0.1621	0.4103	0.5238
Catastrophizing	0.22	0.14	0.1053	0.1702	0.1702	0
Labelling	0.25	0.26	0.2812	0.3333	0.3333	0.4127
Personalisation	0.25	0.30	0.2545	0.2448	0.3019	0.2540
Weighted average	0.23	0.22	0.2220	0.2332	0.2147	0.2274

Table 6.6. illustrates that built classification models are slightly better than random chance baseline (~11%). Generally, F-score for each method are similar, somewhere on the level of 0.21 to 0.23 with the best performance of SVM tf-idf algorithm (F-score weighted = 0.2332).

There are 3 labels that have the highest F-score metric: “Should thinking” (fasttext F-score = 0.5238), “Labelling” (SVM tf-idf F-score = 0.4127*)*, and “Emotional reasoning” *(*fasttext F-score = 0.4127*)*. “Personalisation” is also leaning to have higher F values comparing to the five left classes. “Should thinking” distortion was best detected with distributed based models (logistic regression doc2vec, fasttext, and SVM doc2vec). Logistic regression with BoW model showed the best performance for this class. Potentially, Should thinking, Labelling, and “Emotional reasoning” pattern are easy to classify probably due to the presence of unique features that are not present in other classes, e.g. “should” for should thinking, “feel” for “Emotional reasoning”, “fool, stupid” for “Labelling”. “Black and white thinking” and “Jumping to a conclusion” labels have average results concerning weighted F-score. This probably happened because both of these classes have the larger number of observations in a dataset and good sample representation.

However, there are three labels where almost all the models show lower than guess performance or just above random results: “Overgeneralisation”, “Catastrophizing”, “Disqualifying the positive”. There are a few rationales. First of all, as we annotated whole posts, overlapping problem of multiple distortions could have had a negative impact on classification model performance. Another reason is that besides distorted structures in posts, other sentences adds noise to train dataset. The third reason for poor classification of those labels is related to small dataset size and underrepresented distortions. The last potential issue is that the annotator made a mistake during annotation and wrongly annotated Reddit posts to the selected classes.

7 Summary

In this work, we aimed to broaden NLP techniques usage for mental health-related problems with a specific focus on identifying cognitive distortions based on the real-world posts from Reddit.

Experiments with unsupervised methods showed that they were quite unsuitable for cognitive distortion identification task. Although it was possible to spot some patterns with topic models, the signal in the raw unannotated data is too weak and not traceable to extract meaningful patterns. It seems supervised methods is the more promising approach in mental health tasks.

Supervised learning showed the best F-score of 0.71 for binary classification using a fasttext library with no text pre-processing. It is close to what was achieved in previous works on cognitive distortions detection: Shiekel et al. reported F-score = 0.88 [17] for automatic classification, and Morris et al. [16] reported 89% accuracy in a manual annotation. As for manual annotation, we achieved a 78.74% agreement between two annotators (kappa = 0.569), which might indicate the performance significance of the developed model. For multiclass classification task, the best F-score of 0.23 was reported for SVM using tf-idf vectorisation. SVM able to classify multiple cognitive distortions, but the difference is marginal compared to by chance probability. The automated cognitive distortion classification system requires a larger number of annotated data to achieve a better quality of predictions.

The quality of the annotated data also depends on the quality of the annotation guidelines. As a final contribution of this work, we developed an annotation guideline for manual annotation of cognitive distortions and applied it to annotate 2021 Reddit posts. We achieved moderate agreement for a binary case and multiclass case annotation (kappa=0.424). We also highlight the challenges of annotating mental health data for machine learning tasks. Shared annotation guideline can serve as a reference for further research on cognitive distortions in NLP (e.g., text generation, text-to-voice) and might be adapted to annotate other mental health issues potentially could be identified using text. Sharing annotation guidelines can also be a new approach in solving the problem of the lack of datasets in the mental health domain due to ethical or data privacy reasons.

