Arab American University

Faculty of Graduate Studies

Re-Categorizing the Attention Deficit Hyperactivity Disorder SNAP-IV Diagnostic Tool into Concisely Semantic Groups by the use of Natural Language Processing

By

Zaina Jamal Ziyad Saadeddin

Main Supervisor

Dr. Mohammed A. M. Maree

Co-Supervisor

Dr. Mohammad M. Herzallah

This thesis was submitted

in partial fulfillment of the requirements for

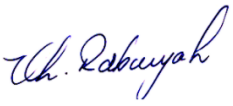
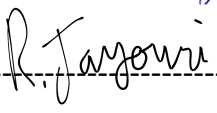
the master's degree in Computer Science

Re-Categorizing the Attention Deficit Hyperactivity Disorder SNAP-IV Diagnostic Tool into Concisely Semantic Groups by the use of Natural Language Processing.

By

Zaina Jamal Ziyad Saadeddin

This Thesis was defended successfully on September 11, 2021 and approved by:

Committee Members	Signature
1. Supervisor Name: Dr. Mohammmed A. M. Maree	----------  ----------
2. Co-Supervisor Name: Dr. Mohammad M. Herzallah	----------  Digitally signed by Mohammad Herzallah Date: 2021.11.29 10:59:03 -05'00' ----------
3. Internal Examiner Name: Dr. Khalid Rabaya	---------  ----------
4. External Examiner Name: Dr. Rashid Jayousi	----------  ----------

Acknowledgments

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisors, Dr. Mohammed Maree and Dr. Mohammad Herzallah, whose experiences were invaluable in formulating the research questions and methodology. The provided insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to thank Dr. Joman Natsheh, ADHD expert from Palestinian Neuroscience Initiative at Al-Quds University for her contribution in providing the real word ADHD dataset and her evaluation judgment throughout the research results. I want to thank you Dr. Joman for your support and for all of the opportunities given to further my research.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my students in the Code for Palestine program, Batool Alsulaibi and Basil Hijaz for their support provided during the classes as teacher assistants to help me during the hard time faced throughout the research.

Thanks for all your patience and encouragement

Declaration

This is to declare that the thesis entitled “Re-Categorizing the Attention Deficit Hyperactivity Disorder SNAP-IV Diagnostic Tool into Concisely Semantic Groups by the use of Natural Language Processing” under the supervision of Dr. Mohammed A. M. Maree and Dr. Mohammad M. Herzallah is my own work and does not contain any unacknowledged work or material previously published or written by another person, except where due reference is made in the text of the document.

Date: August 20, 2021

Name	Signature
Zaina Jamal Ziyad Saadeddin	

Abstract

The diagnostic for the attention deficit hyperactivity (ADHD) disorder based on a symptoms rating questionnaire conducted using interviews is clinically recommended as a diagnostic measure, the SNAP-IV rating scale is a version from Swanson, Nolan and Pelham (SNAP) to measure the core symptoms of ADHD. where it requires answering 90 questions, there is significant semantic overlap across questions, the current questions categories are based on semantically subjective and correlated measures rather than well-defined mathematically independent clusters, the accuracy of non-structured approaches is very limited. The purpose of this study is to improve the utility of SNAP-IV by deriving the hidden semantic dimensions that are encoded in the original questions by understanding the semantic basis based on medical knowledge resources to: i) reduce the length by decreasing questions, ii) Increase the accuracy by grouping questions that express the same medical concept, iii) Enabling healthcare professionals to decide retaining or removing overlapped questions. We conducted several clustering models. Four different prototypes built based on the produced categories from each one of them and been used in a real world completed and scored dataset for ADHD diagnostic. We ran statistical independent-sample t-tests to compare the averages between children with ADHD and matched controls. Results show that all new produced categories from each clustering approach shows significant difference. The semantically-enhanced medical enrichment terms which is based on a medical knowledge graph has proved to produce a new concise concept based categorization with less number of questions without harming the clinical procedure.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADHD – Attention Deficit Hyperactivity Disorder

ADD – Attention Deficit Disorder

BoW – Bag-of-Words

BERT – Bidirectional Encoder Representations from Transformers

BioBERT – Biomedical Bidirectional Encoder Representations from Transformers

cTAKES – Clinical Text Analysis and Knowledge Extraction System

CSS – Cascade Style Sheet

DSM – Diagnostic and Statistical Manual of Mental Disorders

HTML – Hyperlink Text Markup Language

HL7 – Health Level Seven

ICD-CM – International Classification of Diseases-Clinical Modification

LDA – Latent Dirichlet Allocation

MeSH – Medical Subject Headings

MTH – Medical Treatment History

NLP – Natural Language Processing

NCI – National Cancer Institute

ODD – Oppositional Defiant Disorder

SNAP – Swanson, Nolan, and Pelham Rating Scale

SNOMED-CT– Systematized Nomenclature of Medicine-Clinical Terms

Tf–Idf – term frequency–inverse document frequency

UMLS – Unified Medical Language System

Chapter 1 - Introduction

1.1 Background

Attention deficit hyperactivity disorder (ADHD) is a mental disorder with multiple symptoms that include increased hyperactive and impulsive behaviors. People with ADHD can be children or adults suffering from the inability to focus their attention on a single task for an extended period (Weiss et al., 2003). They are often forgetful about completing tasks, easily distracted, feel difficulty in sitting still, and continuously interrupt people while they are talking. There are three main types of ADHD, (i) predominantly inattentive, (ii) predominantly hyperactive-impulsive, or (iii) a combination of both types. As suggested by the name, people with predominantly inattentive ADHD face extreme difficulties in focusing and finishing the assigned tasks. Besides, they face problems in following instructions given to them. This type of ADHD cannot be diagnosed easily because patients do not tend to disturb people around them and are considered introverted and careless by nature. As such patients may suffer from hyperactive and impulsive disorders, including fidgeting, emotional difficulties and extreme changes in mood (Weiss et al., 2003).

