UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Pavle Kerkez

Zaznavanje samomorilnih misli v spletnih objavah z uporabo velikih jezikovnih modelov

Suicidal ideation detection in online posts using Large Language Models

Magistrsko delo

Ljubljana, 2023

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Pavle Kerkez

Mentor: izr. prof. dr. Damjan Škulj

Co-mentor: prof. dr. Gregor Petrič

Zaznavanje samomorilnih misli v spletnih objavah z uporabo velikih jezikovnih modelov

Suicidal ideation detection in online posts using Large Language Models

Magistrsko delo

Ljubljana, 2023

Abstract

The increasing significance of online communication in contemporary society has underlined the need to understand and identify suicidal ideation within these online spaces. Online communities, especially those centered on mental health, frequently feature communications deeply interwoven with expressions of suicidal ideation. While detecting these expressions is important for research, it is also fundamental for proactive moderation and prevention strategies within these platforms. Traditional machine learning methodologies have shown promise in recognizing suicidal tendencies in textual data. However, the emergence of large language models (LLM's) like GPT-4, built on sophisticated deep learning architectures, offers potential for a deeper and more nuanced detection of subtle cues linked with suicidal ideation that are often mingled with other themes and difficult to isolate. The core focus of this research is to examine the capability of LLM's in detecting suicidal content in online content. The objectives include 1) embedding the texts and clustering them based on content similarity, and 2) fine-tuning the models to distinguish and categorize documents based on the presence of genuine suicidal ideation versus general mental health discussions. The results validate the efficacy of LLMs in both tasks, achieving successful clustering of posts based on their content similarities to generate class labels, as well as having high precision and recall in differentiating suicidal ideation from general mental health narratives.

Key words: suicide detection, machine learning, large language models, document embedding, clustering

Povzetek

Naraščajoč pomen spletnega komuniciranja v sodobni družbi je poudaril potrebo po razumevanju in prepoznavanju samomorilnih misli v teh spletnih prostorih. Spletne skupnosti, še posebej tiste, osredotočene na duševno zdravje, pogosto vključujejo komunikacije, ki so tesno prepletene z izrazi samomorilnih misli. Medtem ko je zaznavanje teh izrazov pomembno za raziskave, je tudi ključnega pomena za proaktivno moderiranje in preventivne strategije na teh platformah. Tradicionalne metode strojnega učenja so pokazale obetajoče rezultate pri prepoznavanju samomorilnih nagnjenj v besedilnih podatkih. Vendar pa pojav velikih jezikovnih modelov (LLM), kot je GPT-4, zasnovanih na sofisticiranih arhitekturah globokega učenja, ponuja potencial za globlje in bolj niansirano zaznavanje subtilnih namigov, povezanih s samomorilnimi mislimi, ki so pogosto prepleteni z drugimi temami in jih je težko ločiti. Osrednja tema te raziskave je preučiti sposobnost LLM pri zaznavanju samomorilne vsebine v spletnih vsebinah. Cilji vključujejo 1) vdelavo besedil in njihovo združevanje na podlagi podobnosti vsebine ter 2) fino nastavitev modelov za razlikovanje in kategorizacijo dokumentov glede na prisotnost pristnih samomorilnih misli nasproti splošnim razpravam o duševnem zdravju. Rezultati potrjujejo učinkovitost LLM v obeh nalogah, saj uspešno združujejo objave na podlagi njihove podobnosti vsebine, da ustvarijo oznake razredov, poleg tega pa imajo visoko natančnost in obnovitev pri razlikovanju samomorilnih misli od splošnih pripovedi o duševnem zdravju.

Ključne besede: zaznavanje samomora, strojno učenje, veliki jezikovni modeli, vložitve dokumentov, hierarhično združevanje.

Table of contents

Index of figures

1. Introduction

Perhaps it is noteworthy that detecting suicidal ideation in the context of online communities combines two relevant topics into one. Suicide is a major public health concern (World Health Organization, 2019; Nathan and Nathan, 2020; Feldhege et al., 2022, p. 975) that takes many lives every year troughout the world. At the same time, it is not a new discovery that online communication in various forms has taken unprecedented importance in the lives of many people today, and as time goes on it will almost certainly continue to grow in it's influence (Webster, 1995, pp. 8-32). Because of this, thinking about the ways in which people express suicidal thoughts, feelings and intentions in the context of online communication seems like a worthwile endeavor.

Online communities often serve as virtual places where online communication about various topics- including suicidal thoughts- can take place. Furthermore, online communities can also be focused on topics related to mental health and can serve as safe spaces for people with emotional and psychological troubles that they feel unsafe sharing in their offline environments (Feldhege et al, 2022, p. 975). As such, it is likely that these spaces may contain a lot of content which contains suicidal ideation.

This is relevant for both practical and research reasons. Having such content available for analysis offers a great learning opportunity in terms of being a new data source that helps researching the topic of suicidal ideation. It opens doors for innovative ways of researching suicidal ideation, particularily written expressions of it, where previously the available content to be researched would have been limited. Practically, it would be of use to moderators and others engaged in the governance of such communities to be alarmed if a person's posts indicate suicidal ideation, so that they may approach them early, which is in itself an important step in suicide prevention and could be, in some cases, a part of public health policy.

Statistical methods and technologies that are grouped under the term 'machine learning' could be of great help in this task. Research demonstrating that machine learning algorithms can be used for detecting suicidal ideation in texts already exists (Fatima et al, 2021; Rahaman et al, 2022). However, the new innovations made in areas of large language models (LLM's) that rely on transformers deep learning architecture have opened the door to new opportunities that were previously difficult to handle in this type of suicide research. LLM's, such as the recently popular (and controversial) GPT-4, are powerful pretrained transformer models that have the ability to „understand" text, recognizing context, structure and meaning of words and sentences (Kowsher et al, 2016).

This property of these models, already tested and used in various fields gives them an advantage compared to other machine learning algorithms when it comes to research problems that deal with very specific and non-obvious text features that are not so easy to discern (Vaswani et al, 2017, pp. 1-2).

To the best of my knowledge, the use of new LLM's in detecting suicidal ideation in texts has not been explored yet anywhere, and it is the underlying purpose of this paper to test novel ways of using them for detecting suicidal ideation in online posts. The goal of this thesis is to test the usability of LLM's for classification tasks, specifically in the domain of detecting suicidal ideation in the corpus of general mental health related posts, retreived from various mental health and suicide related subreddit communities. It is a task I expect to be difficult, due to the assumed similarity between discussions about mental health struggles and genuine suicidal ideation.

The basis of this assumption is the inherent connection between suicidality and mental health issues: suicidal ideation usually happens as the part of the suicidal process, which is usually characteristic for individuals in whom the development of a mental illness (such as depression) is already happening (Roškar and Paska, 2021, p. 28). With this in mind, it can be said that suicidal ideation often includes existing mental health struggles, but not the other way around: mental health struggles do not have to imply suicidal ideation. And yet, this probably implies, in my opinion, that any content that contains suicidal ideation may contain large amounts of content that is similar to other non-suicidal mental health related texts.

This possible similarity in content that is about suicidal ideation, and the one that is merely about mental health struggles, is the challenge where LLM's sensitivity to context may prove useful.

For the purpose of this study, GPT models available trough OpenAI's api have been used in Python programming language environment. The entire code was executed and is currently stored in my personal Google Collaboratory notebook, which can be publicly accessed[1].

[1] Full code used in this thesis:
https://colab.research.google.com/drive/1YvRwdWFFdbsSu6eD_sJB8ZWCmlQ75l4v?usp=sharing

2. Suicide and suicidal ideation

2.1 Suicide as a public health problem

Suicide is a worldwide problem that was observed by the World Health Organization since 1950's. Every year, approximately 800,000 people die by suicide, which is estimated to be 10.7 per 100,000 individuals, with variations across age groups and countries. On a global level, suicide is the second leading cause of premature mortality in individuals aged 15 to 29 years, and number three in the age group of 15-44 years. It is also noteworthy that that 78% of suicides were completed in low and middle-income countries (Bachmann, 2018, pp. 1-2).

It is easy to think of suicide as a personal issue that is purely caused by factors that are within the individual- or, at best, his most immediate surroundings such as family. In contrast to this, both sociological theory and empirical findings support a different interpretation.

From a sociological perspective, suicide was one of the earliest subjects to be researched and theorized about. Emile Durkheim, who is considered today to be one of the founders of sociology, published the study „Suicide" (Durkheim, 1897), in which he made the first attempt at a completely sociological approach to understanding the phenomenon of suicide. Durkheim approached suicide by considering wider social factors such as the great societal changes, including the sudden increase in population size and density, developement of cities and the number and speed of new means of communication- an approach that is inherently connected to his wider theoretical contributions to sociology in the study of changes from mechanical to organic solidarity (Roškar and Paska, 2021, p. 185).

Durkheim claimed that the changes from societies of mechanical solidarity, that are based on homogeneity of their members and the overarching collective consciousness, to societies of organic solidarity based on the differentiation of it's members and individual consciousness, faced these societies with intensive changes. Change from one form of solidarity into another is not simple, and in a very short period of time fundamental changes happened, causing the old ways of organizing the society and relations among people to dissapear, with the new society having no time to develop new modes of organizing. The result of this was a new society where a pathological state of ’anomy’ is present, characterized by lack of social integration and regulation. This means that individuals in this society are not connected enough with each other, or with society as such (Roškar and Paska,2021, p. 186). Since Durkhemian anomic suicide is based on the idea of intense societal change and deregulation being the fundamental social cause of suicide, it can be argued that such conditions may indeed apply in contemporary context as well (Khan et al, 2021).

What Durkheim did was provide the first truly sociological approach to examining suicide. Other authors after him, notably Halbwasch(Halbwasch, 1978[2], in Roškar and Paska, 2021, p. 192), have noted that he may have overemphasized the sociological aspects of suicidality and treated individual factors ae either negligible or as a mere expression of the social causes of suicidality, but never as causes themselves. Halbswasch, instead, thought that suicide has both social and individual causes; and, moreover, that they must be studied together, meaning that individual causes themselves have a sociological dimension (Halbwasch, 1978, in Roškar and Paska, 2021, p. 192).

Halbswasch framed his approach of understanding of individual causes for suicide as ’social distribution of misfortunes’. Contrary to both Durkheim’s view that individual factors of suicide are ephimereal and negligent expressions of social forces, and a purely individualistic view that suicide has nothing to do with society, Halbswasch observed that ’unfortunate events’ that are often causes for suicide are not a product of mere chance, but rather the product of the structure of society or social groups. With the increasing complexity of society and it’s associated more complex divisions of labor and individuation, opportunities for different ’accidents’ and conflicts increase, which, especially in conjuction with the weakening of social (religious, family, etc) ties and their support, multiply individual motives for suicide (Roškar and Paska, 2021, p. 192). In other words, society deprives people of support while creating and sustaining social positions and situations where ’accidents’ and ’misfortunes’ are more likely, depending on what social position an individual has.

Much of the contemporary research affirms this intertwined relationship between the social and individual, showing that the risk of suicide depends on both individual and

[2] Halbwachs M. (1978) *The causes of suicide*.

structural conditions of the person. In one such study that was conducted in Netherlands found that various demographic factors contribute to increased risk of suicide (Berkelmans et al, 2021). In another study in the United States that was focusing on youth (participants from ages 10 to 19) over a period of time, the results have found that sex, race/ethnicity, being a member of sexual or gender minority, all contributed to the likelihood of attempted suicide (Ruch and Bridge, 2022, pp. 4-6).

As noted by the authors of the first study, it is tricky to look at these factors as if they were completely independent, since some of them are heavily correlated with each other. For example, the education level and income level are surely in correlation with each other, and in some places ethnicity/race can also be correlated with education and income levels (Robertson et al, 2022). Even so, this still implies that wider social factors interact and influence the final likelihood of attempted suicide, affirming the idea that suicide is, indeed, the product of complex sociological and socio-psychological processes.

It is logical to frame suicide risk as a social problem and an issue of public health and intervention. If anything, looking at suicide risk trough moralistic or individualistic lens is a deception that masks the real process that leads to suicide and takes the responsibility away from institutions and society, and fundamentally sabotages in-depth understanding and prevention of a tragedy. For this reason, researching suicide and developing possible prevention measures is a first-class public health task.

2.2 The relevance of online communities for suicide research

The increase in the amounts, types and availability of data during the previous decade has left it’s mark on virtually almost all academic and professional fields, including public health (Paul and Dredzde, 2017, p. 2).

The popularity and wide acceptance of Big Data is hardly surprising, given the vast amount of benefits that it offers. While Big Data can’t really replace traditional methods of data collection, in comparison to traditional public health data collection, it is fast, cheap, covers a large population, and provides data on topics with little coverage from traditional sources. Importantly, in many cases, it allows for real-time monitoring of large-scale processes, a sharp contrast to traditional methods of monitoring such as surveys, that sometimes take months to finish (Paul and Dredzde, 2017, pp. 12-13).

With new data sources, such as the electronic health records (EHRs) in the United States, researchers had their first large database that could be used to perform advanced statistical analysis in order to answer questions and solve problems that are native to public health. EHRs, electronic records of each patient visit to the physician in the US, contain information about each patient-physician interaction that happens. With over a billion patient visits to a physician over a year, this constitutes a very large amount of data that can be used as a basis for such efforts (Paul and Dredzde, 2017, pp. 2-3).

However, limiting oneself to electronic health records would be underestimating the real availability of data for public health purposes. An already existing source of potentially useful data is the user generated content on the web, in the form of social media posts and discussions (Paul and Dredzde, 2017, p. 3). This social data- data created by users with the goal of sharing with other users- can be monitored, either manually or by using computational tools, in order to learn about opinions, behaviors, sentiments, etc, of the population that generates it (Paul and Dredzde, 2017, pp. 14-15).

Using online data of this sort to monitor and research mental health and- closely related- suicidal ideation is a case that has additional benefits. Mental health fundamentally affects behavior, which can be exibited in online behaviors (Paul and Dredzde, 2017, p. 73), so it can be assumed that textual content of posts made on the internet contains various cues and indicators of the person’s mental health status.

Furthermore, using online content to research and monitor processes related to mental health is useful for another good reason: it bypasses the social stigma and shame related to mental health that often discourages self-reporting (Paul and Dredzde, 2017, p. 73) or other forms of disclosure of mental states to others. Online communities are known for their desinhibition effect- the occurence where some people self-disclose or act out more frequently or intensely when online compared to when they are offline (Suler, 2004). With user-generated posts on the web, we may be able to extract information in a much more ’natural’ environment where users themselves decided to share information with others in a setting that is comfortable for them. Such useful data is most readily found in a variety of online communities, some of which specialize for health and mental health support of it’s users. In such communities, it is possible that members feel ’at home’ and freely share personal stories in a natural way, thus potentially generating vast amounts of textual content that could be used for analysis. A factor that could influence this is the reduced importance of status and authority (Suler, 2004, p. 324).

Connection between suicide, suicidal behavior and suicide prevention and online behavior are well documented: online platforms dedicated to suicide prevention can promote interactions among individuals with shared experiences and enhance knowledge about prevention initiatives, emergency helplines, and various educational and support tools (Luxton et al, 2012, p. 197). Suicidal individuals often reach out for help and seek social support trough social media (Paul and Dredzde, 2017, p. 73), and there are multiple online communities dedicated to suicide support. In this sense, online communities dedicated to mental health, crisis support and suicide prevention can potentially serve as potent data sources for exploring a topic that is otherwise difficult to research.