Among future directions, we hope to focus on developing better annotation guidelines by collaborating with industry experts, and finding ways to speed up annotator performance so that it will be possible to generate large and high-quality datasets.

We are also interested in making the next step from baseline and apply state-of-the-art methods such as neural networks to build models that analyse incoming text and classify distorted thinking patterns with high-quality.

8 References

[1] James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, Abbastabar H, Abd-Allah F, Abdela J, Abdelalim A, Abdollahpour I. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018 Nov 10;392(10159):1789-858.

[2] Sobocki P, Jönsson B, Angst J, Rehnberg C. Cost of depression in Europe. Journal of Mental Health Policy and Economics. 2006 Jun.

[3] Hofmann SG, Asnaani A, Vonk IJ, Sawyer AT, Fang A. The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognitive therapy and research. 2012 Oct;36(5):427-40.

[4] Burns DD. The feeling good handbook, Rev. Plume/Penguin Books; 1999.

[5] Beck AT. Thinking and depression: I. Idiosyncratic content and cognitive distortions. Archives of general psychiatry. 1963 Oct 1;9(4):324-33.

[6] Ingram RE, Kendall PC. The cognitive side of anxiety. Cognitive therapy and research. 1987 Oct 1;11(5):523-36.

[7] Jager-Hyman S, Cunningham A, Wenzel A, Mattei S, Brown GK, Beck AT. Cognitive distortions and suicide attempts. Cognitive therapy and research. 2014 Aug 1;38(4):369-74.

[8] Casabianca S.S. Stuck in the Negatives? 15 Cognitive Distortions To Blame, 2021, https://psychcentral.com/lib/cognitive-distortions-negative-thinking (May 6, 2021)

[9] Levin C, Chisholm D. Cost-effectiveness and affordability of interventions, policies, and platforms for the prevention and treatment of mental, neurological, and substance use disorders. May 272016:219-236.

[10] Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering. 2017 Sep;23(5):649-85.

[11] Natural language processing: How this emerging tool can improve mental health treatment [Internet]. Advisory Board. 2019 [cited 2021May14]. Available from: https://www.advisory.com/blog/2019/05/nlp-mental

[12] Kramer AD, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences. 2014 Jun 17;111(24):8788-90

[13] De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. InProceedings of the 5th annual ACM web science conference 2013 May 2 (pp. 47-56).

[14] Homan C, Johar R, Liu T, Lytle M, Silenzio V, Alm CO. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. InProceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality 2014 Jun (pp. 107-117).

[15] Milne DN, Pink G, Hachey B, Calvo RA. Clpsych 2016 shared task: Triaging content in online peer-support forums. InProceedings of the third workshop on computational linguistics and clinical psychology 2016 Jun (pp. 118-127).

[16] Morris RR, Picard R. Crowdsourcing collective emotional intelligence. arXiv preprint arXiv:1204.3481. 2012 Apr 16.

[17] Shickel B, Siegel S, Heesacker M, Benton S, Rashidi P. Automatic Detection and Classification of Cognitive Distortions in Mental Health Text. In2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE) 2020 Oct 26 (pp. 275-280). IEEE.

[18] Pestian JP, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J, Cohen KB, Hurdle J, Brew C. Sentiment analysis of suicide notes: A shared task. Biomedical informatics insights. 2012 Jan;5:BII-S9042.

[19] De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting Depression via Social Media. ICWSM [Internet]. 2013 Jun.28;7(1). Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/14432

[20] O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interventions. 2015 May 1;2(2):183-8.

[21] Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 shared task: Depression and PTSD on Twitter. InProceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality 2015 (pp. 31-39).

[22] Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, Becker T. Feeling bad on Facebook: Depression disclosures by college students on a social networking site. Depression and anxiety. 2011 Jun;28(6):447-55.

[23] Kotikalapudi R, Chellappan S, Montgomery F, Wunsch D, Lutzen K. Associating depressive symptoms in college students with internet usage using real Internet data. IEEE Technology and Society Magazine. 2012 Dec;31(4):73-80.