To diagnose ADHD, healthcare specialists use tools (in the form of questionnaires) to determine whether a subject has ADHD or not. Generally, a rating scale is used where parents and teachers can answer questions about the subject's behavior in social and schoolwork (Faries et al., 2001). Questionnaire results are interpreted to assess the diagnosis of ADHD symptoms and other conditions related to ADHD. These tools suffer from several limitations such as incorrectly

diagnosing the disorder, the need for using multiple evaluations with repetitions which lead to wrong diagnoses and treatments, the inability to tackle the semantic correlations among ADHD and other related diseases that might be tightly coupled with this disorder, the lengthy and time-consuming procedure that not consistently followed by healthcare professionals worldwide. This has raised concerns over the need for a concise improved questionnaire where questions are grouped into well-defined mathematically independent clusters to remove the overlap of questions to avoid scenarios of having an incorrect diagnosis, untreated, overdosed, or underdosed children. To address these concerns, a combination of natural language processing (NLP) techniques, medical information extraction and retrieval, and the utilization of extrinsic medical ontologies (which provide formal and explicit specifications of shared medical conceptualizations (Díaz-Galiano et al., 2008) is proposed. The goal in this context is to increase the efficiency and effectiveness of ADHD diagnostic, identify semantically-related medical correspondences that are captured in medical semantic resources, and consequently re-categorize and group ADHD questions into a more concise and semantically overt structure. Over the past few years, several semantic resources and classification systems have been developed to address various challenges in the medical information processing domain (Reátegui & Ratté, 2018). Examples of medical resources are: (i) MetaMap[1], which is utilized to map biomedical texts into the (ii) Unified Medical Language System (UMLS[2]) which is a large-scale biomedical thesaurus that provides explicit specifications of biomedical knowledge (Bodenreider, 2004), (iii) cTAKES[3] which extracts clinical information from text, (iv) ScispaCy4 which identifies medical information based on

[1] https://metamap.nlm.nih.gov/
[2] https://www.nlm.nih.gov/research/umls/index.html
[3] https://ctakes.apache.org/
[4] https://allenai.github.io/ScispaCy/

predefined NLP models, and the Medical Subject Headings (MeSH5) metathesaurus which comprises a lexicon of medical terms associated with multiple types of lexical semantic relations (Koopman et al., 2016).

With the prevalence of medical knowledge resources, as well as the recent advancement in NLP techniques, we proposed a new ADHD diagnostic tool that aims at assisting healthcare professionals in (i) discovering repeated patterns among ADHD questionnaires, (ii) effectively minimize the lengthy ADHD diagnosing procedure, (iii) categorize ADHD questions based on their semantic relatedness, and (iv) unify the ADHD diagnostic procedure through the utilization of a commonly-used, trust-worthy and accredited tool. To do this, we integrated multiple medical semantic resources with a unique NLP pipeline and evaluated the effectiveness of the proposed system using a real-world ADHD dataset that is provided by the Palestine Neuroscience Initiative (PNI[6]).

The remainder of this chapter is organized as follows. The motivations behind the proposed system are presented in Section 1.2. In Section 1.3, we provide a preliminary review of the literature and identify the main aspects that characterize our proposed system. Section 1.4 provides the research methodology and steps. Section 1.5 defines our contributions and highlights the obstacles that we attempt to overcome through the proposed system. Finally, the organization of the remainder of the thesis is introduced in Section 1.6.

[5] https://www.nlm.nih.gov/mesh/meshhome.html
[6] https://neuroscience.alquds.edu/

1.2 Motivations

The main motivation behind this research project is the integration between various strands of computer science focused on Natural Language Processing and informatics applied to an available ADHD clinical dataset at the Palestinian Neuroscience Initiative (PNI) at Al-Quds University to solve a key clinical problem and ultimately for people who suffer from ADHD.

1.3 Problem Statement and Research Questions

The diagnosis of ADHD is still carried out based on manual measures, and patient interviews. One of these measures is the SNAP-IV scoring tool (Hall et al., 2020), the SNAP entails answering ninety questions. There is significant semantic overlap across those questions. In addition, there is a semantic gap between the purpose of the questions and the underlying meaning of the provided answers by the patients. The current question categories are based on semantically subjective and weakly correlated measures rather than well-defined mathematically independent clusters. Here, we proposed a query-based symptom assessment coupled with extrinsic medical semantic knowledge resources (Abdi et al., 2017), (Huang et al., 2013) to reduce the lengthy diagnostic procedure, increase questions correlatednes's accuracy by concisely grouping semantically-relevant questions, furthermore to be integrated with a cognitive dataset to feed the neurocomputational modeling to reduce the uncertainty of decisions related to treatment dosage for ADHD. This helps in estimating the best dosage of dopaminergic medications as determining

an accurate dosage after the diagnosis of ADHD in children is still a difficult task as reported by (Ahmadi et al., 2021) and (Colby et al., 2012), and can further affect their activity and movement patterns (Nigg et al., 2012). Recently, statistics about children suffering from this disorder have raised concerns over the need for the right treatment (Kim et al., 2015), (Nylander et al., 2021), leaving much to be done to avoid more untreated, overdosed, or underdosed children. In an attempt to tackle these issues, we present the research questions that we have to examine and address during our research work:

Q1. What are the strengths and weaknesses of the methods and techniques that are employed by current medical resources?

To answer this question, we have conducted a comprehensive analysis of current systems and proposed approaches and studied the features of the implemented methods /techniques for each system.

Q2. How SNAP-IV questions are organized and what alternative grouping of the questions can be proposed by the utilization of NLP pipelines?

To answer this question, we have proposed multiple question reformulation and categorization modules based on a combination of NLP pipelines and medical knowledge resources. Namely, we proposed four different NLP pipelines wherein each pipeline produced a different number of questions and grouped the selected questions into categories that are based on the backbone of each pipeline. We evaluated the resulting categories of each module and compared them against the original SNAP-IV which is used as a baseline in our case. More details on how we developed the various components of each of the proposed pipelines and their quality are presented in Chapters three, four, and five of this thesis.

1.4 Research Methodology and Steps

The following points outline the main steps that we have carried out during our research work:

- **The Construction of an NLP-based Pipeline**

In the preprocessing phase, we have preprocessed SNAP-IV questions through several NLP pipeline phases, including text cleaning, term normalization, stop words removal (based on a predefined list such as a, the, an... etc.), and term lemmatization. During these phases, we tokenized question texts into n-gram tokens through the exploitation of medical knowledge-based resources, namely cTAKES, MetaMap, ScispaCy, and MeSH ontology. Then during this phase, we further enriched the recognized n-gram token with a set of semantically related terms. To do this, we submitted each n-gram token to the UMLS metathesaurus to find whether the token is recognized by the metathesaurus and if so, we extract terms that are semantically related (parents/meronyms) of each term to expand the original questions based on the UMLS semantic network (Schuyler et al., 1993). More details on this pipeline are provided in Chapter 4.

- **Feature Extraction and Representation**

We have used different word representation techniques to represent question texts into word vectors, where each word mentioned in a vector reflects its weight. First, we have used context-free conventional techniques such as the Bag-of-Words (BoW) model and the term frequency-inverse document frequency *tf.idf* weighting scheme (Aizawa, 2003). In addition, we have used techniques that represent the features based on word contextual similarities, such as the Word2Vec model (Mikolov et al., 2013) where similar contextually co-existing words are mapped to their

corresponding weights and represented in a vector space. Second, we have used several pre-trained models, such as BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2019) which utilize a deep network of transformer layers and a new masked language model to assign weights to features extracted from the set of ADHD questions. Third, we constructed semantic networks (in the form of knowledge graphs) which represent the medical terms and their semantically related correspondences in an attempt to conduct similarity measures using the resulting graphs. More details on each of the used feature extraction and representation techniques are presented in Chapters 3 and 4.