Before proceeding forth and commenting on the importance of these communities, it is necessary to more closely define what suicidal ideation is, and how it can be detected in written online text.

Gaining insight into suicide and its associated behaviors requires not just precise terminology but also an understanding of the relationships and potential progression among different types of suicidality. Suicidality is a broad concept that captures both cognitive and behavioral elements, addressing various facets and events associated with suicide. Within the cognitive realm, there are suicidal ideations (or thoughts), intentions, and detailed plans. Suicidal ideations can be any reflection about ending one's life, appearing in multiple forms. These can be passive, where someone might wish they no longer existed without taking action, or active, where the individual actively thinks about self-harm. The clarity of these thoughts can vary, from ambiguous notions about ending one's life to more explicit ideas pointing towards a clear intention or a structured plan. A defined intention to commit suicide is when someone has a concrete and specific idea about it. When this idea provides detailed answers about the timing, method, and location of the act, it's termed a suicidal plan. (Roškar and Paska, 2021, pp. 26-27; Perry et al, 2020, p. 2).

The next question that can be asked is how can suicidal ideation be detected in the content of an online post?

In a study that was focusing on psycholinguistic changes in communication among adolescents in a suicidal ideation online community during Covid-19 (Feldhege et al, 2022), the authors have discussed some of the ways suicidal ideation may express itself in written online content.

The authors of this study have relied on linguistic or lexical style analysis, which studies the language individuals use and has been instrumental in researching suicidal behaviors. This approach provides insights into an individual's psychological state, especially when they're not forthcoming about their emotions. Such analyses have been conducted on various sources, from suicide notes to tweets and artistic works (Feldhege et al, 2022, p. 976).

The specific choice of words or linguistic style is a key area in studying the relationship between language and suicidal ideation. Certain word groups, especially those expressing negative emotions or associated with death, serve as indicators of feelings like hopelessness, a known precursor to suicidal thoughts. Established tools, such as the Linguistic Inquiry and Word Count, help identify these linguistic markers (Feldhege et al, 2022, p. 976).

Another type of linguistic cue that is a marker of suicidal ideation in a text is the use of self-centric words. Linguistic patterns, such as increased use of self-centric words, serve as indicators of social disengagement: As individuals become more isolated, their language reflects a greater self-focus and reduced social references (Feldhege et al, 2022, pp. 976-977).

The existence of these cues and linguistic markers in text implies that written content does, indeed, reflect mental states of people, and that there is a recognizable pattern within texts containing suicidal ideation that can be used for detecting such content.

Moving back to online communities and their content, focusing on content generated in communities dedicated to suicide and crisis support could prove to be a naturally formed pool of data with language and communication that are indicative of suicidal ideation and, as such, it could be used for learning more about these aspects of suicidal ideation, which can then be used for suicidal ideation screening and prediction.

Another potential benefit of having such pools of data is the problem of labeling the data- a process necessary for classification using machine learning about which there will be more word in the next section. Here, I would like to focus on the advantage gained by such natural grouping of texts that happen in online support communities such as ones related to suicide. In short, the existence of a community dedicated to a specific mental health problem, or- in our case- suicide, allows us to reasonably assume that the content of posts in such a community is representative of mental states, communication and behavior related to said problem- an assumption that was already confirmed by mental health experts that checked the content in another study (Choudhury et al, 2016). This may seem as a far-fetched statement, but assuming that the rules of the online community are such that they limit the conversations to specific content, and the rules are enforced by moderators, then this can serve as initial labeling of posts as relating to a specific type of content. While additional validation of data may indeed be necessary is some cases, it is undoubtedly easier to do once we have such ready pools of data that is likely to be valid for our research questions or hypotheses. This perk of using online communities as data sources is something that can be further exploited with machine learning, which will be discussed in further text.

3. Machine learning

The term machine learning refers to the automated detection of meaningful patterns in data (Shalev-Shwarts and Ben-David, 2014, p. 7). As the name suggests, it is a computational way of having machines use principles of ’learning’ in order to be able to perform tasks.

Roughly speaking, learning is the process of converting experience into expertise or knowledge. A human being may use his or her prior experience with events, movements, sensory input, etc, in order to learn or improve their performance at a specific task. In the case of machine learning, the learner is the machine (a computer), while the experience is the input data, and the output is some expertise at a task we have fine-tuned the machine for (Shalev-Shwarts and Ben-David, 2014, p. 19).

A successfull learner should be able to progress from individual examples to broader generalizations, also known as inductive reasoning or inductive inference (Shalev-Shwarts and Ben-David, 2014, p. 20). In the same sense, in machine learning, the algorithm being trained is using the training data inputs in order to find generalizable patterns in those inputs, that can then be used on other, unseen data.

Broadly speaking, there are two major types of machine learning- supervised learning and unsupervised learning (Shalev-Shwarts and Ben-David, 2014, pp. 22-23; Bonaccorso, 2017, p. 10).

The term ’supervised’ and ’unsupervised’ refer to the relationship between the learning process and the environment in which it is happening and data provided. For supervised machine learning, the data provided includes relevant information such as labels of cases in the training set. Based on these provided labels, the algorithm can connect patterns in data with specific labels, and in this way learn to predict labels based on underlying patterns in the data (Shalev-Shwarts and Ben-David, 2014, p. 23).

The main concept of supervised learning is the ’teacher’ or ’supervisor’ (the environment providing the labels), whose main purpose is to provide the learner with precise measure of it’s error. Using this information, the learner can correct the parameters in order to better the accuracy of it’s predictions. The main goal of supervised learning is to be able to learn and generalize the patterns from the training data in order to use these patterns to process new, unseen data. The algorithm that learns the patterns of training data ’too well’, to the point of learning specific and non-generalizable patterns that are unique to the training set at hand, is said to have overfit and is largely not useful for new data (Bonaccorso, 2017, p. 10).

Supervised learning can do predictions in two distinct ways: regression and classification. In the case when the desired output is a continuous value, while in the case when there is a discrete number of categories it is called classification. Common supervised learning applications include: predictive analysis based on regression or categorical classification; spam detection; pattern detection; sentiment analysis; automatic image classification; automatic sequence processing (for example, music or speech); and, finally, natural language processing (Bonaccorso, 2017, pp. 11-12). Since natural language processing (shortened as NLP) is of crucial importance for our analysis, there will be more word about it later.

Unsupervised learning, on the other hand, there are no labels or similar information provided in the learning environment, and indeed, there is no ’training’ and ’test’ data (in most cases, at least). Instead, the learner is provided with data and tasked to find patterns within data and subsequently find similar subgroups (or anomalies) in the data (Shalev-Shwarts and Ben-David, 2014, p. 23). In situations where we need to group data without previously provided labels, unsupervised learning becomes extremely useful: it can be used to learn how a set of elements can be grouped based on the statistical similarity of their features. Common unsupervised learning applications include object segmentation (for example, users, products, movies, songs, and so on) ,similarity detection, and automatic labeling (Bonaccorso, 2017, pp. 13-14).

Finally, it can be said that there is a third type of learning termed reinforcement learning. If machine learning is framed trough the prism of having a supervisor that gives labels to data and patternt, versus the situation where there are no labels but only data and patterns, reinforcement learning can be considered an ’intermediate’ learning environment. In the case of reinforcement learning, the training exampes contain more information than the test examples (for example, labels provided), the learner is required to predict even more information for the test examples (Shalev-Shwarts and Ben-David, 2014, p. 23):

> For example, one may try to learn a value function that describes for each setting of a chess board the degree by which White’s position is better than the Black’s. Yet, the only information available to the learner at training time is positions that occurred throughout actual chess games, labeled by who eventually won that game. Such learning frameworks are mainly investigated under the title of reinforcement learning.

In other words, in reinforcement learning, the learner receives feedback from the environment in the form of rewards and penalties that help the learner understand whether an action is positive or negative in a specific situation. Bonaccorso (Bonaccorso, 2017, pp. 14-15) describes reinforcement learning as:

> Even if there are no actual supervisors, reinforcement learning is also based on feedback provided by the environment. However, in this case, the information is more qualitative and doesn't help the agent in determining a precise measure of its error. In reinforcement learning, this feedback is usually called reward (sometimes, a negative one is defined as a penalty) and it's useful to understand whether a certain action performed in a state is positive or not. The sequence of most useful actions is a policy that the agent has to learn, so to be able to make always the best decision in terms of the highest immediate and cumulative reward. In other words, an action can also be imperfect, but in terms of a global policy it has to offer the highest total reward. This concept is based on the idea that a rational agent always pursues the objectives that can increase his/her wealth. The ability to see over a distant horizon is a distinction mark for advanced agents, while short-sighted ones are often unable to correctly evaluate the consequences of their immediate actions and so their strategies are always sub-optimal. Reinforcement learning is particularly efficient when the environment is not completely deterministic, when it's often very dynamic, and when it's impossible to have a precise error measure.

A variety of machine learning algorithms exist of both types. Some of the known supervised learning algorithms are linear and logistic regression, random forest, naive bayes; while commonly used unsupervised learning models include kmeans and hierarchical clustering.

3.1 Deep learning

Deep learning is a term used to describe more advanced machine learning models, particularily artificial neural networks (ANN's), that have proved to be remarkably powerful algorithms for complex tasks. Artificial neural networks are designed to mimic the natural neural networks found in the human brain (Kinsley and Kukiela, 2020, p. 13; Kumar and Garg, 2018, p. 23).

Simply speaking, what this means is that they are composed of layers of interconnected 'neurons' that rely on weights, biases, activation functions and backpropagation in order to learn very complex interdependencies and patterns in the data that is fed to them.

In biological systems, neurons are interconnected and communicate with each other. The transmission of signals occurs through axons, which carry input-output signals. These signals, in the form of electrochemical impulses, rapidly travel through the network. Neurons can either store information or transmit it to neighboring neurons via their dendrites.

Similarly, artificial neural networks (ANNs) mimic the functioning of biological neural networks. ANNs consist of interconnected artificial neurons. Each neuron in a layer is connected to every neuron in the previous and next layers. The connections between neurons are assigned weights (Kumar and Garg, 2018, p. 23).

During operation, each neuron receives inputs, which are the outputs of neurons from the previous layer. The neuron processes these inputs and generates an output, which is then forwarded to the neurons in the next layer. Activation functions, present in each neuron, collect and sum the inputs to generate the output (Kumar and Garg, 2018, p. 23).

A basic artificial neural network comprises three layers: the input layer, the hidden layer, and the output layer. The input layer receives input vectors, and the number of neural nodes in this layer corresponds to the number of input attributes. The output of each neuron in the input layer is transmitted to every neuron in the hidden layer, where the main processing occurs. The number of nodes in the hidden layer is initially chosen randomly and may be adjusted during training. The outputs of the hidden layer neurons are then forwarded to the output layer, where they serve as inputs. The output layer generates the final output of the network. The number of nodes in the output layer depends on the type of output desired. For classification problems, the number of nodes matches the number of classes, while regression problems may have a single output node for producing a continuous output value (Kumar and Garg, 2018, p. 23).

Another important concept for understanding ANN’s and deep learning is backpropagation- the process by which the neuron weights are updated, based on error in prediction, so that their performance is increased.

Firstly, the data is fed to the network and it goes trough a forward pass, where activations of each neuron are activated layer by layer, untill the output is obtained. After that, an error in predicted value versus the real value is calculated. Based on this calculated error (called’loss’), starting at the output layer, the error is propagated back trough the network, and the impact of each neuron on the error is calculated, based on it’s weight. The weights are then updated in order to minimize error. This process is repeated multiple times untill the network reaches a significant improvement in it’s performance (Kinsley and Kukiela, 2020, pp. 180-214). This algorithm, inherent to ANN’s, is the key component in their abilitiy to solve complex issues so efficiently.

3.2 Natural language processing

Since the topic of this thesis directly concerns the application of deep learning language models to other texts in order to understand and classify them, it is necessary to have a more in-depth discussion about Natural Language Processing, or NLP, here.

Natural Language Processing (NLP) is a field that explores how computers can understand and manipulate natural language to perform useful tasks.It involves computational techniques for analyzing and representing texts to achieve human-like language processing. The term "natural" distinguishes human language from more formal languages like mathematical notations or programming languages, which have limited vocabulary and syntax (Joseph et al, 2016, pp. 207-208).

Unlike artificial languages such as programming languages, natural languages have a much more complex and broad vocabularies, meanings and syntaxes, and processing them in a computational was is difficult.

Language serves as a means to share meanings rather than merely encoding information. NLP draws from various disciplines such as computer science, linguistics, mathematics, engineering, psychology, artificial intelligence, and robotics. (Joseph et al, 2016, pp. 207-208). Written text, as a storage of meaning, is then an extremely useful data source. For this purpose, NLP has developed many quantitative, statistical techniques and methods of extracting this meaning in ways that can be systematically analyzed.

Due to the complexity and ambiguity of natural language, the process of developing these ways of extracting features and analyzing them has been lengthy and dependent upon advances in computational power and statistics. The origins of NLP can be traced back to the 1950s when it emerged at the intersection of artificial intelligence and linguistics. Initially, NLP and text information retrieval (IR) were separate fields. IR utilized statistics-based techniques to efficiently index and search through extensive text collections, as described in Manning et al.'s comprehensive introduction to IR. Over time, NLP and IR have experienced some convergence. Presently, NLP draws inspiration from a wide range of diverse fields, necessitating researchers and developers to expand their knowledge base considerably (Chapman et al, 2011, p. 544).

The early NLP approaches, such as the simple word-for-word translation machines, have revealed the difficulty of the task at hand when confronted with homographs, the words that are written in the same way, but have multiple meanings. Furthermore, natural language poses challenges in standard parsing approaches that rely solely on symbolic, hand-crafted rules due to its vast size, lack of restrictions, and ambiguity. Extracting meaning from text, known as semantics, requires expanding grammars to address natural language semantics by incorporating additional rules and constraints. However, because the rules and grammars of actual natural language are so vast, this approach often leads to an overwhelming number of rules that interact unpredictably and result in more frequent ambiguous interpretations. In addition to this, handwritten rules struggle to handle "ungrammatical" spoken prose and the telegraphic prose commonly found in medical progress notes, despite being comprehensible to humans (Chapman et al, 2011, p. 544).

In the 1980s, a fundamental reorientation occurred in the approaches to extract meaning from text. Fundamentally, it involved replacing ’deep’ analysis with simple, robust approximations, adopting more rigorous evaluation practices, and embracing machine-learning methods that utilize probabilities.

The advent of statistical NLP marked a significant milestone in this reorientation. Statistical parsing, for example, addresses the proliferation of parsing rules by employing probabilistic context-free grammars (CFGs). These CFGs assign probabilities to individual rules based on machine learning from annotated corpora. This approach replaces detailed rules with a smaller set of broader rules, using statistical frequency information to disambiguate. Other approaches construct probabilistic "rules" from annotated data, which build decision trees from feature-vector data. The goal is to determine the most likely parse for a given sentence or phrase, with the notion of "most likely" being context-dependent (Chapman et al, 2011, p. 545).

In other words, instead of having detailed, precise rules that attempt to process the totallity of natural language- a task that has shown itself to be too complex and unpredictable- statistical NLP learns broad rules of language by relying on massive amounts of text examples and draws context-dependent probabilities from such examples by relying on machine learning. Statistical NLP benefits from learning with abundant real data, enabling it to handle common cases effectively. (Chapman et al, 2011, p. 545).