[24] Perušic A, Kustura D, Matak I. Using linguistic metadata for early depression detection in social media. Text Analysis and Retrieval 2018 Course Project Reports. 2018:87.

[25] Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences. 2017 Dec 1;18:43-9.

[26] J. M. Grohol, “15 Common Cognitive Distortions,” 2018. [Online]. Available: https://psychcentral.com/lib/15-common-cognitive-distortions/

[27] Zhao12 X, Miao12 C, Xing Z. Identifying Cognitive Distortion by Convolutional Neural Network based Text Classification. International Journal of Information Technology. 2017;23(1).

[28] Mowery DL, Bryan C, Conway M. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using Twitter data. InProceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality 2015 (pp. 89-98).

[29] DSM-IV-TR. AP. Diagnostic and statistical manual of mental disorders. Washington, DC: American Psychiatric Association; 2000.

[30] Homan C, Johar R, Liu T, Lytle M, Silenzio V, Alm CO. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. InProceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality 2014 Jun (pp. 107-117).

[31] Chen S. Getting Started with Text Vectorization [Internet]. Medium. Towards Data Science; 2020 [cited 2021May14]. Available from: https://towardsdatascience.com/getting-started-with-text-vectorization-2f2efbec6685

[32] Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT press.

[33] Brownlee J., A Gentle Introduction to the Bag-of-Words Model, 2017, https://machinelearningmastery.com/gentle-introduction-bag-words-model/ (August 7, 2019)

[34] Stecanella B., What is TF-IDF?, 2019, https://monkeylearn.com/blog/what-is-tf-idf/ (May 11, 2019)

[35] Neskorozhenyi R., Word embeddings in 2020. Review with code examples, 2020, https://towardsdatascience.com/word-embeddings-in-2020-review-with-code-examples-11eb39a1ee6d (July 24, 2020)

[36] Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

[37] Shperber G., A gentle introduction to Doc2Vec, 2017, https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e (Jul 26, 2017)

[38] Budhiraja A. A simple explanation of document embeddings generated using Doc2Vec, 2018, https://medium.com/@amarbudhiraja/understanding-document-embeddings-of-doc2vec-bfe7237a26da (May 14, 2018)

[39] Dabbura A., K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks, 2018, https://towardsdatascience.com/k-means-clustering-

algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a (Sep 17, 2018)

[40] Scikit-learn Machine Learning in Python, K-means: Clustering, https://scikit-learn.org/stable/modules/clustering.html#k-means

[41] Maklin C., BIRCH Clustering Algorithm Example In Python, 2019, https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9 (July 1, 2019)

[42] Tian Zhang, Raghu Ramakrishnan, Maron Livny BIRCH: An efficient data clustering method for large databases. https://www.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf

[43] Tirthajyoti Sarkar, Clustering metrics better than the elbow-method, 2019, https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6 (Sep 7, 2019)

[44] Drakos G., Silhouette Analysis vs Elbow Method vs Davies-Bouldin Index: Selecting the optimal number of clusters for KMeans clustering, 2020, https://gdcoder.com/silhouette-analysis-vs-elbow-method-vs-davies-bouldin-index-selecting-the-optimal-number-of-clusters-for-kmeans-clustering/ (March 4, 2020)

[45] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

[46] Ajitesh Kumar, KMeans Silhouette Score Explained With Python Example, 2020, https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam (Sep 17, 2020)

[47] Halkidi, Maria; Batistakis, Yannis; Vazirgiannis, Michalis (2001). "On Clustering Validation Techniques" Journal of Intelligent Information Systems, 17(2-3), 107-145. doi:10.1023/A:1012801612483.