- **Term Re-weighting and Identification of Overlapping Questions**

As we discussed earlier in the motivation section, there are several overlapping questions in the SNAP-IV. The reduction of the repeated questions, especially when they semantically convey the same idea becomes more crucial. To address this problem, we conducted a term re-weighting technique to identify question terms that express the same medical concept and re-assigned their weights to be further considered in the question classification task. In this context, terms that pertain to questions that express the same medical concept (more specifically, an ADHD-related concept) are grouped under the same class, enabling healthcare professionals to visually depict such similarities and accordingly decide on retaining/removing overlapping questions. Indeed, after the implementation of the proposed system's prototype, it was clear for healthcare experts that some overlapping questions can be grouped under the same category and even be minimized to a smaller number of questions without harming the quality of the ADHD diagnostic procedure. More details on this process are provided in Chapters 4 and 5.

- **Topic Modeling and ADHD Questions Clustering Techniques**

To cluster the SNAP-IV questions into more concise and semantically related groups we have conducted four different unsupervised machine learning models. First, we have developed our medical knowledge graph based clustering technique in the same manner as proposed by (Pham et al., 2019), where we have constructed a knowledge graph (using the NetworkX[7] python library) that links n-gram tokens both extracted from question texts and the medical knowledge resources using the parenty/meronymy relationships In this context, each sub graph represents a group of questions with a shared parent concept in a semantics-based hierarchical structure. More details on the implementation for each model are provided in Chapter 4. (For more details on knowledge graph visualization using NetworkX, we refer the reader to Appendix D). Second, we employed LDA (Blei et al., 2003) topic modeling technique using the Gensim[8] library. The goal of this step is to identify the various topics under which each question term can be classified. Then, we utilized two different models, namely K-means (Kalra et al., 2018) and Gaussian Mixture (Reynolds, 2008) clustering techniques using scikit-learn[9] machine learning library. The goal of utilizing both techniques is to group similar questions under different categories, without the help of any knowledge resource. Our goal here was to demonstrate how the exploitation of medical knowledge resources can assist in improving the quality of such clustering approaches.

- **Evaluation of the Proposed Techniques**

To evaluate the quality of the proposed models, we have carried out a series of experiments using the SNAP-IV questions as our reference baseline. The NLP pipeline phases and their associated

[7] https://networkx.org/
[8] https://radimrehurek.com/gensim/index.html
[9] https://scikit-learn.org/

question classification techniques were implemented gradually in order to explore the impact of each addition/update on the full pipeline phases. In this context, we have utilized multiple term weighting and vector representation techniques starting from the very basic ones, which included conventional NLP preprocessing steps and moving forward towards the more complex term representation models, such as Word2Vec and BERT. As far as the clustering techniques are concerned, we have employed the following clustering models:

- Text clustering without the utilization of any term re-weighting technique.
- Text clustering with the help of term re-weighting techniques.
- Text clustering using semantically-enhanced medical terms that we obtained from the exploited medical knowledge resources.
- Text clustering based on medical knowledge graph based weighting technique.

- **Results Evaluation**

The clustering results of the different techniques are evaluated and compared using several metrics including, the run time of each model, degree of purity (accuracy), precision, recall and F-Score, Confusion matrix. We visualized the clustering results for assisting healthcare professionals in visualizing areas of overlap and to assist them in validating the reduction of the repeated questions, networkX visualization provided in Appendix D and t-SNE visualization provided in Appendix E. Using each of the developed clustering models, we applied question re-categorization with reference to the baseline real-world dataset which comprises ADHD and healthy control (HC) subjects and their scoring results. We conduct a comparative analysis among the developed models for the scores obtained manually in the dataset which were all based on the manual ADHD

diagnostic procedure and using the SNAP-IV full questions dataset. In addition to the comparison, we developed a Web-based application using Google Sites[10] where we designed a Webpage for visualizing the new categorization of each clustering model In addition to visualization dashboard. Our aim in this context was to assist healthcare professionals visually explore the various categorization patterns that resulted from each model, and to evaluate them in comparison with each other. In Chapter 5, we provide more details on the developed application and a link for the live version of the newly proposed ADHD diagnostic GUIs.

1.5 Contributions

The proposed clustering models have introduced a number of advantages to the current ADHD diagnostic procedure which is normally implemented either manually or with the help of SNAP-IV. First, they produced a new concise categorization of the questions that are involved in diagnosing ADHD subjects. The newly proposed question clusters can be mainly characterized by their conciseness, as well as their strong semantic co-relatedness. This was confirmed with the help of healthcare professionals who conduct ADHD diagnostic procedures using conventional methods and tools. The results from text clustering using the semantically-enhanced medical terms re-weighting technique which is based on medical knowledge graphs has proved to produce a new categorization of questions that are close in their semantic distance on the one hand, and have less overlap or redundant questions on the other.

[10] https://sites.google.com/

1.6 Structure of the Thesis

The rest of this thesis is organized as follows, Chapter 2 provides a comparative analysis between current ADHD diagnostic approaches and discusses related studies starting from conventional to semantically-enhanced ADHD diagnoses. A general overview of the system architecture of our proposed system is presented in Chapter 3. We introduce a detailed implementation of the methods that we have utilized in the proposed system's prototype in Chapter 4. Chapter 5 presents the conducted experiments with a discussion on the obtained results, in addition to the evaluation of the effectiveness of the proposed models. Finally, in Chapter 6, we discuss the conclusions and outline the future extensions of our research work.

Chapter 2 - Literature Review

2.1 Background

ADHD is among the most common neurodevelopmental disorders of childhood (Doernberg & Hollander, 2016). It is usually first diagnosed in childhood and often lasts into adulthood. Children with ADHD may have trouble paying attention, controlling impulsive behaviors (acting without considering results of actions), or be overly active. It is normal for a child to have difficulty in focusing and behaving at one time or the other. However, a child with ADHD cannot just grow out of the symptoms. These can continue, can become severe, and can cause trouble at home or school. A child with ADHD might daydream a lot, forget or lose things a lot, squirm or fidget, talk too much, make careless mistakes or take unnecessary risks, have a hard time resisting the temptation, have trouble taking turns, or have difficulty getting along with other children. There are three different types of ADHD depending on which symptom has the highest severity and predominance in a child (Weiss et al., 2003). These are:

1. Inattentive.

   It is hard for the individual to organize or finish a task, to pay attention to details, or to follow instructions or conversations. The child is easily distracted or forgets details of daily routines.

2. Hyperactive-impulsive.

   The child fidgets and talks a lot. It is hard to sit still for long (e.g., for a meal or while doing homework). Smaller children may run, jump or climb constantly. The child feels restless

and has trouble with impulsivity. An impulsive child may interrupt others a lot, grab things from people, or speak at inappropriate times. It is hard for the child to wait their turn or listen to directions. A child with impulsiveness may have more accidents and injuries than others.