3.3 Semantic representation and text feature extraction

For texts to be processed trough a machine learning algorithm, certain steps need to be taken beforehand. Besides ensuring that we have a large enough and representative collection of texts, called ’corpus’, the textual data needs to be transformed into a shape and form that can be understood and learned by the machine.

In the book „Embeddings in Natural Language Processing“ (Pilehvar and Camacho-Collados, 2021), Pilehvar and Camacho-Collados state that, from a certain perspective, Natural Language Processing (NLP) can be divided into two main subfields: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU focuses on comprehending the meaning of human language, typically conveyed through text. For example, when a Question Answering system (such as ChatGPT) is presented with the question, it needs to first grasp the meaning of the words in the text, which is no easy task: „For instance, when a Question Answering (QA3) system is asked ’do penguins fly?’, the very first step is for it to understand the question, which in turn depends on the meaning of penguin and fly, and their composition.“ (Pilehvar and Camacho-Collados, 2021, p. 1)

In the book, the authors proceed to list challenges that make natural language understanding an extremely difficult problem for machine learning algorithms. Describing each of the ambiguities in detail would take too much space and doesn’t seem necessary for the purpose of this thesis, so they will be briefly listed with short descriptions, on which more can be found in the original text.

The most basic and fundamental problem of NLU is ambiguity. One of the major complexities of human language lies in its inherent ambiguity, which manifests in various forms. For instance, lexical ambiguity refers to the fact that individual words can belong to multiple syntactic classes (parts of speech) simultaneously (for example, both a noun and a verb). Then, there is syntactic ambiguity: a sentence can be syntactically parsed in multiple ways, so that it allows for multiple interpretations. Metonymic ambiguity, as the name suggests, is based on metonymy, which involves substituting a concept, phrase, or word with a semantically related one; and, lastly, anaphoric ambiguity, which pertains to the interpretation of pronouns (Pilehvar and Camacho-Collados, 2021, pp. 1-2).

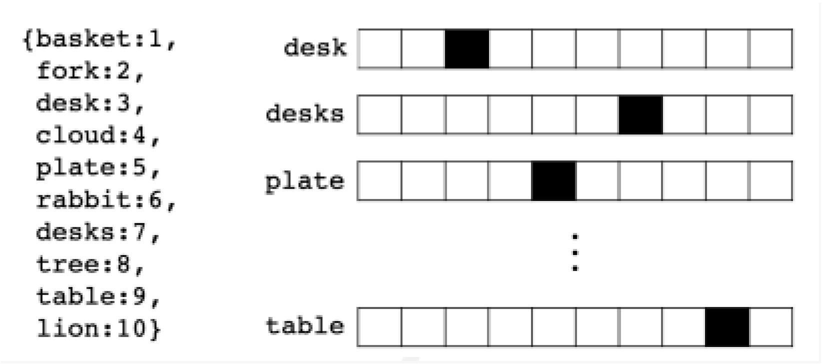
In themselves, words and sentences seem to carry too much ambiguity for their meaning to be knowable just from themselves. In the case of humans and everyday life, resolving many of these ambiguities necessitates knowledge that is not explicitly present in the context; it requires world knowledge or reasoning. Referential ambiguities that demand background knowledge for resolution are the focus of the Winograd Schema Challenge, considered an alternative to the Turing Test for assessing machine intelligence: „Similarly, it would be easy for humans to identify the intended meaning of mouse in “I ordered a mouse from Amazon” given their background knowledge from the marketplace.“ (Pilehvar and Camacho-Collados, 2021, p. 2)

Another problem for NLU are figurative expressions such as idioms ("fingers crossed," "all ears") and sarcasm. These forms of figurative language are extensively used by humans in both spoken and written communication, and their meanings often cannot be deduced directly from the individual words comprising them (Pilehvar and Camacho-Collados, 2021, p. 2).

And so, a question arises: in what way can language, specifically it’s written form, be transformed and processed so that the meaning and information it carries is understandable to the machine? There are several approaches and solutions to this challenge.

One popular method is the use of one-hot representations, which provide a simple yet effective way to encode words. For example, we can imagine a vocabulary with 100 words, and our goal is to represent each word as a so-called „one-hot“ vector. In the one-hot representation, each word is assigned a unique index ranging from 1 to 100. We create a fixed-sized array-like representation for each word, with a dimensionality equal to the size of the vocabulary (in this case, 100). All dimensions in the array are set to zero, except for the dimension corresponding to the word's index, which is set to one. This encoding scheme is why it is called "one-hot," as there is a single "hot" or active value (1) in the array, while the remaining values are zeros (Pilehvar and Camacho-Collados, 2021, p. 4). An attempt of visualizing this can be found in Figure 3.1.

Figure 3.1: An example of a 10 word vocabulary and corresponding one-hot encoding



Source: Pilehvar and Camacho-Collados, 2021, p. 4

In practice, for example, if we had a corpus of documents, each document would have different one-hot vectors assigned to it, depending on the word content of the document itself. This would constitute a set of different numerical arrays for different texts, with texts with a more similar word content having more similar vectors, thus allowing the machine to learn to discern among them based on their similarity or difference expressed in their numerical attributes.

While one-hot encoding serves as a basic representation, it has its limitations. It partially addresses the second limitation mentioned earlier, which is the lack of "similarity" between word representations. However, it still suffers from the first limitation, which is the absence of capturing the semantic relationships and similarities between words. For example, with one-hot encoding, there is no inherent way to encode the conceptual similarity between synonyms.

Another drawback of one-hot encoding is the increase in representation size as the vocabulary grows. In practical scenarios, such as natural language processing tasks, vocabularies can consist of countless words instead of just 100. Because vector size in one-hot encoding depends on the vocabulary size, representing each word using a one-hot vector in such cases becomes highly storage-intensive and computationally challenging (Pilehvar and Camacho-Collados, 2021, p. 5).

In summary, while one-hot encoding is a simple and intuitive representation of words, it lacks the ability to capture semantic relationships and suffers from the challenges of large vocabulary sizes. More sophisticated techniques like word embeddings address these limitations and have become widely used in various natural language processing tasks, and something more needs to be said about them.

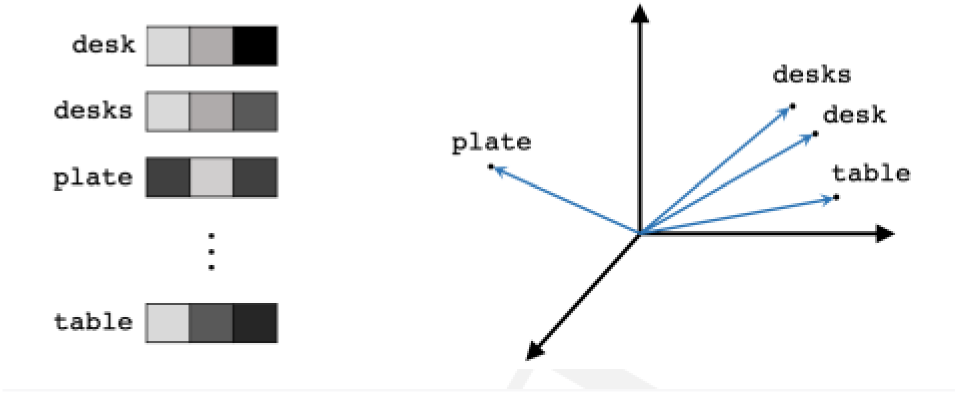
The Vector Space Model (VSM) offers a solution to overcome the limitations of one-hot encoding. It introduces a representation approach where objects are represented as vectors in a continuous multi-dimensional space known as the semantic space.

To illustrate the idea, we can imagine a simple 3-dimensional semantic space that represents four words with their respective vectors. This example highlights that one-hot encoding can be seen as a specific case of distributed representation, where each word is associated with a vector positioned along one of the axes in the semantic space. In practice, the semantic space consists of n dimensions, corresponding to the number of words in the vocabulary (Pilehvar and

Camacho-Collados, 2021, p. 6). Such a large number of dimensions is impossible for the human mind to imagine, but it surely is possible in the mathematical sense.

Shifting from the discrete and localized nature of one-hot encoding to distributed and continuous vector spaces brings several advantages. Notably, it introduces the concept of measuring similarity between words by evaluating their spatial distance in the semantic space, as visualized in Figure 3.2. Furthermore, employing a lower-dimensional space can accommodate a larger vocabulary, potentially addressing the size constraints of one-hot encoding (Pilehvar and Camacho-Collados, 2021, p. 6).

Figure 3.2: Representing the words from figure 3.1 in 3-dimensional vector space



Source: Pilehvar and Camacho-Collados, 2021, p. 5

One thing that was mentioned multiple times is that in Vector Space Model, localized semantic representation of words is replaced by distributed representation. This refers to the distributional hypothesis (Harris, 1954[3], Firth, 1957[4], in Pilehvar and Camacho-Collados, 2021, p. 7), the idea that words that occur in the same contexts tend to have similar meanings. However, while this foundational idea is indeed the basis of all Vector Space Models, the interpretation of the hypothesis and way of collecting „similarity clues" have undergone large changes (Pilehvar and Camacho-Collados, 2021, p. 7).

Around 2011, the field of natural language processing experienced a significant impact with the arrival of the deep learning revolution. This transformative wave brought forth various breakthroughs, and among them, Word2vec- a shallow neural network model available trough various programming libraries and designed to create vector space representations of

[3] Harris, Z. (1954). Distributional structure. Word, 10(2-3), 146-162.

[4] Firth, J. R. (1957). A synopsis of linguistic theory. Studies in Linguistic Analysis, 1952-1959, 1-32.

documents and words- emerged as a powerful force that propelled research in semantic representation to new heights. Despite its lack of depth in terms of network architecture, Word2vec showcased remarkable efficiency in generating concise vector representations by harnessing the capabilities of shallow neural networks. Following this influential development, the term "embedding" gained prominence and largely overshadowed the traditional notion of "representation," exerting a dominant influence over the realm of lexical semantics (Pilehvar and Camacho-Collados, 2021, p. 7).

Finally, the most recent surge in the field of natural language processing revolves around contextualized representation. This approach tackles the inherent limitations of static word embeddings by enabling the embedding to dynamically adjust itself based on the surrounding context. Unlike traditional word embeddings that operate in isolation, these models take into account the context in which words appear (Pilehvar and Camacho-Collados, 2021, p. 8).

3.4 Machine learning and NLP

In the context of natural language processing, machine learning models can be broadly classified as generative or discriminative. Generative methods aim to create comprehensive models of probability distributions and have the ability to generate synthetic data. Discriminative methods, on the other hand, focus on estimating posterior probabilities based on observations. As an example, generative approaches for identifying an unknown speaker's language rely on deep knowledge of multiple languages, while discriminative methods rely on differences between languages to find the closest match (Chapman et al, 2011, p. 546). The difference seems subtle, yet it is an important one- in the case of generative machine learning (or, generative AI, as it is often called), what is being learned is the probability distribution of the entirety of data, which is not the case with discriminative machine learning. Generative approach is much more knowledge-intensive, and has additional features such as generating synthetic data.

While there are many natural language processing algorithms and machine learning models, it is beyond the scope of this paper to go into the detalis on each and every one of them. Instead, at this point, I will say something about the ’transformer’ model, which is the basis of GPT models that will be used for this analysis.

3.5 Transformer models- from sequence-to-sequence convolutional network to attention mechanism

The Transformer model was first introduced in the paper „Attention is all you Need“ (Vaswani et al, 2017), authored by researchers from Google and University of Toronto. Prior to the introduction of Transformer models, recurrent neural networks (RNNs) such as long short-term memory (LSTM) and gated recurrent neural networks (GRU) have been widely used in sequence modeling and transduction tasks like language modeling and machine translation (Vaswani et al, 2017, pp. 1-2).

In contrast to individual words, it is not practical to pre-train embeddings for all possible sequences in a natural language since their number is infinite. Therefore, text sequences are typically represented by combining the embeddings of their constituent words.

The most basic approach, known as the bag of words (BoW) model, involves computing the sequence representation by averaging the embeddings of its words. Essentially, the set of words is represented by the centroid point in the vector space. However, the BoW representation treats all words equally in the final sequence representation, lacking consideration for words that may hold greater semantic significance. To address this, variations of BoW assign weights to words during the combination process using TF-IDF or other information measuring schemes.

Furthermore, an inherent issue with the BoW representation is its disregard for the order of words, which can have semantic implications:„For instance, the semantically different sequences ’rain stopped the match’ and ’match stopped the rain’ (and many other ungrammatical sequences constructed using these words) will have an identical BoW representation.“ (Pilehvar and Camacho-Collados, 2021, p. 13)

To tackle these challenges in natural language processing, Recurrent Neural Networks (RNNs) have emerged as a powerful solution. RNNs possess unique architecture characterized by recurrence, enabling the network to exhibit temporal dynamics and retain information from the past. Unlike feedforward networks, such as fully connected and convolutional neural networks, RNNs can capture the sequential order of data, making them suitable for NLP tasks.

In feedforward networks, input is typically processed all at once, lacking the ability to capture sequential relationships unless specific measures are taken. However, RNNs with their recurrent connections offer a means to effectively model sequential data, preserving the order of words and enabling better representation of their contextual meaning (Pilehvar and Camacho-Collados, 2021, pp. 13-14).

For the tasks such as translation or question answering, a model belonging to the category of RNN’s and relying on encoder-decoder architecture was developed. They are known as sequence transduction models, or Seq2Seq models. The architecture of a typical Seq2Seq model follows the encoder-decoder structure. In this structure, two recurrent neural network (RNN) modules are employed as the encoder and decoder. The encoder takes an input sequence (x1, ..., xn) and produces a sequence of continuous representations (r1, ..., rn). The decoder's role is to decode the representation sequence (r) into an output sequence (y1, ..., ym). Initializing the decoder with the final output and state of the encoder, it generates the first output token. Each output token is selected based on the highest probability assigned by the softmax layer, which spans across the vocabulary (Pilehvar and Camacho-Collados, 2021, pp. 18-19).

The preceding descriptions of RNN’s and seq2seq models are brief and there is no more space in this thesis to explain them in detail, since the reason they are mentioned is solely because they are crucial for understanding the Transformer model. Their ability to capture sequential order of data and create word embeddings that capture the context of the words was a gread advancement in natural language processing and allowed for deeper machine „understanding“ of natural language. However, the models had some issues and limitations. As the authors of „Attention is all you Need“ point out, recurrent models perform computations sequentially based on the positions of symbols in input and output sequences. This sequential nature limits the ability of the model to process computations simultaneously or in parallel, and becomes problematic with longer sequences due to memory constraints (Vaswani et al, 2017, p. 2).

The Transformer model is, in truth, a specific instance of a seq2seq model that relies on the encoder-decoder architecture (Amatriain, 2023, p. 3). The difference with the Transformer is that instead of relying on RNN’s in encoder and decoder parts, it consists of feedforward networks, thus circumventing the parallelization problem. However, since it was already said that feedforward networks, unless additional measures are taken, cannot capture sequential data, the question arises what may the „additional measures“ be in the case of the Transformer?

What truly makes the Transformer stand out is it’s reliance on the Attention mechanism (Amatriain, 2023, p. 5; Jurafsky and Martin, 2023, p. 2). Transformers are powerful models that can map input sequences (x1,...,xn) to output sequences (y1,...,yn) of the same length.

Composed of transformer blocks, which are multilayer networks, transformers utilize self-attention linear layers, feedforward networks, and self-attention layers. Self-attention is a crucial aspect of transformers, allowing the model to directly extract and incorporate information from larger contexts without relying on intermediate recurrent connections found in traditional recurrent neural networks (RNNs)(Jurafsky and Martin, 2023, p. 3).