[48] Scikit-learn Machine Learning in Python, Davies-Bouldin Index: Clustering, https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index

[49] Namratesh Shrivastav, PCA vs t-SNE: which one should you use for visualization, 2019, https://medium.com/analytics-vidhya/pca-vs-t-sne-17bcd882bf3d (Dec 28, 2019)

[50] Ramakrishnan Thiyagu, Everything About t-SNE, 2020, https://medium.com/swlh/everything-about-t-sne-dde964f0a8c1 (Dec 2, 2020)

[51] Geoffrey Hinton, Laurens van der Maaten, Visualizing Data using t-SNE, 2008, https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[52] t-SNE Corpus Visualization: Text Modeling Visualizers, https://www.scikit-yb.org/en/latest/api/text/tsne.html

[53] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. the Journal of machine Learning research. 2003 Mar 1;3:993-1022.

[54] Selva Prabhakaran, Topic Modeling with Gensim (Python), https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[55] Buenaño-Fernandez D, González M, Gil D, Luján-Mora S. Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. IEEE Access. 2020 Feb 19;8:35318-30.

[56] Wright RE. Logistic regression. 1995

[57] Logistic Regression: Classification, https://saedsayad.com/logistic_regression.htm

[58] OpenCV, Introduction to Support Vector Machines: Machine Learning (ml module), https://docs.opencv.org/3.4/d1/d73/tutorial_introduction_to_svm.html (May 14, 2021)

[59] SVM in R for Data Classification using e1071 Package: Tutorials, https://techvidvan.com/tutorials/svm-in-r/

[60] Support Vector Machine: Chemical engineering, 2017, https://www.sciencedirect.com/topics/chemical-engineering/support-vector-machine

[61] fastText. Facebook; [cited 2021May14]. Available from: https://fasttext.cc/

[62] Gensim: topic modelling for humans [Internet]. FastText Model - gensim. 2021 [cited 2021May14]. Available from: https://radimrehurek.com/gensim/auto_examples/tutorials/run_fasttext.html#sphx-glr-auto-examples-tutorials-run-fasttext-py

[63] Jeremy Jordan. Hyperparameter tuning for machine learning models. [Internet]. Jeremy Jordan. Jeremy Jordan; 2018 [cited 2021May14]. Available from: https://www.jeremyjordan.me/hyperparameter-tuning/

[64] Malik F. What Is Grid Search? [Internet]. Medium. FinTechExplained; 2020 [cited 2021May14]. Available from: https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a

[65] Markham K. Simple guide to confusion matrix terminology [Internet]. Data School. Data School; 2020 [cited 2021May14]. Available from: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[66] Ghoneim S. Accuracy, Recall, Precision, F-Score Specificity, which to optimize on? [Internet]. Medium. Towards Data Science; 2019 [cited 2021May14]. Available from: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

[67] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759. 2016 Jul 6.

Acknowledgements

Plutarch says: "The mind (of a student) is not a vessel to be filled but a fire to be kindled." I would like to start with huge gratefulness and appreciation to my supervisor Kairit Sirts. Thank you for creating an impulse to think independently and fuel an ardent desire to seek the truth with your expert guidance. You are the role model for me of a great, patient and wise mentor.

I also would not make this work if it were not for some very important people in my life. Beginning with my Father, Mother, brother and my other family. I have been encouraged, sustained, hosted on the couch, inspired by the greatest friends anyone could ever had: Rodion Kharabert, Kateryna Kryvenko, Katsiaryna and Kiryl Lashkevich, Tatyana Korotkova, Elena Pasiachnik and Kirill Savchenko, Karina Kulish. Thank you! I would also like to praise my newly formed friendship with many great people, and I hope it will bloom in the nearest future.

Finally yet importantly, I would like to thank one more important person in my life - Anna Dovlatova. Thank you for showing me there are no limits to a person's strengths and how one person alone overcome challenges after challenges can stay humane even in the darkest hours of life. Thank you for showing me that there is dignity and growth from waking up every day and giving your best honest effort in everything you do, even if you do not succeed. Thank you for teaching me that the worst thing one can do in their life is not even to try because of the fear to fail. You and your lessons will always be a part of my life.