3. Combined.

   Symptoms of the above two types are equally present in the child. Since symptoms are supposed to change over time; the presentation may change as well, and is not mandatory for the person to belong to only one category of the aforementioned symptoms.

2.2 Factors that Contribute to ADHD Development in Children

Many types of research have been conducted and scientists are still in the course of detecting the actual cause of ADHD among children and adults (Killeen et al., 2012), (Núñez-Jaramillo et al., 2021). Below are several factors in the child that might be taken as contributors to the development of ADHD.

- Genes.
- Cigarette smoking, alcohol use, or drug use during pregnancy.
- Exposure to environmental toxins during pregnancy (or at a young age), such as lead.
- Low birth weight.
- Brain injuries.

2.3 Existing Diagnostic Systems

The medical diagnosis for ADHD among adults as well as children is symptom-based. The people supposed to have the disorder are recommended to go through a symptom questionnaire where their behavioral attributes are noted by the medical practitioner. Based on their answers to the

questions, they are diagnosed to have ADHD or not. The questions carry respective weights based on their importance. The overall score of the individual for ADHD is calculated using a set of formulas, and the diagnosis could be made.

2.3.1 Manual ADHD Diagnosis

The DSM method describes two techniques for the calculation of scores for the individual. Technique 1 focuses on the total calculation of the score irrespective of which category it belongs to or not. Technique 2 calculates the score category-wise, it calculates the scores for each category and then compares with their respective threshold values. This method is more efficient than the previous one as the categories with the same weights are treated as a group and can be fairly compared with their respective threshold values.

In the starting phase, DSM III (American Psychiatric Association, 1980) was introduced, in which the questionnaire criteria were based on 16 symptoms listed in 3 domains -- 5 inattention (IN), 6 impulsivity (IMP), and 5 hyperactivity (HY) symptoms. These symptoms and domains were obtained from clinical observations. To measure if there is a significant difference between (IN) symptoms and (HY) symptoms or there are a combined 10 symptoms, the same number of items provided for these 2 domains.

After the release of DSM-IV, Swanson, Nolan, and Pelham's (SNAP) rating scale was reorganized (Swanson et al., 2001).

The term "often" was used for each DSM item from the symptoms list in which a psychopathology evaluation was needed based on the degree of presence which is unstable and inconsistent with the development level.

2.3.2 Versions of SNAP

The SNAP questionnaire has gone through different development phases, where the categorization of the questions and their representations were improved. SNAP was developed for use with DSM-III defined ADHD and has been updated with each DSM revision (Swanson et al., 2012). The longer version includes subscales along with items selected from other scales measuring ADHD and associated features, including Corners Index Questionnaire, the IOWA Corners Questionnaire, and the Swanson, Kotkin, Agler, M-Flynn, and Pelham rating scale.

- **SNAP III**

Soon after the release of DSM-III, the Swanson, Nolan, and Pelham (SNAP) rating scale was developed. The SNAP-III was designed to evaluate response to treatment with stimulant medication. It used a 4-point rating scale with the following weights (Not at All = 0, Just a Little = 1, Pretty Much = 2, and Very Much = 3). There were no details regarding the assessment in classroom observations. The variance and rating measurement for norms like age was not included in this version. As such, an improvement with further assessment details were needed for its validation as a diagnostic tool.

- **SNAP IV**

SNAP-IV consisted of the 18 items for DSM-IV-Work-in-Progress category of ADHD, separated into sets of 9 symptoms for the domains of Inattention (INN, to distinguish the DSM-IV set of symptoms from the DSM-III set labeled IN) and Hyperactivity/Impulsivity (HY/IMP). Also, the 14 items not included in DSM-IV but previously included in the symptom lists for Attention

Deficit Disorder (ADD) and ADHD in DSM-III and ADHD in DSM-III-R were retained. Also, 12 items for Oppositional Defiant Disorder (ODD) from the DSM III, III-R, and IV criteria were included (Bussing et al., 2008).

The SNAP-IV method of scoring is an advanced version of the DSM-IV method. The main difference lies in the categorization of questions under consideration for both techniques. The logical procedure remains the same: to compare the total score calculated by the threshold value of the category.

Research indicates that SNAP-IV parent rating scale for children is a good predictors of research diagnosis in both inattention and hyperactivity. Furthermore, the SNAP-IV teacher rating scale is a very good assessment of both clinical and research diagnosis in hyperactivity (Hall et al. 2020). This accentuates the validity of the tool for use in clinical settings. Studies have also been conducted to evaluate the impact of demographics, such as gender and socioeconomic status on responses provided on SNAP-IV (Bussing et al, 2008). Furthermore, research indicates contradiction in the factor structure of this tool. An ADHD and Normal behavior rating scale called SWAN was later developed to address the deficiencies of SNAP (Swanson et al. 2012).

2.4 From Conventional to Semantically-enhanced ADHD Diagnosis

In an early study by Zhang et al., the authors conducted a comparative analysis of how different semantic similarity measures of terms including path-based similarity measure, information content-based similarity and feature based similarity affect document clustering (Zhang et al., 2008). Furthermore, they evaluated the integration between a domain ontology and term re-weighting to the clustering process. The K-means clustering applied on PubMed generated corpus

shows that term re-weighting increases the clustering performance when documents include verbose sentences. In a similar research work that was proposed in (Logeswari & Premalatha, 2013), the authors proposed concept-based clustering technique, where term weights are calculated dynamically based on their hierarchical relationships. The proposed technique first converts the documents into bags of concepts. Then, the K-means clustering algorithm is employed where it shows that concept-based clustering performs better performance than clustering based on conventional techniques.

In another early study by Tar & Nyaunt (Tar & Nyaunt, 2011), the authors proposed a system that presents concept weighting for text clustering that was developed based on the K-means algorithm with the utilization of ontology. After document preprocessing, weights for the words are calculated based on four factors: word length, term frequency, and weights of terms recognition by the exploited ontology. An improvement on precision and recall measures showed that the proposed conceptual representation of word mentions in documents improves the clustering.
Saiyad et al., (Saiyad et al., 2016) provided a survey of 23 papers in traditional and semantic clustering techniques. The authors address the problem of semantic relationships that may appear among the terms of documents implicitly, such as synonyms and polysemous terms, term ambiguity, term high dimensionality, etc. There are several ways to solve this problem that takes place due to the use of traditional approaches. Different ways to solve the problem include the use of Latent Semantic Indexing (LSI), Lexical Chains, and Ontology. The authors find that combination of approaches ontology and lexical chain provides better results for semantic document clustering. Proposed lexical feature selection based on some heuristics and filter out

the non-nouns, in addition to addressing the use of RapidMinor[11] Tool for text processing, Web Mining, and WordNet[12] dictionary facilitate performing various document clustering operations.