At the heart of attention-based approaches is the ability to compare an item of interest with other items to determine their relevance in the given context. In the case of self-attention, these comparisons occur within a sequence itself. The results of these comparisons are then used to compute the output for the current input (Jurafsky and Martin, 2023, p. 4). By employing this technique, the decoder in a sequence generation task can perform a flexible exploration to identify the key words necessary for generating the current output token. This enables the decoder to concentrate on specific segments of the input sequence where relevant information resides. In essence, the encoder is not constrained to compress all the information from the source sentence into a single fixed-size vector. Instead, it encodes the input sentence into a sequence of vectors, which are subsequently utilized by the decoder to generate the output sequence (Pilehvar and Camacho-Collados, 2021, pp. 20-22).

In other words, based on input sequence, the encoder generates a sequence of vectors, each representing a specific position or word in the input sequence. During the decoding process, the decoder performs a soft search, assigning attention weights to these encoded vectors based on their relevance to the current output token being generated. The attention weights are made by calculating ’scores’ for the current output token and all other words in the sequence, and the closer the words are, the bigger the score. At this point, I think it is important to highlight the fact that, prior to being fed to the Transformer, input sequences of words already have embeddings and vector representation, and mentioned closeness refers to the closeness in semantic space, and all calculations are between values of these vectors.

By assigning higher attention weights to the most important words or positions in the input sequence, the decoder can focus on the specific parts of the input that are crucial for generating accurate and meaningful output. This adaptive attention mechanism allows the model to capture the dependencies and relationships between different parts of the input and output sequences more effectively.

In summary, the attention mechanism enables the decoder to selectively attend to different elements of the input sequence, improving the model's ability to generate accurate and contextually relevant output.

3.6 Advent of ChatGPT

The Transformer model, relying on Attention mechanism, is the current state-of-the art model in most NLP tasks. Relying on large language models, pretrained at enormous amounts of text, and using advanced mechanisms to process inputs and generate outputs, this model has the ability of truly deep understanding of the language it processes, which gives it great advantage for multiple tasks involving language processing, in comparison to ordinary machine learning algorithms.

During the last couple of months at the moment of the writting of this thesis, the popularity of the model has exploded, mostly thanks to the success of OpenAI's ChatGPT (GPT being the model at the basis of the chatbot, shortened for 'General Pretrained Transformer').

ChatGPT showed unparalleled ability of understanding and generating natural language, and it has virtually set off an 'arms race' between big companies such as Bing, Google, and others in who can develop better and more advanced models of the same type.

However, an interesting feature that OpenAI made available to developers is the access to their language models, embeddings and Transformers of varying power. These models and embeddings, available trough an API in python programming interface for a small fee per processed token, can be utilized for a variety of tasks. Importantly, the very powerful embeddings provided by OpenAI's language model can be used for embedding any novel texts; and several transformer models of varying power can be fine-tuned for classification tasks involving texts and documents. These two functionalities will be of primary importance for the analysis of this thesis.

4. Machine learning application in detecting suicidal ideation – problems and solutions

Coming back to the topic of this thesis, machine learning and natural language processing have proven to be a suitable approach to detecting suicidal content in online user posts. To my knowledge, there are countless researches written on this topic, with various machine learning methods used, social media and post types analyzed, and different data preprocessing techniques (Ji et al, 2021; Fatima et al, 2021; Rahaman et al, 2022; Chatterjee et al, 2022). Since the number of research papers written during the last couple of years is too large, I think it is excessive to comment and review every type of research on the topic just for the sake of it.

Instead, in the following text, I will try to lay out the purpose and goal of this thesis, as well as the problems encountered in reaching it, and the relevant research that encountered similar problems. I will try to discuss the solutions they offered, and offer a different one.

The main goal of this thesis is using Large Language Models to detect content with suicidal ideation in a corpus of texts, with the focus on discerning the content which contains suicidal ideation against the content that contains discussions about mental health problems, but not suicidal ideation. The reason why I think this is a challenging task has been discussed in the introductory part of the thesis, but it may be helpful to reiterate it here as well.

The connection between suicidality and mental health issues certainly exists. Suicidal ideation usually happens as the part of the suicidal process, which is usually characteristic for individuals in whom the development of a mental illness already exists (Roškar and Paska, 2021, p. 28). In fact, it can be argued that suicidal ideation is usually the result of many mental health conditions, but not the other way around.

This closeness of mental health issues and suicidal ideation implies that the two topics are similar, and so, any data and written expressions of the two would be closer to each other compared to many other topics. This similarity in data would mean that the models trained to discern and classify the two topics (suicidal ideation and generic mental health issues) would need to be sensitive to context and able to detect the subtle linquistic changes and markers that characterize content with suicidal ideation, such as the context specific usage of death-related words or even more subtle increase in self-centric words indicative of social disengagement (Feldhege et al, 2022, pp. 976-977).

For any model- no matter how powerful- to be able to do this, it would first require a high-quality data with proper labels that it can be trained on. A notable hurdle in the field of suicidal ideation detection, however, is the absence of a readily available public dataset. This challenge arises from the societal stigma associated with mental illnesses and suicidal thoughts, making it traditionally arduous to obtain relevant data. Nonetheless, there is a growing trend of individuals utilizing the internet as a platform to express their frustrations, seek assistance, and engage in conversations surrounding mental health issues, which has proven to be a useful source of data in a number of research projects (Haque et al, 2022, p. 4).

Thus, scraping the data from internet platforms and online communities is a potentially good source of data. However, if the data is scraped from online communities, a problem exists: how should the proper labels be created? If we want to discern suicidal ideation from generic mental health issues content, but those two are inherently connected (as discussed previously) and probably co-occur together, how do we know which of these should we label as containing suicidal ideation and which one as being only mental health issues?

Broad solutions such as roughly estimating which is which are not suitable here, since the distinction between the two is relatively subtle.

A possible solution is having a mental health or suicide professional to evaluate each text in the corpus as containing suicidal ideation or not, and using their judgement as labels. While this is probably the solution that would offer highest quality of labeling, it is important to note that we are attempting to train a deep learning model, which requires very large amounts of data to be adequately trained. Using the services of a mental health professional for such a task could be both expensive and time-consuming.

4.1 Using subreddits of origin as the basis of the labels

At this point, looking at another study that had a similar aim as this thesis might be of use, since the authors propose an alternative solution. I am reffering to an interesting study that tried to detect shifts from mere depressive online content without suicidal ideation to more suicidal posts. In this study (Choudhury et al, 2016), the authors attempt to detect and analyze the linguistic features of posts made by authors who post only on mental health related subreddits, and those who proceed to post on the subreddit Suicide Watch- the subreddit dedicated to suicide support. The main idea of the paper was to identify subtle indicators of shift from mental health posting to suicidal posting (Choudhury et al, 2016, p. 2). The main metric used for this was whether a poster posted only in mental health subreddits, or if the poster later proceeded to post in Suicide Watch subreddit as well.

Considering posting to Suicide Watch as an indicator of suicidal ideation in the study is not without basis.. Reddit is an established online community known for curating social news and facilitating discussions. It encompasses various topic categories, with each specific area referred to as a subreddit. The Suicide Watch subreddit is a community dedicated to people sharing their suicidal thoughts and feelings, and as such, it contains texts that probably bear features that can be considered representative of suicidal content (Ji et al, 2021, p. 221).

There are substantial reasons to believe that Suicide Watch content can be considered representative as a source of content with high levels of suicidal ideation. Besides the rules of the subreddit clearly limiting the content to suicidal posts, it is noteworthy that the subreddit was subjected to expert evaluation during this study by Choudhury and colleagues, which confirmed that the content is, indeed, representative for content with suicidal ideation (Choudhury et al, 2016, pp. 5-6). For this reason, the content of this subreddit is often used as a source of posts with suicidal ideation, and it's posts are labelled as such. As for the non-suicidal content, in many research papers posts are sourced from other popular subreddits (Ji et al, 2021, p. 221).

The paper, however, raises an important concern: the users may post on mental health related subreddits, and never engage in posting on Suicide Watch, even though they may indeed have suicidal thoughts and feelings (Choudhury et al, 2016, p. 6). This would, in turn, corrupt the data and make any analysis relying on Suicide Watch alone as the default 'suicide' label lose it's validity. This also constitutes a concern more broad than this specific study, reffering to any future study attempting to analyze the data from mental health subreddits and Suicide Watch, as it is truly possible- even likely- that mental health subreddits could contain posts that are, in fact, suicidal. Training any sort of machine learning models on data labelled purely according to their subreddit of origin would, thus, be fallacious in these case, and would mislead the algorithms, causing their performance to drop.

The authors of the paper argue against this based on vaguely defined practice of channeling the suicidal contents from mental health subreddits to Suicide Watch, as well as famousness of Suicide Watch (Choudhury et al, 2016, p. 6), but I think it can be said that their argumentation is flawed for several reasons: it assumes a) moderators can correctly distinguish between all suicidal and merely depressive content, and b) that the suggested moving of posts with suicidal content to their proper subreddit (Suicide Watch) is strictly enforced - both contentious assumptions.

4.2 Alternative solution

The issue of adequate labelling of data scraped from online communities for the purpose of this study is a difficult one. If we want to train a model to discern between suicidal ideation and mental health issues contents, we have to somehow create labels for a large quantity of data. Relying on a simpler logic of labeling everything within Suicide Watch as ’suicidal ideation’ and everything outside of it as not belonging to that label would work only if we could prove that all posts with suicidal ideation end up being posted in Suicide Watch – a statement that is almost certainly not true.

However, having at least one half of the picture- the knowledge that all posts within Suicide Watch are probably containing very high levels of suicidal ideation, based on arguments previously discussed - can be used as a leverage.

What I suggest as the solution, and what I will attempt to do in the practical part of the thesis, is using this knowledge in combination with LLM-based document embedding also available trough OpenAI’s API, and clustering in order to create new labels. If we 1) know with sufficient certainty that Suicide Watch posts contain suicidal ideation, 2) we embed all the posts (both the ones from Suicide Watch and those from mental health subreddits) and compare them in the vector space, and 3) we find the cluster of posts that share the vector space with the largest number of Suicide Watch posts, then 4) we can infer that they are similar in content, and 5) we can reasonably assume that this cluster is the ’suicidal ideation’ cluster, for it includes both Suicide Watch posts, and other posts with suicidal content from outside of that subreddit. As such, it can be used as the new, ’corrected’ label for posts that are considered to contain suicidal ideation, and the classification models can be trained on it.

5. Research goals and questions

The goals of this thesis are as follows:

1. To explore the utility of large language models in solving the issue of creating adequate labels for posts containing suicidal ideation in a corpus of data retreived from online communities;

2. To explore the utility of large language models in accurately detecting suicidal ideation in such a corpus;

Research questions at the base of this thesis are:

1. Can embeddings from Large Language Models help group and cluster corpus contents in a way that helps distinguish between content with suicidal ideation and purely mental health issues content?

2. Can Large Language Models accurately classify texts based on whether they are merely mental health related or containing suicidal ideation?

6. Methods

6.1 Data used

In order to collect the data, Reddit’s official Pushshift API has been used. Due to the recent changes in regard of the API’s pricing, it is worth noting that, at the moment of data collection, the API and access to all subreddits’ data was still free of charge. Therefore, there were no costs related to collecting the data, although any new research that would seek to gather new data would probably need to pay.

The content scraped from Reddit originates in different subreddits. Subreddits are forums dedicated to specific topics on the Reddit website, and each of them has their rules, moderators, and content closely related to the specified topic.

The subreddits scraped for the purpose of this study were separated into three segments: suicide-related subreddits, mental health related subreddits, and ’other’ subreddits. Suicide ideation-related subreddits include only r/suicidewatch- a subreddit specifically dedicated for participants struggling with suicidal thoughts. Mental health subreddits encompass a variety of subreddits dedicated to specific mental health issues: r/depression, r/ptsd, r/mentalhealth, r/bipolar. Lastly, the ’other’ category contains generic subreddits that cover a variety of topics from everyday life and can include variety of information. It includes r/teenagers, r/askreddit, r/randomthoughts, r/nostupidquestions, r/outoftheloop, r/explainlikeimfive, r/latestagecapitalism, and r/cryptocurrency.

The result of the scraping is a total of 8896 posts, of which 933 belong to the suicide category, 3917 are from mental health subreddits, and 4149 belonging to ’other’ subreddits.

The reason subreddits were structured into these three groups is closely connected to the purpose of this thesis. The main purpose and goal of this thesis is to train an algorithm that relies on deep language understanding in order to distinguish between content that contains suicidal ideation and content that only contains mental health struggles (but no suicidal ideation). This task can be considered difficult, because suicidal messaging and mental health struggling are, indeed, very closely related, and yet, distinguished things. SuicideWatch provides content that is almost exclusively suicide-related, while mental health related subreddits provide content dealing with mental health struggles, and differentiating between them is the major challenge the model faces.

Meanwhile, ’other’ subreddits are there to provide generic content that is bound to be present in conversations and internet posts. This type of data, though not directly connected to detecting suicide within mental health context, is added to the dataset for the reason of additional testing complexity.

Reflecting on the number of posts belonging to each category, we have a dataset in which suicide posts are a a small fragment compared to other types of posts. I consider this to be a positive thing about the dataset, since it more closely reflects the realistic proportions: In many realistic situations, it can be reasonably assumed that suicidal posts compose a minority compared to the rest of the posts, In this sense, it should also be more difficult for the model to accurately detect suicidal ideation in the test set, which is a good thing for assessing how successfull the model is.

6.2 Creating the labels

So, with the primary model task being classification of texts based on whether they belong to the suicide category or to some other category, the mentioned texts need to be labeled in order for the model to learn.

In the previous research done on the topic that was mentioned previously, especially the one done by Choudhury (Choudhury et al, 2016), the labels were created based solely on subreddits of origin- meaning, all posts from Suicide Watch are labeled as ’suicide’ and those from mental health subreddits are labeled as ’mental health’. Their argumentation is that subreddits have rules that direct all users to post suicidal content to Suicide Watch, since it is a reputable and well known subreddit for these things. I argue against this, since it is 1) unlikely that a suicidal person will be so prudent about whether they post in r/SuicideWatch or r/Depression, and 2) even casual browsing of mental health subreddits reveals that there is a significant amount of posts that appear to be quite suicidal by content. In other words, while we can be sure that posts originating from r/SuicideWatch can indeed be considered suicidal both because of the design of the virtual community and the evaluation done by experts (Choudhury et al, 2016, pp. 5-6), we cannot be certain that all posts from mental health subreddits are non-suicidal- in fact, it would seem to be quite likely that a good amount of them could potentially be suicidal.

In practice, this creates a problem: machine learning algorithms are only as good as the data fed to them is good. Data in which a large amount of texts is possibly mislabeled as belonging to one of the non-suicide categories is bad data, and any model trained on it would be a bad model.

As a solution to this problem, I have created a workaround that could be a way to correct this labeling problem. There are two steps to the solution: 1) embedding the documents, and 2) clustering them and assigning labels according to similarity.