I wish everyone to win a war that no one sees and conclude my thesis by quoting Randle Patrick "Mac" McMurphy, the main antagonist of One Flew Over the Cuckoo's Nest. In the bathroom scene, he attempted to **:***"Well, I tried, didn't I? God damn it. At least I did that."*

Appendices

I. Annotation guideline

The annotation guide was used to form a dataset of cognitive distortions for the detection and classification task of this master's thesis.

1. **NOT DISTORTED**

Description: A person is considered not having distorted thinking when describing the current situation, expressing their needs (e.g. person to talk), how they feel about their life without making conclusions.

Example: For instance, person might miss playing with their friends, being nostalgic, expressing how they feel.

2. **BLACK AND WHITE THINKING**

Description: A person sees life or events in terms of two mutually exclusive categories with no middle options in between. This "either-or" thinking habit may result in self-recrimination or anxiety. Also knows as All or Nothing, polarised thinking.

Examples from literature: *"I never do a good enough job on anything."*[21], *"I've blown my diet completely."*[22] when a person ate a chocolate bar, *"I did not finish writing my thesis today so it was a complete waste of time.", "They didn't show, they're completely unreliable"*[23]

Follow up questions for complex cases:

- Is it possible to identify what the polarities are?
- Did a person mention anything good while describing the current life situation?
- What type of outcomes did person mention? Are these outcomes extreme?

3. **OVERGENERALISATION**

Description: A person views a negative event as a never-ending pattern to defeat. For instance, a friend's inconsiderate response means there is no caring for you, even when there have been other examples of consideration.

Examples from literature: *"I felt awkward during my job interview. I am always so awkward."*[24], *"Every time I have a day off from work, it rains."*[25], *"Today I talked to my boss about promotion. She said because of some company budget limits, it is impossible to do it right now. I'll never get that promotion."*

Follow up questions for complex cases:

- Did person mention anything about frequency?
- Did person mention when the last time this pattern occur was?

4. MENTAL FILTER (MF) AND DISQUALIFYING THE POSITIVE (DP)

MF Description: A person picks out single negative detail and dwells on it exclusively so that their perception becomes distorted.

[21] https://www.therapistaid.com/worksheets/cognitive-distortions.pdf
[22] http://www.pacwrc.pitt.edu/curriculum/313_MngngImpctTrmtcStrssChldWlfrPrfssnl/hndts/HO15_ThnkngAbtThnkng.pdf
[23] https://www2.aston.ac.uk/staff-public/documents/hr/OD/Cognitive%20Distortions.pdf
[24] https://niamhlynchcounselling.com/2020/06/03/identifying-your-thinking-patterns/
[25] De Oliveira, I. R. (2014). *Trial-based cognitive therapy: a manual for clinicians*. Routledge.

Examples from literature: *"That doesn't count, anyone could have done it", "I've only cut back from smoking 40 cigarettes per day to 10. It does not count because I've not fully given up yet."*

DP Description: A person rejects positive experiences by insisting they *"don't count"* for some reason or another. Persons maintain a negative belief, although their daily experiences contradict that.
Examples from literature: *"My boss said he liked my presentation, but since he corrected a slide, I know he did not mean it.", "We had a great evening and dinner with Jack and Jill, but my chicken were overcooked. It spoiled the whole evening. I hate this restaurant now."*

Follow up questions for complex cases:

- Did person describe any other positive/negative/neutral side of the outcome?
- Did person repeat any negative details very often?
- Did person mention their positive experience?

5. JUMPING TO CONCLUSIONS

Description: A person makes a negative interpretation even though there are no definite facts to support their conclusion. Jumping to conclusions includes different variations:

- **MIND READING:** A person arbitrarily concludes that someone is reacting negatively to them without validating these conclusions.
  Examples from literature: *"I know she hates me and does not want to be friends with me.", "My co-worker in Philly never says hello to me at work, I think it's because she has a better job than me, and she thinks she's better than me."*
- **THE FORTUNE TELLER ERROR:** A person anticipates that things will turn out badly.
  Examples from literature: *"She would not go on a date with me. She probably thinks I'm ugly.", "Housing market is good, but I think I won't be able to sell my house and I'll be stuck here."*

Follow up questions for complex cases:

- What the conclusion is based on?
- Did person actually try to do something?
- Did person mention the outcome of their attempts?