In a recent study by Rashid et al. (Rashid et al., 2019), the authors proposed a topic modeling technique using multiple biomedical datasets and fuzzy K-means clustering. The feature extraction was performed based on local and global term weights as proposed in (Salton & Buckley, 1988). The proposed model showed an improved clustering performance than its counterparts such as the FLSA, LDA and LSI baseline models.

To our knowledge, the work on semantics-based clustering of ADHD questions is new, and the idea of incorporating medical semantic knowledge resources for the purpose of semantically-enhancing the current categorization of ADHD questions has not been addressed in the literature before. Therefore, we aim to investigate the impact of utilizing knowledge captured by medical knowledge resources in this context, in an attempt to re-organize ADHD questions in a more concise and semantically-related structure. Our goal in this context is to assist healthcare professionals in their diagnostic procedures using a new rigorous ADHD grouping of questions on the one hand, and also helps them in finding patterns among similar questions and their semantic co-relatedness.

In Table 1, we provide a summary of the discussed approaches and models and highlight the main features that characterize each approach.

[11] https://rapidminer.com/

[12] https://wordnet.princeton.edu/

Table 1: Semantically-enhanced Clustering Approaches

	System	Category	Implementation Techniques
1	A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering (Zhang et al., 2007)	Ontology-based Clustering	● Path based similarity measure. ● Information content based similarity measure. ● Feature based similarity measure. ● Term re-weighting integrated with domain ontology. ● K-means clustering.
2	Biomedical document clustering using ontology-based concept weight (Logeswari & Premalatha, 2013)	Concept-based Clustering Technique	● The phrases in the document are tokenized using the tri-gram technique. ● The keywords are searched in the MeSH ontology for their existence. ● The weight for the terms and phrases can be calculated dynamically based on the parent-child relationship. ● Concept-based weighting scheme computes the significance of the underlying text by converting the documents into bags of concepts.
3	Ontology-Based Concept Weighting for Text Documents (Tar, H.H., & Nyaunt, T.T., 2011)	Ontology-based Clustering	● Defined an ontology ● Conceptual weighting ● K-means clustering
4	A Survey of Document Clustering using Semantic Approach (Saiyad et al., 2016)	Concept-based Clustering Technique	● Latent Semantic Indexing (LSI) ● Lexical Chains ● Ontology
5	Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means	Fuzzy K-means based Clustering Technique	● Text documents preprocessing ● Bag of words model representation ● Local term weight ● Global term weight ● Fuzzy K-means clustering ● Calculate probability of documents

	Clustering (Rashid et al., 2019)		● Calculate probability of documents over topics ● The probability of words over documents ● The probability of words topics
6	Medical Document Clustering Using Ontology-Based Term Similarity Measures (Zhang et al., 2008)	Concept-based Clustering Technique	● Term re-weighting of document vectors ● Integrated domain ontology to the clustering process

2.5 ADHD Diagnostic Approaches Beyond SNAP IV

In this section, we compare between current ADHD diagnostic approaches that didn't rely on the SNAP IV in terms of the techniques employed for identifying ADHD symptoms on the one hand, and diagnosing subjects whether they have ADHD or not.

In the work that was conducted by Zhang et al. (Zhang et al., 2017), an automatic extraction of psychiatric symptoms from clinical text using unsupervised distributional representations was carried out using three main modules. First, the seed symptoms were collected from three external domain-specific knowledge databases (i) MedlinePlus, (ii) Mayo Clinic and (iii) American Psychiatric Association. Second, the 2016 CEGS-N-GRID challenges data with paragraph2vec were used to generate the distributional representation of terms. Third, the semantic similarity between the distributional representations and candidate phrasal syntax symptoms were calculated. Results showed that the semantic measure based on the distributional representation of terms outperforms the random ordering list and *tf.idf* -based cosine similarity.

(Ahmadi et al., 2021) in their study proposed a novel deep learning approach to classify the EEG signal of Healthy children (normal) from ADHD children (ADHD - Inattention, ADHD -

combined). The input data is the raw ECG records. The implicit feature extraction (spatial and frequency bands) is performed using the deep neural network architecture. The model achieves high classification accuracy.

ADHD and standard control groups have been identified using a broad confidence network in the analysis. Different ADHD subtypes have been classified by resting-state fMRI evidence from stable individuals. (Kuang & He, 2014) used a learning algorithm to try to identify the ADHD target utilizing fMRI graphics. In this research, the author uses a deep belief network for the distinct classification of ADHD and regular monitoring types. Information collected from the restorative status fMRI has been used to distinguish various ADHD subgroups from normal subjects. The findings were confirmed using a 72 per cent precision interdisciplinary cross-validation process.

CEGS N-GRID 2016 task 2 RDoC classification: Prediction of symptom severity using information in psychiatric notes (Liu et al., 2017) in their study used the Natural Language Processing for text normalization and question-answers transformation into multiple data types. Bag-of-Words feature extraction used to train one support vector machine regression for the severity score prediction to feed gradient tree boosting classifier. The final submission achieved 81.75% of the normalized score.

2.6 ADHD Questionnaire Diagnostic Tools

In this section, we compare between current ADHD diagnostic approaches that rely on subjective reports and observation and compare them with the SNAP IV tool in terms of the psychometric

properties, scale, diagnosis and rating. In Table 2, we provide a comparison of the ADHD diagnostic tools and highlight the main differences.

Table 2: Comparison between ADHD diagnostic rating tools

Tool	Strengths	Weaknesses	Psychometric Properties	Versioning Changes	Scale	Diagnosis	Rating
SWAN[13] (Swanson et al., 2012)	Captures variance in populations that is not detected by other scales	Use is genetic studies is undetermined	High validity and reliability	Analysis of evidence for effectiveness is in progress	0-7 scale (18 items)	Raw scores are first converted to T-scores.	Rated based on frequency
SNAP-IV[14] (Hall et al., 2020; Bussing et al, 2008)	Efficiently assesses core symptoms of ADHD	Parent scale is not a good predictor of clinical diagnosis	High validity and reliability	Validity for use in randomized controlled trials is being assessed	0-3 scale (26 items)	High score on all items identifies ADHD diagnosis	Rated based on frequency
ADHD-RS[15] (Dobrean et al., 2019)	Wide use in research and practice	Cannot be used for diagnosis alone	High validity and reliability	Updated for DSM-5 criteria	0-3 scale (9-items)	Sum of scores determines ADHD score	Adding odd and even numbers

[13] https://www.amerihealth.com/pdfs/providers/resources/worksheets/prevhealth_swan.pdf

[14] https://www.adhdfoundation.org.uk/wp-content/uploads/2019/04/ADHD-screening-with-SNAP.pdf

[15] https://pcptoolkit.beaconhealthoptions.com/wp-content/uploads/2016/01/cms-quality-child_adhd_rating_scale_screener.pdf

CAARS[16] (Harrison et al., 2019)	Wide use in research and practice	Low diagnostic validity since it is a self-rated scale	High Cronbach alpha reliability	Adaptations and Translations are in progress	0-3 scale	Raw scores are first converted to T-scores. Higher scores indicate ADHD	Rated on how true a question might be for the child

2.7 Summary

In this chapter, we have studied various aspects of ADHD, its diagnostics, and persistent problems with ADHD patients in two broader categories of adolescents and early age kids. There are various techniques applied in the papers as shown in Table 2.2. However, the common pattern among them is the use of symptoms and questionnaires of ADHD patients. It can be generalized as the symptoms and responses in the ADHD patients vary with nature, nurture, and provided environment, many clinicians conduct a repetition of lengthy questionnaires for diagnosis. This results in inappropriate medication and dosage decisions that may lead to certain complexities among the individuals. To address this issue, well-defined mathematically independent clusters are necessary for the SNAP-IV questions. There exists a reasonable amount of literature for the diagnosis of ADHD. However, in the domain of text mining, a limited amount of work is done and it needs to be addressed in order to cover up the gap.