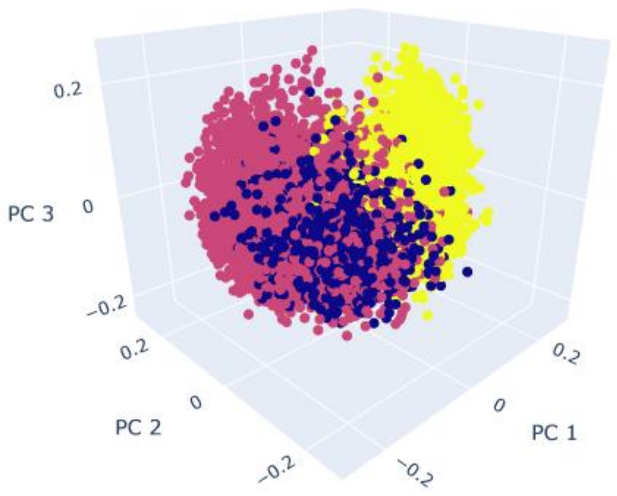
For embedding the documents, OpenAI's text-embedding-ada-002 model has been used. Currently, it is their most advanced document embedding model that is available trough their API (OpenAI, n.d.- b).

The assumption of embedding the texts in this way is that the texts that have similar context will have embeddings that are more similar, and that it will be possible to group them together in a sensible manner based on these similarities.

A way to show these similarities is by visualizing the embeddings in space. Since the embedded documents have too many dimensions to be feasibly visualized, first I have performed principal component analysis and used the first three components to visualize the data in a 3D space.

In Figure 6.1, each dot represents a post scraped from Reddit, while different colors indicate different subreddit of origin/category (dark blue = SuicideWatch/Suicide, pink = Mental Health, yellow = Other). Different positions alond the component dimensions indicate different embedding values, and the closer the dots are in the space, the more similar their content is:

Figure 6.1: Positions of embedded posts in 3D vector-space



Source: Author

- r/SuicideWatch
- Mental Health subreddits
- Other subreddits

Based on this visualization of posts in 3D space, a few things can be concluded:

- All three categories have visibly distinct groupings in the 3-dimensional space. This indicates that most posts from different subreddits share similar content, distinct from other visible groupings;
- The posts from r/SuicideWatch (dark blue) have visible grouping in one of the corners;
- However, as previously suspected, there appears to be a significant overlap between posts from SuicideWatch and a part of posts from Mental Health subreddits (pink);

The overlap of a segment of Mental Health content and Suicide Watch content seems to affirm the previously expressed suspicion that, while content from Suicide Watch may be confirmed to be about suicide, it would also be likely that some posts from Mental Health subreddits could also contain suicidal ideation.

This conclusion is based on two premises: one, that Suicide Watch content contains suicidal ideation; and two, that embedded documents that are similar to each other content-wise also have similar embedding values; from which follows that that embedded posts that have similar values to the embedding values of SuicideWatch posts, are similar to SuicideWatch posts- and, consequently, their content is similar to content containing suicidal ideation.

From the visualization above, it is visible that there are embedded posts from Mental Health that are deeply in the ’space’ of SuicideWatch, indicating that their content is also filled with suicidal ideation.

This leads me to my second point: clustering. Since there is a visible ’region’ of embedded posts that group around and intertwine with what is confirmed to be suicidal content, they could be clustered and grouped according to this similarity in content: all posts sharing the embedded space with SuicideWatch posts could be considered as posts containing suicidal ideation. Creating clusters and inspecting them could give us a new, similarity-in-embedding way of labeling posts as suicidal or non-suicidal.

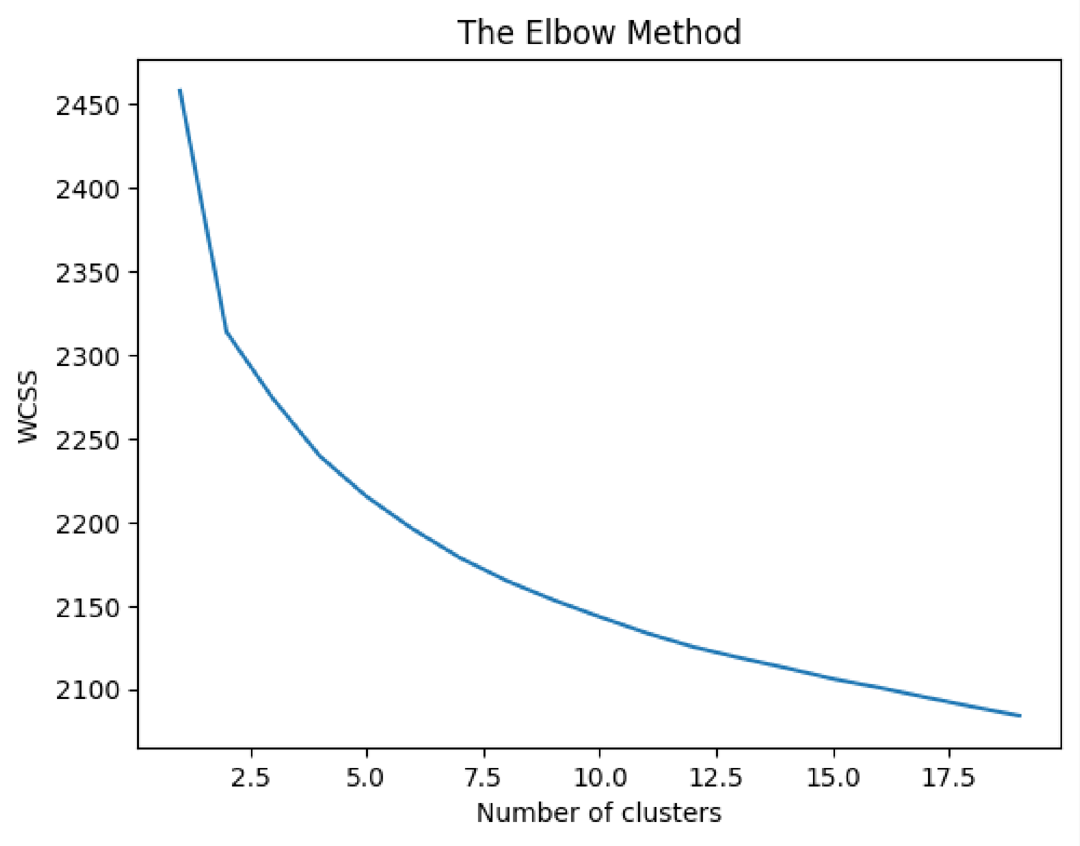
For this, I have used agglomerative clustering with cosine distance.

Agglomerative clustering is a hierarchical clustering method that builds a dendrogram by iteratively merging clusters. It starts by treating each data point as a single cluster and then repeatedly merges the closest pairs of clusters until only one cluster remains (Praveen et al, 2020, p. 2; Santoso et al, 2018, p. 38). Notably, it is common to use it in document clustering: in text analytics, it can be used to group similar documents (Prihatini et al, 2019, p. 2104).

Cosine distance is a metric used to measure the cosine of the angle between two non-zero vectors. It's derived from the cosine similarity formula and is particularly useful when dealing with high-dimensional data, like text data represented as word vectors. (Wang et al, 2023, pp. 6-8).

Before chosing the number of clusters and clustering, I did a quick preliminary analysis on the potentially best number of clusters using elbow method (Figure 6.2) and silhouette (Figure 6.3):

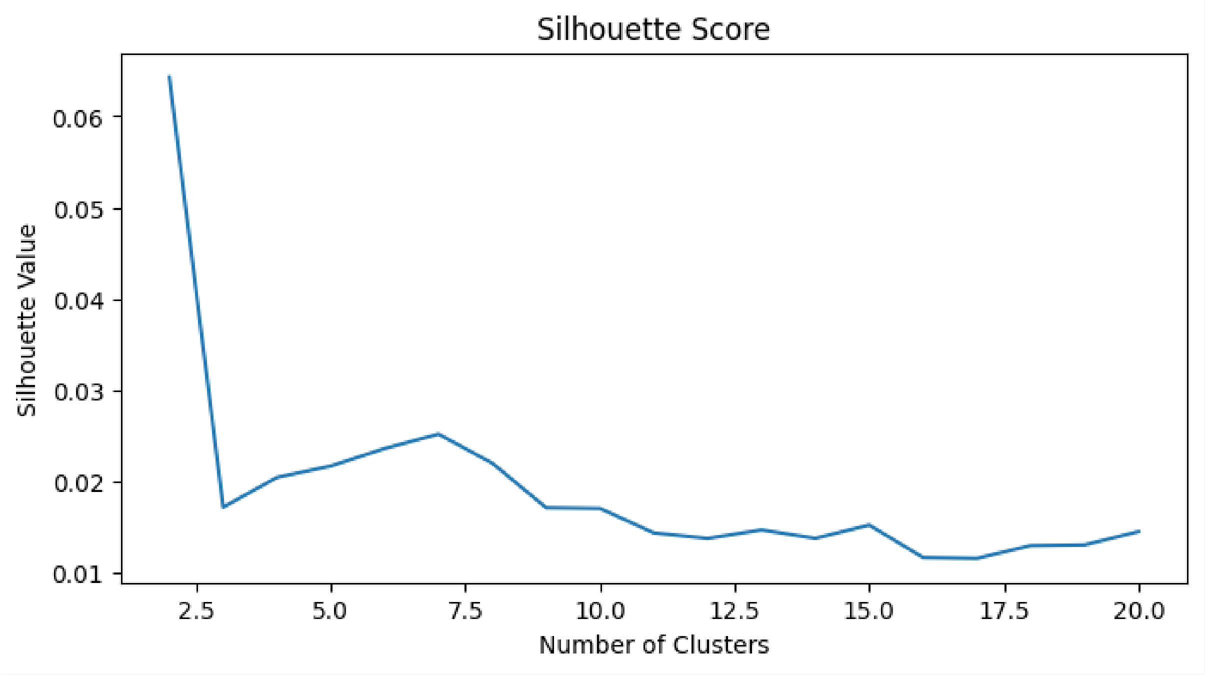
Figure 6.2: The Elbow method results



Source: Author

The elbow method does not provide much insight into the optimal number of clusters- there is a faint break at around two clusters, but this is hardly helpful, since it could simply be the break between 'other' and mental health/suicide parts of data, which does not cover the most importand distinction we are trying to find- the one between suicide and mental health posts. Because of this, silhouette metric was also tried:

Figure 6.3: The Sillhouette results



Source: Author

The results of silhouette appear more interesting, as they show some clear increases in score across different numbers of clusters, and indicate that- besides the two clusters- seven clusters may be the optimal number of clusters. It is worth noting, however, that the differences in silhouette scores are not dramatic, so this metric, on it’s own, should not be taken too seriously.

After some experimentation and tweaking the clustering parameters, I have decided on going for 5 clusters.

The reasoning for this is based on exploration of clusters and trying to find the one whose content is aligned with previously discussed conditions. The main criteria I assesed this was how much content from in each cluster originates from each of the three subreddit categories, with the assumption that one of the clusters will contain the majority of r/SuicideWatch posts, as well as large amounts of posts from Mental Health subreddits, and desirably low amounts of posts from ’Other’ subreddits, given how their content is supposed to be generic and unrelated to suicidal ideation.

The Figures 6.4 and 6.5 below visualize the contents of each of the clusters in terms of amount of posts from each subreddit category:

Figure 6.4: Number of posts from subreddit categories in different clusters (post counts)

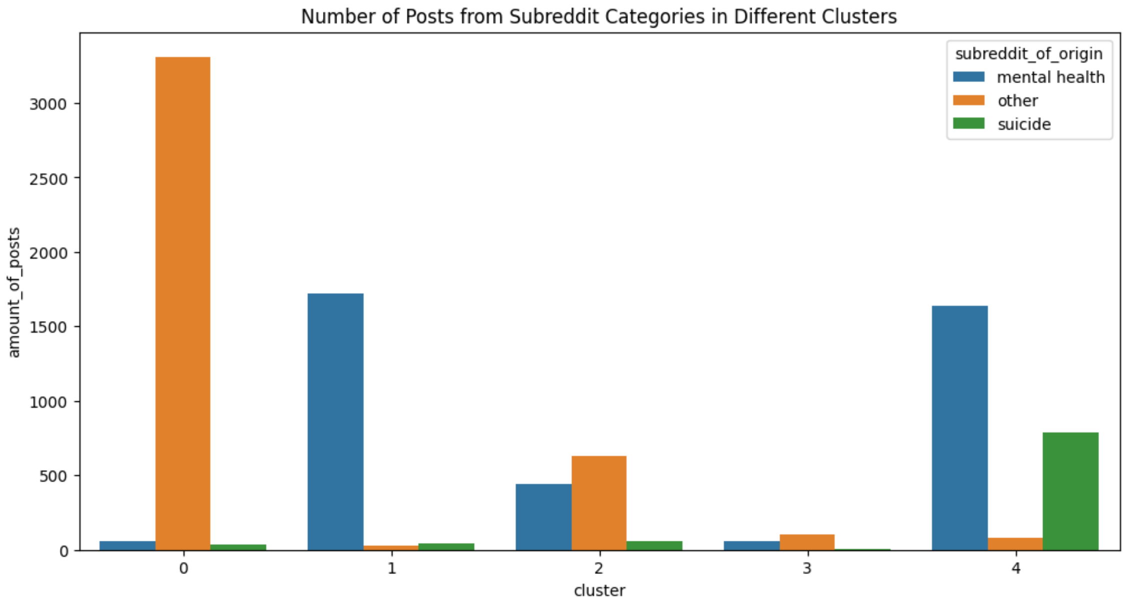
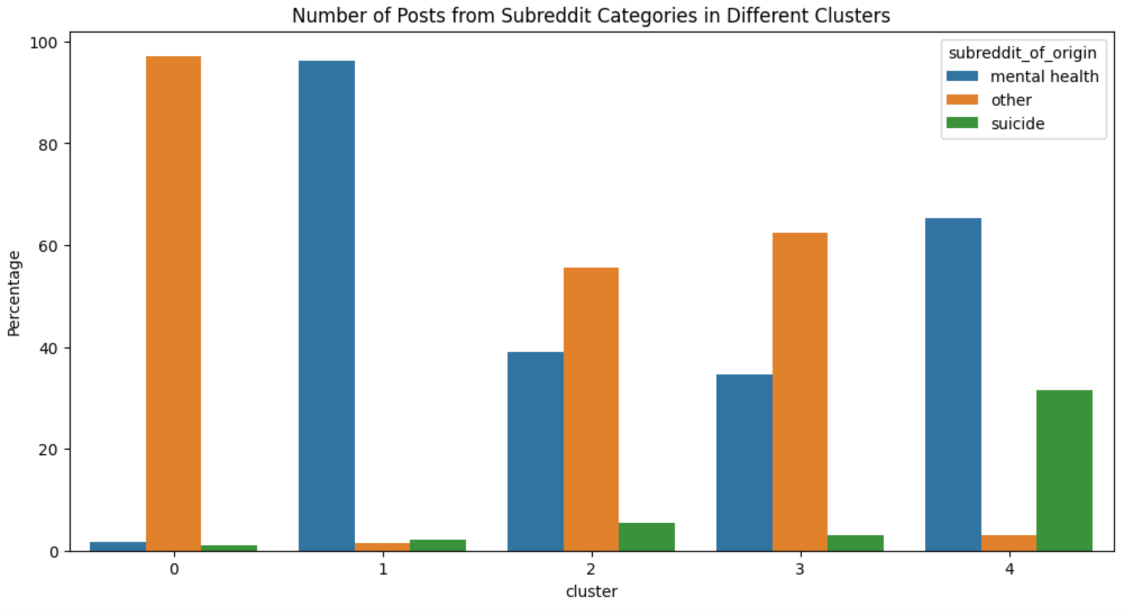


Figure 6.5: Number of posts from subreddit categories in different clusters (percentage of cluster)