6. EMOTIONAL REASONING

Description: Person assume that their negative emotions\thinking necessarily reflect the way things really are. Examples of emotional reasoning include feeling hopeless and concluding that a problem is impossible to solve, or feeling angry and concluding that another person is acting badly.
Examples from literature: *"I feel terrified about going on airplanes. It must be very dangerous to fly.", "I feel angry. This proves I'm being treated unfairly.", "I don't feel clean, even though I've washed my hands three times. Therefore, I should wash my hands again.", "I feel really bad for yelling at my partner, "I must be really selfish and inconsiderate.", "I feel unable to cope today; therefore, I will feel unable to cope tomorrow."*

Follow up questions for complex cases:

- Did person express her/his opinion or facts?
- Did person mention any other person noticing this?
- Did person see any other solution?

7. CATASTROPHIZING (MAGNIFICATION AND MINIMISATION)

Description: A person exaggerates the importance of things or single-time events (such as personal goof-up or someone else's achievement) or inappropriately shrinks their problems until they appear tiny (desirable things or another person's imperfections).

Examples from literature: *"Yes, I won an important award—but that still doesn't really mean I'm accomplished in my field.", "I ruined my presentation by mispronouncing a couple words", "I did not complete my master's thesis in time, but it's no big deal.", "I forgot that email! That means my boss won't trust me again, I won't get that raise and my wife will leave me."*

8. SHOULD STATEMENTS

Description: A person tries to motivate themselves with *"should"* and *"shouldn't"*. "Musts" and "ought" are also issues.

Examples from literature: *"I should always be friendly.", "I should be married by now. All of my peers are!", "I shouldn't have made so many mistakes, "They ought to have been more considerable of my feelings, they should know that would upset me."s*

Follow up questions for complex cases:

- Does a person say/mention that the current situation is bad?
- Does a person describe any other potential "good" situation?
- Words like "must", "ought", "should", "other way", "differently" or their forms are used
- (Ask when was the last time what she/he wants (not should/must/ought))

9. LABELING AND MISLABELING

Description: Instead of describing an error, a person attach a negative label to themselves: Slapping simplistic labels on things to explain them, rather than looking at the unique facets of the situation. "I'm a loser."

Examples from literature: *"I didn't stand up to my co-worker, I'm such a wimp", "What an idiot, he couldn't even see that coming!", "I'm a loser", "He's an SOB."*

Follow up questions for complex cases:

- Can person specifically describe the situation?
- Did person mention an exact reason/facts of failure?
- Does person repeat one word in a different form? *(e.g. loser, failure)*

10. PERSONALISATION

Description: A person sees themselves as the cause of some negative external event (that already happened) for which they were not primarily responsible, **OR** person blames others.

Examples from literature: *"My mom is always upset. She would be fine if I did more to help her.", "It was my fault because I didn't say no"*[26]*, "If only I hadn't said that, they wouldn't have an argument", "We were late to the dinner party and caused everyone to have a terrible time. If I had only pushed my husband to leave on time, this wouldn't have happened."* [27]

Follow up questions for complex cases:

- Did person mention anyone else in the text?
- Was there any other person involved?
- Did person mention any other person responsible for the situation? (e.g, "Loser", "failure")

[26] https://www.psychologytools.com/articles/unhelpful-thinking-styles-cognitive-distortions-in-cbt/
[27] https://psychcentral.com/lib/15-common-cognitive-distortions

II. Code

The source code for this thesis is available on GitHub by the link below: https://github.com/SochynskyiStas/Master-s-Thesis-code

III. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, Stanislav Sochynskyi

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, Automated cognitive distortion de-tection and classification of Reddit posts using machine learning supervised by Kairit Sirts, PhD.
2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.
3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.
4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Stanislav Sochynskyi

14/05/2021