[16] https://documents.acer.org/caars-self-l-int1.pdf

Chapter 3 - System Overview

3.1 System Architecture

In our proposed system, we developed multiple clustering techniques with the help of a number semantic and contextual terms embedding algorithms for re-categorizing the set of ADHD questions. As part of the implementation of the first technique, we constructed lexical semantic knowledge graphs to represent the relationships that link the original terms of each ADHD question their semantically-related medical terms, such hypernyms/meronyms that are defined in the exploited medical knowledge resources. To do this, we developed a term re-weighting algorithm to select those terms that share common medical concepts in the subsumption hierarchical structure of the used medical resources. As far as the second and third techniques are concerned, we utilized NLP-based pre-trained language models, such the BERT and BioBERT where candidate medical expansion terms are excluded based on their contextual relationships to the original question terms in each ADHD question. In this context, and unlike the first technique, we take advantage of the pre-training process to find additional expansion terms that can be used to further improve the quality of the clustering results. It is import to point out that we have utilized the Latent Dirichlet Allocation (LDA) technique for the purpose of defining topics of the resulting clusters by each of the employed clustering algorithms. Figure 1 depicts the overall architecture of the proposed techniques.

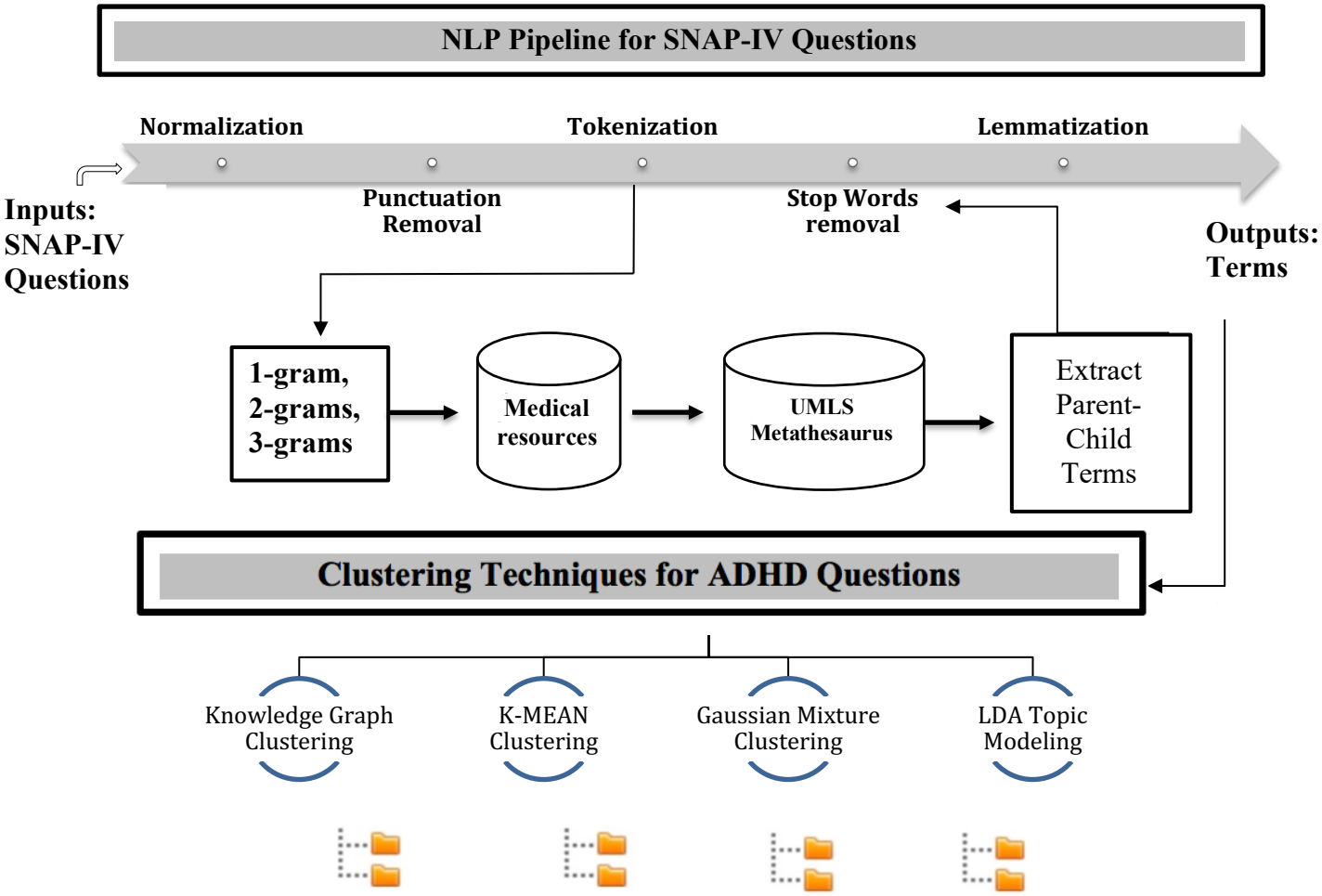


Figure 1: System Architecture

3.2 Detailed Characterization of the Proposed NLP Pipeline

The proposed NLP pipeline comprises a number of phases that are sequenced according to their priorities and execution precedence. As depicted in Figure 2, the full set of ADHD questions are submitted as input to the NLP pipeline. Accordingly, the first phase of text pre-processing starts. In this phase, the text of each question is cleaned using a series of steps, such as text normalization, unwanted words and special characters' removal. The next sections provide more details on each of these steps, in addition to the rest of the phases of the proposed NLP pipeline.