Source: Author

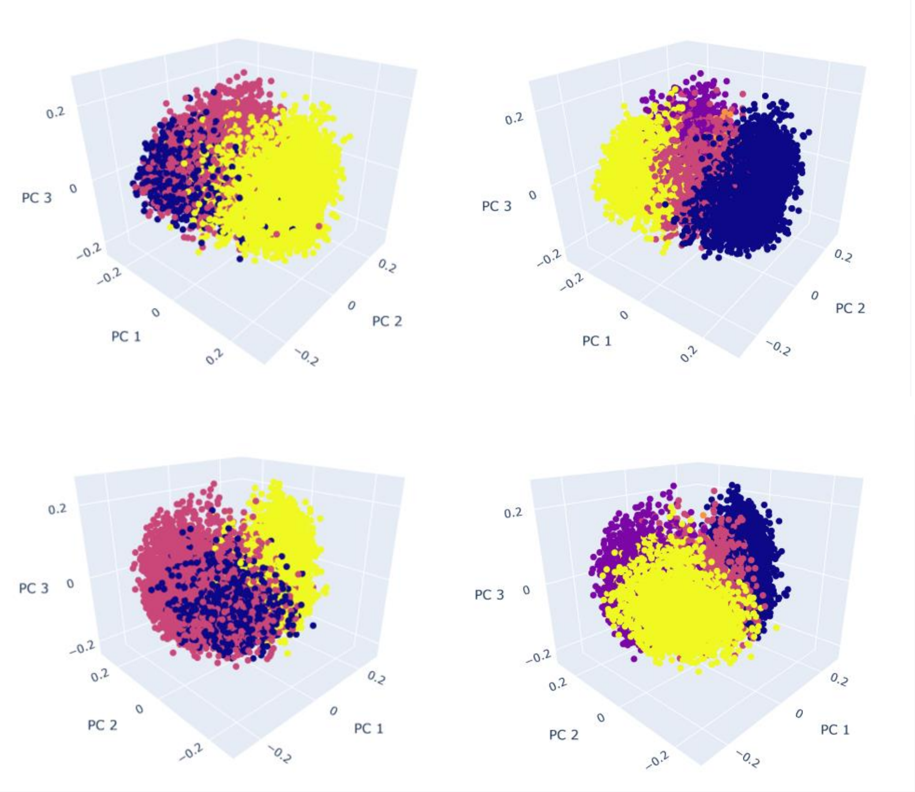
The clusters have some characteristics worth noting. All five clusters differ somewhat in content compared to each other. Firstly, Cluster 0 is almost exclusively composed from posts originating from ’Other’ subreddits. In a similar fashion, cluster 1 mostly contains posts from Mental Health subreddits. These two clusters seem to be well distinguished from each other, as their content has a clear pull to one or the other side.

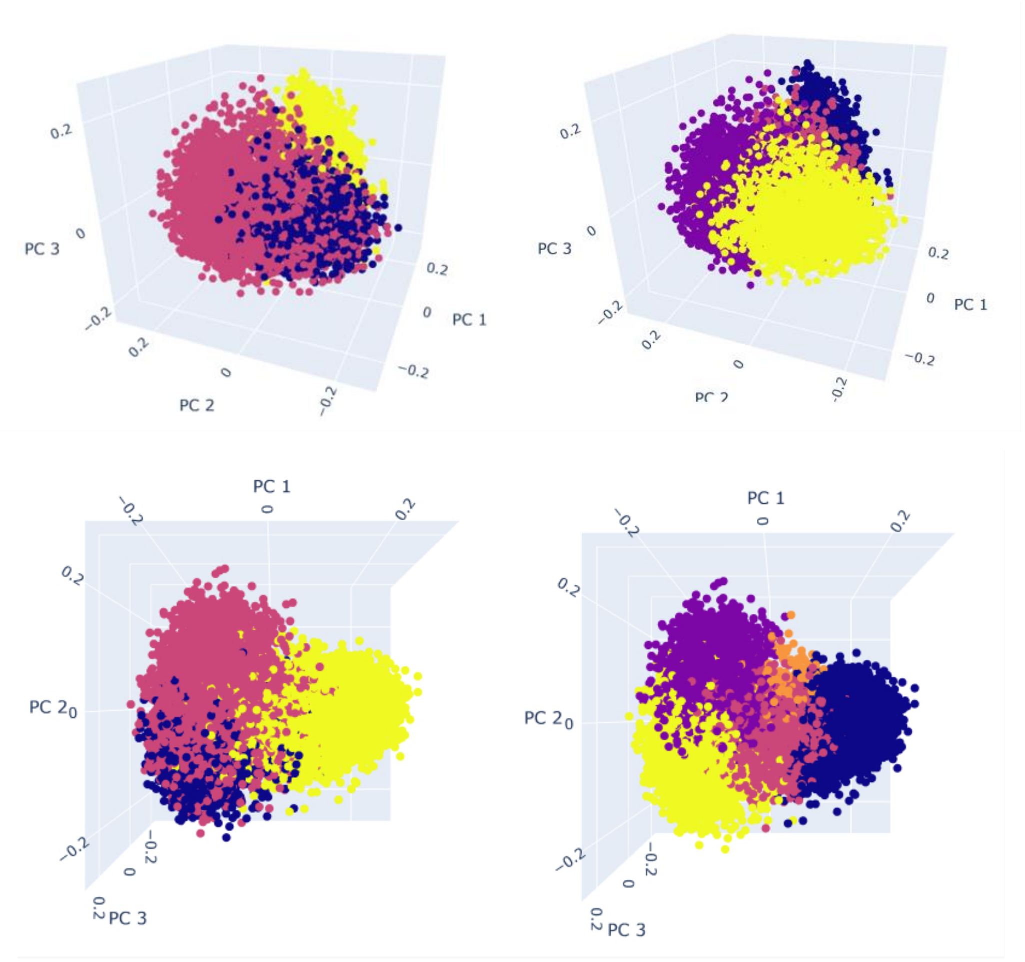
Next, clusters 2 and 3 are somewhat ambiguous, as their content has a number of posts from both ’Other’ and Mental health subreddits, but also a small but considerable amount of posts from r/SuicideWatch.

Lastly, the thing that is important to notice is that cluster 4 contains almost the entirety of r/SuicideWatch posts, as well as large portion of Mental Health subreddits posts. At the same time, it contains a minimal amount of ’other’ posts. As such, this cluster appears to be the one that contains posts which contain suicidal ideation, both from SuicideWatch and part of Mental Health subreddits.

If we visualize the clusters in 3D vector space like before and compare the new clusters with subreddits of origin categories, will give us Figure 6.6 below.

Figure 6.6: Comparison of Positions of embedded posts in 3D vector-space with colored subreddit of origin (left) and newly created clusters (right)





Legend:

● r/SuicideWatch	● Cluster 0
● Mental Health subreddits	● Cluster 1
● Other subreddits	● Cluster 2
	● Cluster 3
	● Cluster 4

Source: Author

In the 3D visuals above, it is visible that Cluster 4 covers the area in which the intersection of SuicideWatch posts and Mental Health posts happened. As indicated by the previous plot, it would seem that this cluster is adequate to be considered as the one that captures the suicidal content from the given corpus of posts, and can be therefore used for the ’corrected’ labelling of data.

However, it needs to be noted that it does not perfectly capture all of the SuicideWatch posts- which are assumed to be suicidal based on the subreddit’s purpose and content- leading to the question what do we do with them. Should the new ’suicide’ label contain only the content within Cluster 4, while all posts outside of it- including the small amount of those that originate from SuicideWatch- are excluded from the label? Or, alternatively, should the new ’suicide’ label include the Cluster 4 posts, plus the remaining SuicideWatch posts that were left out of it?

While I didn’t think the inclusion or exclusion of the number of SuicideWatch posts that were left outside of Cluster 4 would make much of a difference, I have decided to explore the SuicideWatch posts that were in other clusters- specifically, in clusters 2 and 3, as they were the most ambiguous ones.

Below are some of the examples from both clusters.

Cluster 2:

- „Hi im Dylan (m22) im going thru a rough patch rn and would just like to talk it out with someone if anyone has sum time to spare“
- „I’ve tried this once two days ago. I sat in my garage for about an hour just waiting. I have a 2019 dodge charger rt. Did i do something wrong? I want to try again tonight. All i got was a headache. So upsetting.“

Cluster 3:

- „Hello Everyone I am trying night night method. So i tried to compress my carotid arteries, At first nothing happened only headaches as i was pressing some other veins, After few hours of trying, I pressed some artery and after few seconds my heart started pumping faster and faster. Anxiety kept on increasing with little vision change and as soon as I left that artery my anxiety went away and heart also goes normal, I pressed that for about 5 to 7 sec. So i want to know is that the carotid artery or i need to try more or something else, I mean what are the indications that the artery I am pressing is the carotid one“
- „I'm thinking of taking 40 30mg dxm tabs and 14 238 mg AdvilPM along with some alcohol and 10 acetaminophen tabs 500 mgs“
- „AdvilPM 200 mg and diphenhydramine 38 mg“
- „If I manage to make my head go first into the water, would that kind of impact cause me unconscious?“

To have a better idea of how these differ from the majority of SuicideWatch posts that ended up within Cluster 4, here is an example of a typical post from Cluster 4:

- „What's the point of sticking around physically when I just feel so drained? I don't enjoy things anymore. I'm ready for things to end. I go to sleep every night hoping I won't wake up the next day. I'm tired of being passive and waiting for the universe to end things for me. I just want to go now. The person I love the most hates my guts. He finds distractions with other girls. I feel too hurt to move past this and move on with my life. I just want to die“

While the selected posts from clusters 2 and 3 do originate from SuicideWatch, they do seem somewhat different from the typical SuicideWatch content found in cluster 4. While they are, indeed, discussing suicide-related things, without the context that is known to the reader (the subreddit of origin, in this case) it would not be clear whether they are talking about suicide or not. For example, the posts from Cluster 2 could also just be asking someone to chat, or discussing technical difficulties with some machine in the garage; while posts from Cluster 3 could be having a calm discussion about medications or technicalities about some physical techniques.

The only thing revealing that all these questions discussions are not, in fact, calm and neutral, but related to suicidal ideation and attempts is that their subreddit of origin is SuicideWatch, and this information would be unavailable to the machine learning model, unless we include them under the ’suicide’ label. On the other hand, including these very technical or vague posts under the label ’suicide’ might cause the model to start generalizing them to the other equally technical and vague posts that- it can be reasonably assumed- do not always reffer to suicidal ideations or attempts. This would make the model’s accuracy in predicting to decrease.

Because technical discussions in posts can, at the same time, be a lagitimate expression of suicidal ideation if the context is provided, but they might also confuse the model and make it generalize and detect suicidal ideation where it is not present, the question whether to include the SuicideWatch posts outside the Cluster 4 under the ’suicide’ umbrella was a difficult one. After some consideration, I have decided to train two models and compare their success- one based on labels in which only Cluster 4 is labelled as ’suicide’, and one where the mentioned posts are included under the label as well.

To sum this part up, I have used document embedding and agglomerative hierarchical clustering to create a corrected ’suicide’ label that would not be based only on the post originating on r/SuicideWatch. This was done out of suspicion that many posts from Mental Health subreddits could be containing suicidal content, but would mistakenly be labeled as not being ’suicide’ solely for not originating from r/SuicideWatch. By creating a cluster that covers both the vast majority of r/SuicideWatch posts and all the Mental Health posts that share similar embeddings in the vector space, I created a new way of labelling posts as ’suicide’ based on whether they are in this cluster or not. I have additionally created another system of labelling that also uses this cluster as the new ’suicide’ label, but also includes the remaining r/SuicideWatch posts under the label as well.

In the next section, I will discuss the models that were trained and fine tuned on this data.

6.3 Fine-tuning the models

The data, labelled in both ways, was used to train the models.

For the task of fine-tuning the models, OpenAI’s API was used. Trough OpenAI’s API, it’s pretrained models can be accessed and fine-tuned for variety of tasks. The API is accessed trough the openai library in Python, and fine-tuning the models is a paid service.

Fine-tuning is a process that involves training a pre-trained model on a new dataset to adapt it to specific tasks. OpenAI's base models are trained on a large corpus of text from the internet, but they don't have specific knowledge about the data they were trained on. Fine-tuning helps in adapting these base models to specific tasks or domains by training them on custom datasets. For example, the models currently used for ChatGPT are fine-tuned for chatting, but the models can be fine-tuned for different tasks as well. The process involves creating an adequate dataset, training the model, and then evaluating its performance (OpenAI, n.d.-a).

Classification is one of the tasks that can benefit from fine-tuning. For classification tasks, the dataset should be structured with a column for inputs and another for labels. The model can be trained to predict the correct label for a given input. Once the model is fine-tuned, it can be used to classify new unseen data based on the training it received (OpenAI, n.d.-a).

The idea behind fine-tuning a pretrained model is that the model learns to connect certain inputs with specific outputs, relying on it’s deep understanding of language. In the case of fine-tuning it for chatting, it would connect prompts with generated responses in the form of text. In the case of classification, it connects prompts with a very limited number of responsed, which are specified classes – in our case, the prompts are posts, and desired responses are ’suicide’, ’mental health’, and ’other’.

For a pretrained model to be fine-tuned, it requires a dataset of prompts and the desired response to those prompts. Technically speaking, the process of fine-tuning a model is teaching it to connect the prompt with the response. For classification, this means connecting the texts with the classes they belong to (OpenAI, n.d.-a).

The datasets I have provided for the model to learn from contained two columns- one with posts, and the other one with the class labels I created. In order to fine-tune a model on this dataset, OpenAI’s API has special requirements that must be fulfilled in order for the dataset to be accepted by the API as fitting for fine-tuning. These requirements include 1) specific labeling of columns as ’prompt’ and ’completion’ for the model to recognize what the prompt and expected completion are; and 2) transforming the dataset into JSONL format. While the latter step can be done trough various Python libraries and transformations, it is convenient that the API itself offers free tools for transforming the dataset into the apropriate format.

Trough the tools provided trough the API, I have 1) transformed the dataset into the appropriate format, and 2) split it into the training set and test set (1000 units). After that, trough using API commands, I have chosen the desired model to be fine-tuned, as well as the number of classes and if I want metrics to be calculated. After completing this step, the datasets and configuration are sent to OpenAI’s servers, where the fine-tuning process is being carried in entirety, which is useful considering that such process would probably not be possible on a local computer. The tuning of a single model can take several hours (3 hours on any session that I had), depending on how clogged the network is and how big the queue for using the service is, and the status of the job can be checked trough an API command, after which the results can be downloaded and analyzed.

At the time of fine-tuning the models on my data, four models were available trough the API, ranked by their power from weakest to the most powerful: Ada, Babbage, Curie and Davinci (OpenAI, n.d.-a). Since the company has been announcing depreciation of some of the models, it is not guaranteed all of them will be available in the future for replicating the research.

I have used Curie model and fine-tuned it twice, for both systems of labeling. Thus, as a result, I have two different fine-tuned models with slightly differently labelled data whose results I will present in the next section.

7. Results

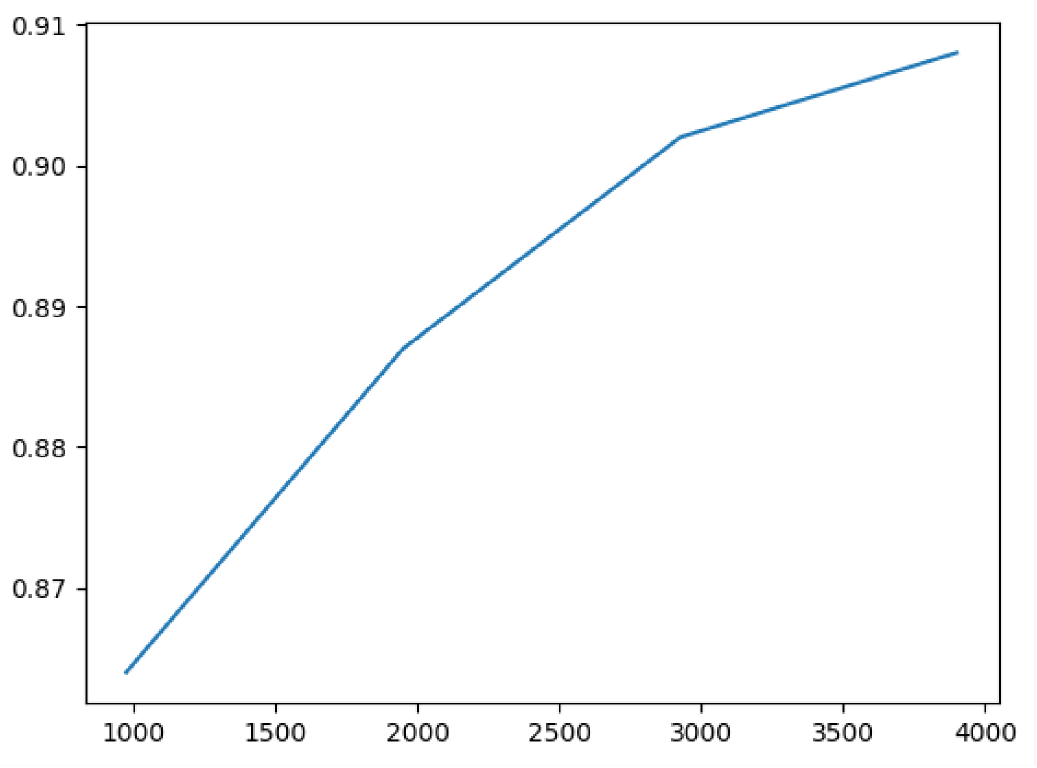
In the text that follows, results for both models will be presented- first, the model trained on data where only cluster 4 is used as ’suicide’ label(I will refer to it as Model 1), and second the model trained on data where cluster 4 and remaining SuicideWatch posts are taken as ’suicide’ label (Model 2).

The first of the shown plots represents the model’s accuracy over ’steps’ during model training. ’Steps’, found on the x-axis, represent a single update of the model’s weights.

Apart from this plot, the confusion matrix is also included, along with some of the metrics- accuracy for the models overall, as well as precision and recall for each of the classes.

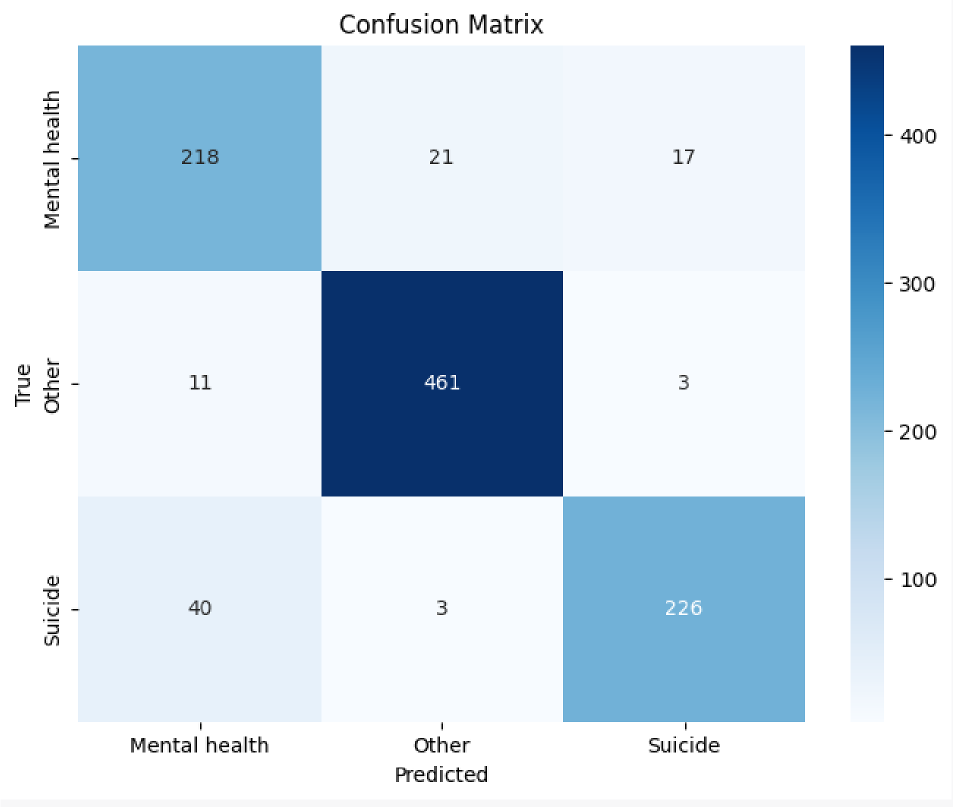
- Only cluster 4 as ’suicide’

Figure 7.1 Accuracy curve during steps for Model 1



Source: Author

Figure 7.2: Confusion matrix for Model 1



Source: Author

Accuracy: 0.905

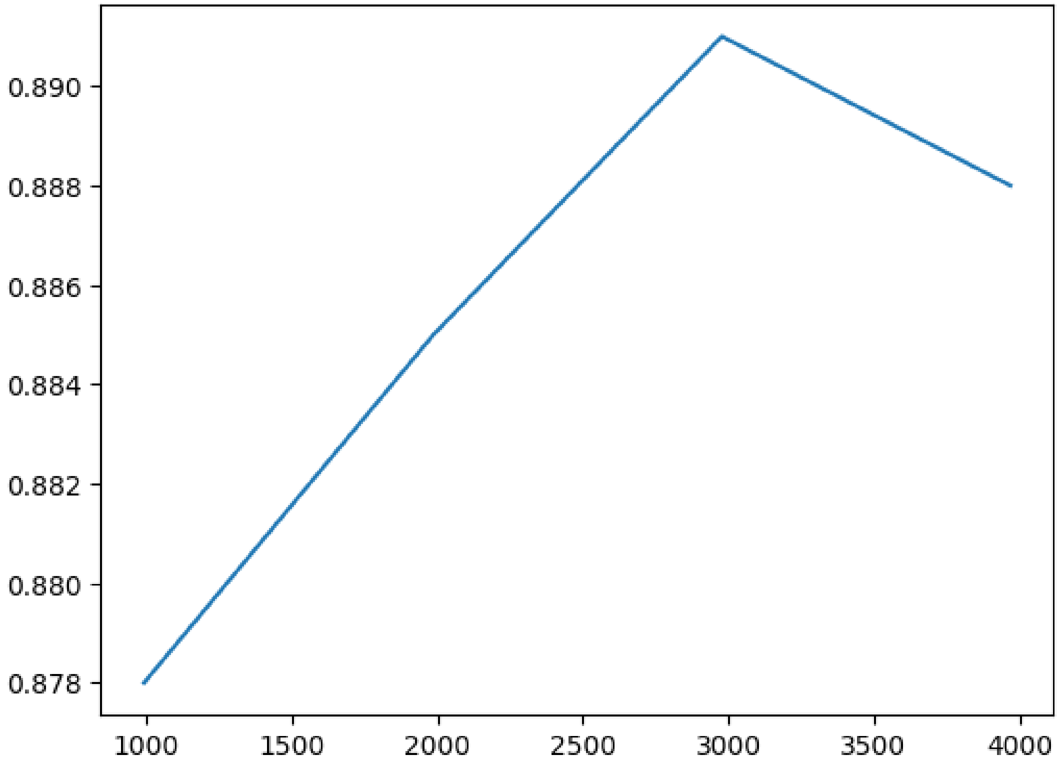
Suicide: Precision- 0.92, Recall- 0.84

Mental Health: Precision-0.81, Recall- 0.85

Other: Precision-0.97, Recall-0.95

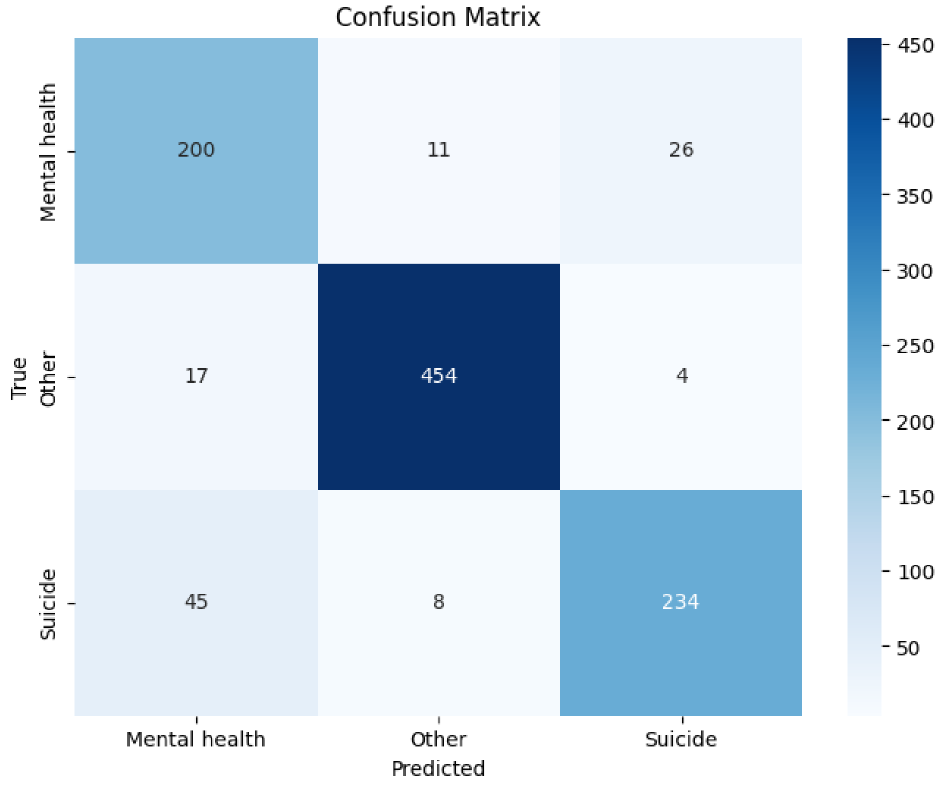
- Cluster 4 and remaining SuicideWatch posts as ’suicide’

Figure 7.3: Accuracy curve during steps for Model 2



Source: Author

Figure 7.4: Confusion matrix for Model 2



Source: Author

Accuracy: 0.88

Suicide: Precision- 0.89, Recall- 0.81

Mental Health: Precision-0.76, Recall- 0.84

Other: Precision-0.95, Recall-0.96

8. Discussion and conclusion

Overall, it can be said that both models performed quite well- both have quite high metrics on all fields. However, the model trained on data in which only cluster 4 was used as the ’suicide’ label performed better than the one which included both cluster 4 and the remaining SuicideWatch posts. This, however, was something that was anticipated, to a degree: in the discussion about the remaining, ’excluded’ SuicideWatch posts in the section about creating the labels, it was found that they are, in fact, quite vague, generic or technical, and as such could confuse the model and make it generalize and include other vague, generic or technical posts that have nothing to do with suicide.

It would seem that this prediction came true, and is reflected both in the lower metrics (Figures 7.2 nd 7.4) and in the chart depicting the progression of the model’s accuracy during training (Figures 7.1 and 7.3)- in the first model, the progression goes upwards as the model becomes better with each iteration and weight adjustment. For the second model, on the other hand, there is unexpected worsening of the model’s accuracy late during training- probably caused by the problems with the labelling system and data used on this particular model that I was reffering to.

At this point, it is useful to reflect on the research questions of this thesis:

1. Can embeddings from Large Language Models help group and cluster corpus contents in a way that helps distinguish between suicidal content and purely mental health content?

Trough the embedding and plotting the locations of posts in 3D space, It became apparent that creating the labels for classification that rely purely on the subreddit of origin is flawed. While the vast majority of posts from Other subreddits were truly well-separated from the ones from either Mental Health or SuicideWatch, the significant overlap between parts of Mental Health and SuicideWatch was evident. This leads to the conclusion that using subreddit of origin as the only basis of labels is inadequate, as the contents of Mental Health and Suicide overlap.

For this reason, an alternative approach that relies on content similarity (measured trough embeddings) was chosen. By using hierarchical clustering with cosine similarity on embedded posts, I was able to cluster content in a way that captured content similarity in a satisfactory way. The cluster analysis showed that clusters can be relatively clearly distinguished from one another: from five clusters, one contains the vast majority of ’Other’ content, one contains most of Mental Health posts, one contains the vast majority of SuicideWatch posts (along with a good chunk of Mental Health, as expected). Two smaller clusters lie between ’Other’ and ’Mental Health’ regions, but these clusters are relatively small compared to the rest of the clusters and are not relevant for the analysis.

By using cluster 4 – the cluster that contained most of SuicideWatch posts, as well as the Mental Health posts similar in content- and doing additional checks and comparing it’s content to the content of other clusters, I have succeeded in creating an improved labelling system compared to what was used in the study that was mentioned earlier.

Because of this, I believe it can be concluded that reliance on document embedding that uses powerful pretrained language models capable of capturing deep structures and relations in texts is of great help in all tasks that require segmenting textual. They provide a remarkably good numeric expression of written content, and in combination with other techniques (such as clustering) they have shown themselves as a powerful solution for a problem in suicide research that relies on online data that would otherwise be very difficult, or very expensive, to solve.

2. Can Large Language Models accurately classify texts based on whether they are merely mental health related or suicidal?

Looking at the performance of the models in classifying the posts based on the labels I created, I think it can be concluded that the models are, overall, very successfull in differentiating between Suicide and Mental Health, and in differentiating both of these from ’Other’ content.

In both models, the ’Other’ class has very high precision and recall, meaning that the models are quite clearly distinguishing between the content that lies outside of the domains of mental health and suicidal ideation, and the one that lies within. This part of data was also easily distinguishable even in the clustering stage, considering that cluster 0 contained most of ’Other’ data and extremely small amounts of data originating from Mental Health or SuicideWatch subreddits, so this result is not surprising.

Since the main goal of this thesis was to see whether the model is capable of distinguishing between content with suicidal ideation, and content that is merely about mental health struggles, it is useful to address the metrics associated with these classes.

The metrics for suicide are overall quite good, especially in the first model. The amount of false positives is especially low, judging by the precision score.

However, in practice, the proportion of false negatives- expressed in recall- could be much more important in this particular case. Since we are discussing suicide detection, and-presumably- prevention, detecting more people who are actually contemplating suicide is more important than if we wrongly detect someone as suicidal. In this regard, the recall for suicide is smaller than precision, but still quite high.

Another thing to note is a relatively high confusion between Suicide and Mental Health classes-out of 269 suicide posts, 40 were missclassified as Mental Health, thus being the largest number of false negatives for Suicide class. While this implies that the major challenge for the model lies in distinguishing between Suicide and Mental Health (which is to be expected), I still think that, overall, it can be said that the model performs quite well in the major task it was designed to do.

In order to improve the model's performance in the future, a few things should be considered:

1. Using a more powerful model: The model which was fine-tuned for this study is Curie model, the second most powerful model available trough the API. However, the most powerful model-Davinci- was not used for pricing reasons. However, using it could have major effect on model's performance.