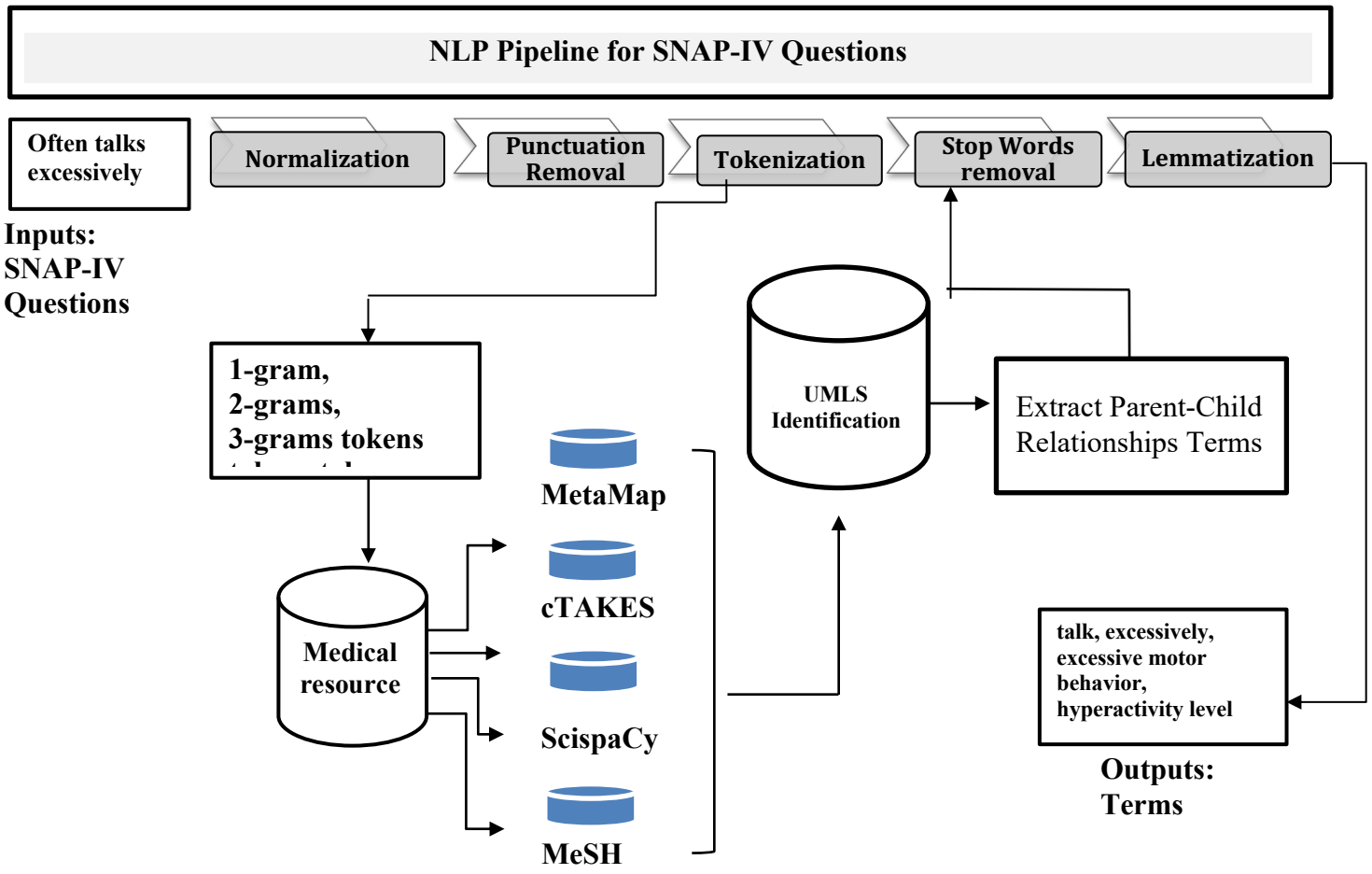


Figure 2: NLP Pipeline

- **Text Normalization**

Similar to other NLP pipelines, and as a conventional NLP step, we apply word normalization on the tokens of each and every question among the full set of ADHD questions, regardless of its initial category that it belongs to. As such, token are convert to a consistent common form where superficial differences among tokens will have no impact on the quality of the clustering

techniques. In this context, all tokens in our text are canonicalized and standardized in order to be further submitted to the exploited knowledge resources to extract semantically-relevant medical entities that can be used for expansion purposes in the same manner as reported in References (Ibrahim et al., 2019) and (Naseem et al., 2019). Below is an example from SNAP-IV questions set (Q#5, Q#6, Q#7).

Before Text Normalization:

5. Often has difficulty organizing tasks and activities.

6. Often avoids, dislikes, or reluctantly engages in tasks requiring sustained mental effort.

7. Often loses things necessary for activities (e.g., toys, school assignments, pencils, or books).

After Text Normalization and Punctuation Removal:

5. often has difficulty organizing tasks and activities

6. often avoids dislikes or reluctantly engages in tasks requiring sustained mental effort

7. often loses things necessary for activities toys school assignments pencils or books

- **Tokenization**

The process of splitting text into sequences of tokens is normally referred to as Tokenization. Tokens are divided according to a set of boundaries. These boundaries are normally whitespace characters (Afsharizadeh et al., 2018). Using this process, we can split paragraphs into phrases or phrases into single sentences. Further, these sentences can be divided into words by the same procedure. As we mentioned earlier in this section, normally this splitting is performed through white spaces (Ibrahim et al., 2019) and (Basile et al., 2014). In our research and in order to build

the knowledge graph, we will identify the n-gram medical entities through the use of medical knowledge bases (MetaMap, cTAKES, ScispaCy and MeSH Ontology). The MetaMap tool is used in highlighting medical concepts in healthcare. Medical professionals use the MetaMap to know the relationships between medical terms found in clinical notes and their corresponding semantic groups of symptoms. The tool uses UMLS in the extraction process (Chiaramello et al., 2016) to identify semantic relations among the medical terms. cTAKES is another tool that is used to process clinical notes. cTAKES uses unstructured information management applications to produce and extract large-scale terminologies based on certain medical results. As highlighted by Ghiasvand in (Ghiasvand, 2014), both MetaMap and cTAKES are used to identify the names of diseases and disorders. Another tool, the ScispaCy is used to process biomedical data and clinical text, highlighting all the necessary keywords to extract useful medical concepts. By outlining all the keywords, a medical practitioner or a patient can familiarize themselves with different medical requirements (Neumann et al., 2019). Furthermore, the MeSH tool is mostly applicable in scenarios where there are advanced medical terms that need to be identified (Batet et al., 2011). The MeSH tool is only applicable in a few categories as compared to the other three tools. Compared to the four tools, the most effective tool concerning mental health would be the MetaMap as it covers a wider range of medical concepts associated with a richer level of semantic and lexical relationships. In the case of the MeSH tool, it has minimal benefits when used in the mental health sector (Ghiasvand, 2014), given the limitedness of the terminologies that it covers. The MetaMap also extracts advanced terminologies from all the categories as compared to the MeSH tool which is compatible with fewer mental health categories. Below is an example from SNAP-IV questions set (Q#5, Q#6, Q#7) for the tokenization step.