2. Training dataset size: As is the case with all machine learning models, the larger the data, the more the model can learn. Increasing the dataset size for training could improve performance.

3. Training dataset structure: In the case of my model, the training dataset was not perfectly balanced- the posts originating from r/SuicideWatch were underrepresented. This may have reduced the model's ability to distinguish between Mental Health and Suicide.

4. Lastly, the labelling system should be explored more. While the current system which performs better- using solely cluster 4- works well, it is still unclear if some texts outside of that cluster should be granted special status and included within the 'suicide' label; it is also unclear what would be the best way to assess which texts should get this special status. A combination of additional 'nested' clustering, or manual inspection, could be helpful.

Finally, this research contributes to science in several ways.

As far as I know, at the time of writing this thesis there were no cases where LLMs and transformer models were used for suicide detection. As such, the exploration of their utility and efficiency in this task constitute a contribution to science that can be the basis for further research.

This research shows that the transformer models and LMMs are not only useful for detecting suicidal ideation as such, but actually capable of differentiating between mental health struggles and specifically suicidal content- a task that is made difficult due to the similarity between the two types of posts. The subtle difference between a post that contains discussion about mental health, versus the one containing suicidal ideation, however, is an important one. Creating a model that is able to distinguish between the two with relatively high recall shows that transformer models are capable of very fine discernment, which seems to me as something of great importance for suicide detection research. It is something that can be further explored in the future.

This research also takes innovative steps in creating more adequate suicide labels for online data. By using data that has been confirmed as containing suicidal ideation by experts in previous researches, and clustering other data around these posts based on embeddings, I was able to create a more accurate labeling of data. To my knowledge, this approach to creating labels was not used in previous research, and I think it is an interesting way of approaching the problem, and that other scientific research that deals with processing and classifying text data can have much use from.

9. Summary in Slovene

Naraščajoči pomen spletnega komuniciranja v sodobni družbi je izpostavil potrebo po razumevanju in prepoznavanju samomorilnih misli v teh spletnih prostorih. Spletne skupnosti, še posebej tiste, osredotočene na duševno zdravje, pogosto vsebujejo komunikacije, ki so tesno prepletene z izrazi samomorilnosti. Medtem ko je zaznavanje teh izrazov pomembno za raziskave, je tudi temeljnega pomena za proaktivno moderacijo in preventivne strategije na teh platformah.

Tradicionalne metode strojnega učenja so pokazale obetavne rezultate pri prepoznavanju samomorilnih nagnjenj v tekstovnih podatkih. Vendar pa pojav velikih jezikovnih modelov (kot je GPT-4), zgrajenih na sofisticiranih arhitekturah globokih učenj, ponuja možnost globljega in bolj niansiranega odkrivanja subtilnih namigov, povezanih z samomorilnimi mislimi, ki so pogosto prepleteni z drugimi temami in jih je težko izolirati.

Osrednje vprašanje te raziskave je preučiti sposobnost velikih jezikovnih modelov pri zaznavanju samomorilnih vsebin v spletnem okolju. Glavni raziskovalni problem, ki ga to delo rešuje, je usposabljanje modela, ki je sposoben razlikovati med splošnimi razpravami o duševnem zdravju in resnično samomorilno vsebino. Naloga je zahtevna, saj so razlike med tema vsebinama lahko zelo subtilne.

To je proces v dveh korakih. Najprej je treba objave ustrezno označiti, kar samo po sebi predstavlja velik podproblem te raziskave. Pri delu z veliko količino spletne vsebine (ki je potrebna za usposabljanje takih modelov strojnega učenja) to lahko predstavlja problem, saj bi za ustrezno označitev besedila kot vsebine, ki vsebuje samomorilno vsebino, potrebovali potrditev strokovnjaka ali uporabo druge rešitve. V delu predlagam rešitev, pri kateri uporabljam vgradnjo dokumenta, ki jo zagotavljajo OpenAI-jevi jezikovni modeli, v kombinaciji z vsebino iz podskupine SuicideWatch na Redditu. Ugotovljeno je bilo, da podskupina SuicideWatch, namenjena uporabnikom s samomorilnimi mislimi, vsebuje samomorilno vsebino. To ugotovitev izkoristim tako, da vgradim te objave skupaj z drugimi objavami v korpusu in uporabim hierarhično združevanje, da združim objave glede na njihovo podobnost v vsebini. Če najdem skupino, ki vsebuje veliko večino objav iz podskupine SuicideWatch, sklepam, da je to skupina, ki je značilna za vsebino z samomorilnimi mislimi, in vse objave v njej označim kot "samomor".

Drugi korak je uporaba zdaj označenih podatkov za natančnejše usposabljanje modela za klasifikacijo. Natančnejše usposabljanje je postopek usposabljanja jezikovnega modela, da na podlagi vhodnega poziva proizvede določene izhode. V tem primeru so pozivi različne objave in želeni izhodi so njihove oznake razredov. Natančnejše usposabljanje je postopek usposabljanja modela, katero vrsto pozivov povezati s katero vrsto izhodov.

Zato so raziskovalna vprašanja zastavljena v te magistrskem delu naslednja: 1) Ali lahko vgradnje iz velikih jezikovnih modelov pomagajo združiti in razvrstiti vsebino korpusa na način, ki pomaga razlikovati med samomorilno vsebino in zgolj vsebino o duševnem zdravju, in 2) Ali lahko veliki jezikovni modeli natančno klasificirajo besedila glede na to, ali so zgolj povezana z duševnim zdravjem ali so samomorilna?

Posledično cilji vključujejo 1) vgradnjo besedil in njihovo združevanje na podlagi podobnosti vsebine ter 2) natančnejše usposabljanje modelov za razlikovanje in kategorizacijo dokumentov glede na prisotnost pristnih samomorilnih misli v primerjavi s splošnimi razpravami o duševnem zdravju.

Rezultati potrjujejo učinkovitost velikih jezikovnih modelov pri obeh nalogah, saj dosegajo uspešno združevanje objav na podlagi njihove podobnosti v vsebini za ustvarjanje oznak razredov, pa tudi visoko natančnost in priklic pri razlikovanju samomorilnih misli od splošnih pripovedi o duševnem zdravju.

10. Sources:

1. Amatriain, X. (2023). *Transformer Models: An Introduction and Catalog*. https://www.researchgate.net/publication/368540509_Transformer_models_an_introduction_and_catalog

1. Bachmann, S. (2018). Epidemiology of Suicide and the Psychiatric Perspective. *International Journal of Environmental Research and Public Health, 15(7).* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6068947/pdf/ijerph-15-01425.pdf

3. Berkelmans, G., van der Mei, R. D., Bhulai, S., & Gilissen, R. (2021). Identifying socio-demographic risk factors for suicide using data on an individual level. *BMC Public Health, 21*. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-11743-3

4. Bonaccorso, G.(2017). *Machine learning algorithms textbook*. Packt Publishing. https://balasahebtarle.files.wordpress.com/2020/01/machine-learning-algorithms_text-book.pdf

5. Chatterjee, O., Kumar, P., Samanta, P., & Sarkar, D. (2022). Suicide ideation detection from online social media: A multi-modal feature based technique. *International Journal of Information Management Data Insights, 2(2), 100103.* https://www.sciencedirect.com/science/article/pii/S2667096822000465

6. Chapman, W.(2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association.* https://www.researchgate.net/publication/51576224_Natural_language_processing_An_introduction

7. Choudhury, M., Kıcıman, E., Dredze, M., Coppersmith, G. A., & Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5659860/pdf/nihms909062.pdf

8. Durkheim, É. (1897). *Le suicide: Étude de sociologie*. Paris: Félix Alcan.

9. Fatima, A., Ying, L., Hills, T., & Stella, M. (2021). DASentimental: Detecting depression, anxiety and stress in texts via emotional recall, cognitive networks and machine learning. *Big Data and Cognitive Computing, 5(4).*

https://www.researchgate.net/publication/357018212_DASentimental_Detecting_Depression_Anxiety_and_Stress_in_Texts_via_Emotional_Recall_Cognitive_Networks_and_Machine_Learning

10. Feldhege, J., Wolf, M., Moessner, M., & Bauer, S. (2022). Psycholinguistic changes in the communication of adolescent users in a suicidal ideation online community during the COVID-19 pandemic. *European Child & Adolescent Psychiatry*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9415261/

11. Haque, R., Islam, N., Islam, M., & Ahsan, M. (2022). A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. *Technologies*. https://www.mdpi.com/2227-7080/10/3/57

12. Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2019). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. *IEEE Transactions on Computational Social Systems*. https://www.researchgate.net/publication/345024088_Suicidal_Ideation_Detection_A_Review_of_Machine_Learning_Methods_and_Applications

13. Jurafsky, D. & Martin, J.H. (2023). Transformers and Pretrained Language Models. *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/10.pdf

14. Joseph, S.R., Hlomani, H., Letsholo, K., Kaniwa, F., Sedimo, K. (2016). Natural Language Processing: A Review. *International Journal of Research in Engineering and Applied Sciences*. https://www.researchgate.net/profile/Sethunya-Joseph/publication/309210149_Natural_Language_Processing_A_Review/links/5805ea1f08ae03256b75d965/Natural-Language-Processing-A-Review.pdf

15. Khan, A., Shimul, S., & Arendse, N. (2021). Suicidal behaviour and the coronavirus (COVID-19) pandemic: Insights from Durkheim's sociology of suicide. *International Social Science Journal*. https://onlinelibrary.wiley.com/doi/10.1111/issj.12269

16. Kowsher, M., As Sami, A., Prottasha, N. J., Arefin, M., Dhar, P. K., & Koshiba, T. (2016). Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding. *IEEE Access*. https://www.researchgate.net/publication/362574897_Bangla-BERT_Transformer-based_Efficient_Model_for_Transfer_Learning_and_Language_Understanding

17. Kinsley, H., & Kukiela, D. (2020). *Neural Networks from Scratch in Python*. https://www.haio.ir/app/uploads/2021/12/Neural-Networks-from-Scratch-in-Python-by-Harrison-Kinsley-Daniel-Kukiela-z-lib.org_.pdf

18. Kumar, V., & Garg, M.L. (2018). Deep Learning as a Frontier of Machine Learning. *International Journal of Computer Applications*. https://www.researchgate.net/publication/326429676_Deep_Learning_as_a_Frontier_of_Machine_Learning_A_Review

19. Luxton, D.D., June, J.D.,Fairall, J.M.(2012). Social Media and Suicide: A Public Health Perspective. *American Journal of Public Health.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3477910/pdf/AJPH.2011.300608.pdf

20. Nathan, N., & Nathan, K. (2020). Suicide, Stigma, and Utilizing Social Media Platforms to Gauge Public Perceptions. *Frontiers in Psychiatry*. https://www.frontiersin.org/articles/10.3389/fpsyt.2019.00947/full

21. OpenAI (n.d.-a) *Fine tuning- legacy*. https://platform.openai.com/docs/guides/legacy-fine-tuning

22. OpenAI (n.d.-b) *Embeddings*: https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

23. Paul, M. J., & Dredze, M. (2017). *Social Monitoring for Public Health*. https://www.cs.jhu.edu/~mdredze/publications/2017_social_monitoring_preprint.pdf

24. Perry, A., Lamont-Mills, A., du Plessis, C., du Preez, J., & Pyle, D. (2020). Suicidal behaviours and moderator support in online health communities: Protocol for a scoping review. *BMJ Open*. https://bmjopen.bmj.com/content/bmjopen/10/1/e034162.full.pdf

25. Pilehvar, M.T., Camacho-Collados,J. (2021) *Embeddings in Natural Language Processing – Theory and advancements in vector representation of meaning*. Morgan and Claypool publishers. http://josecamachocollados.com/book_embNLP_draft.pdf

26. Praveen, P., Ranjith kumar, M., Shaik, M., Ravikumar, R., & Kiran, R. (2020). The comparative study on agglomerative hierarchical clustering using numerical data. *IOP Conference Series: Materials Science and Engineering, 981(2).* https://iopscience.iop.org/article/10.1088/1757-899X/981/2/022071/pdf

27. Prihatini, P., Putra, I., Giriantari, I., & Sudarma, M. (2019). Complete agglomerative hierarchy document’s clustering based on fuzzy luhn’s gibbs latent dirichlet allocation. *International Journal of Electrical and Computer Engineering (IJECE), 9(3), 2103-2111*. https://www.researchgate.net/publication/333538281_Complete_agglomerative_hierarchy_document's_clustering_based_on_fuzzy_luhn's_gibbs_latent_dirichlet_allocation

28. Rahaman, M. N., Chaki, S., Biswas, M. S., Biswas, M., Ahmed, S., Mahi, M. J. N., & Faruqui, N. (2022). Identifying the Signature of Suicidality: A Machine Learning Approach. *The International Conference on Emerging Trends in Artificial Intelligence and Smart Systems, THEETAS 2022*. https://www.researchgate.net/publication/361193793_Identifying_the_Signature_of_Suicidality_A_Machine_Learning_Approach

29. Robertson, R. A., Standley, C. J., Gunn, J. F. III, & Opara, I. (2022). Structural indicators of suicide: An exploration of state-level risk factors among Black and White people in the United States, 2015-2019. *J Public Ment Health, 21(1), 23-34*. Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9524014/

30. Roškar, S., Paska, A.V.(2021). *Samomor v Sloveniji in Svetu*. https://www.zadusevnozdravje.si/wp-content/uploads/2022/06/Monografija-Samomor-v-Sloveniji-in-po-svetu.pdf

31. Ruch, D.A., Bridge, J.A. (2022). Epidemiology of Suicide and Suicidal Behavior in Youth. In: Ackerman, J.P., Horowitz, L.M. *Youth Suicide Prevention and Intervention*. https://link.springer.com/chapter/10.1007/978-3-031-06127-1_1

32. Santoso, M. A., Susanto, B., & Virginia, G. (2018). The Application of Agglomerative Clustering in Customer Credit Receipt of Fashion and Shoe Retail. *International Journal of Industrial Research and Applied Engineering 3(1)* . https://www.researchgate.net/publication/324891426_The_Application_of_Agglomerative_Clustering_in_Customer_Credit_Receipt_of_Fashion_and_Shoe_Retail

33. Shalev-Shwartz, S., Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

34. Suler, John. (2004). The Online Disinhibition Effect. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society, 7.* https://www.researchgate.net/publication/8451443_The_Online_Disinhibition_Effect

35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.(2017) Attention is all you Need. *Conference on Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

36. Wang, L., Mistry, S., Abdulahi Hasan, A., Omar Hassan, A., Islam, Y., & Junior Osei, F. A. (2023). Implementation of a Collaborative Recommendation System Based on Multi-Clustering. *Mathematics, 11(6).* https://www.researchgate.net/publication/369118938_Implementation_of_a_Collaborative_Recommendation_System_Based_on_Multi-Clustering

37. Webster, F. (1995) *Theories of the Information Society, Third Edition*. https://cryptome.org/2013/01/aaron-swartz/Information-Society-Theories.pdf

38. World Health Organization. (2019) *Suicide worldwide in 2019 Global Health Estimates*. https://iris.who.int/bitstream/handle/10665/341728/9789240026643-eng.pdf?sequence=1

11. Appendices:

Appendix A: Google Collaboratory notebook containing the full Python code used for this thesis

https://colab.research.google.com/drive/1YvRwdWFFdbsSu6eD_sJB8ZWCmlQ75l4v?usp=sharing