Before Tokenization:

5. often has difficulty organizing tasks and activities.

6. often avoids dislikes or reluctantly engages in tasks requiring sustained mental effort.

7. often loses things necessary for activities toys school assignments pencils or books

After Tokenization (tokens separated by "|"):

5. often | has | difficulty organizing tasks | and | activities

6. often | avoids | dislikes | or | reluctantly | engages | in | tasks | requiring | sustained mental effort

7. often | loses | things | necessary | for | activities | toys | school assignments | pencils | or | books

- **Stopword Removal**

In many cases, and as part of any conventional NLP pipeline, there are words that get removed from the text as they appear to be of little significance and don't carry much contribution to actual meanings of sentences. Such words are normally pre-defined in a list that can be extended either manually or automatically depending on the task at hand. For instance, the NLTK[17] and Spacy[18] libraries include such lists and provide them to the programming for further editing and update when necessary. In other works, and more specialized domains, stopword lists can be enriched using the *tf.idf* technique which is employed in this context to find terms of low weights (those are of weights less than a threshold value) and accordingly they get removed. In our research work, we have employed the pre-defined lists that are defined in both third-party libraries (NLTK and

[17] https://www.nltk.org/
[18] https://spacy.io/

Spacy) in order to remove all unwanted stopwords from the ADHD questions set. Below is an example from SNAP-IV questions set (Q#5, Q#6, Q#7) for the stopwords removal step.

Before stopwords removal (tokens separated by "|"):

5. often | has | difficulty organizing tasks | and | activities

6. often | avoids | dislikes | or | reluctantly | engages | in | tasks | requiring | sustained mental effort

7. often | loses | things | necessary | for | activities | toys | school assignments | pencils | or | books

After stopwords removal (tokens separated by "|"):

5. often | difficulty organizing tasks | activities.

6. often | avoids | dislikes | reluctantly | engages | tasks | requiring | sustained mental effort

7. often | loses | things | necessary | activities | toys | school assignments | pencils | books

- **Lemmatization**

An important phase among the NLP pipeline phases is the term lemmatization phase. The lemmatization process is utilized for finding the lemma of each word that appears in the list of processed tokens of each ADHD question. We employed this technique specifically to obtain semantically-related terms from the exploited medical knowledge resources. As such, when extracting the lemma of a given token, it is this what gets submitted to the medical knowledge base to find whether it has hypernyms/meronyms or any other relevant terms. However, it is important to point out that despite the fact the lemmatization can assist in accurately linking tokens to their correspondences in the medical resources, the process can hinder the run-time efficiency of the

overall clustering process as reported in (Basile et al., 2014). Below is an example from SNAP-IV questions set (Q#5, Q#6, Q#7) for the lemmatization removal step.

Before lemmatization (tokens separated by "|"):

5. often | difficulty organizing tasks | activities.

6. often | avoids | dislikes | reluctantly | engages | tasks | requiring | sustained mental effort

7. often | loses | things | necessary | activities | toys | school assignments | pencils | books

After lemmatization (tokens separated by "|"):

5. often | difficulty organize task | activity.

6. often | avoid | dislike | reluctantly | engage | task | require | sustain mental effort

7. often | lose | thing | necessary | activity | toy | school assignment | pencil | book

3.3 Clustering Techniques for ADHD Questions

To cluster the full set of ADHD questions, we have utilized four different clustering models as depicted in Figure 3. These are: i) *tf.idf* based term re-weighting and medical knowledge graph based clustering, ii) K-means clustering model based on the BERT/BioBERT pre-trained term weighting model, iii) Gaussian Mixture based clustering using the BERT/BioBERT pre-trained term weighting model, and iv) BoW and Glove/Word2Vec term weighting in combination of LDA to naming the produced clusters.

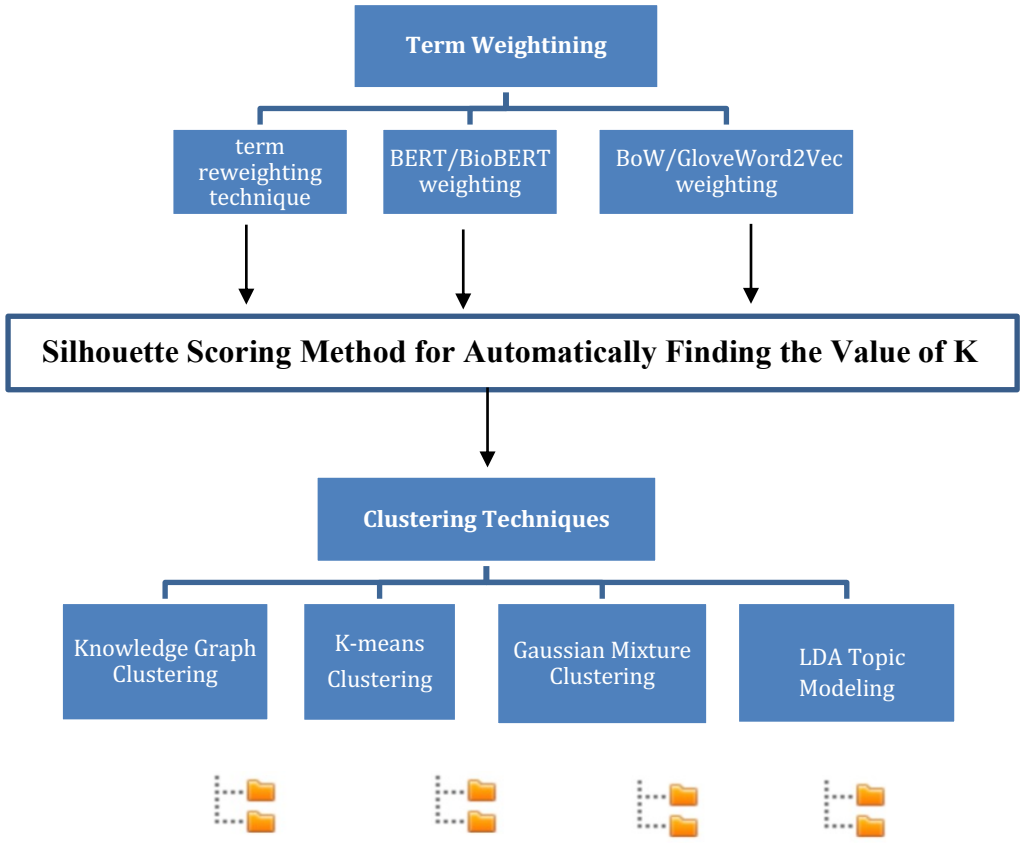


Figure 3: Proposed ADHD Questions Re-Categorization Models

- **Identifying the value of *k* in the Proposed Clustering Models**

One of the important steps in any clustering task is to specify the value of *k* clusters based on which the clustering model will operate. Manual assignment of the value of *k* is error-prone and time-consuming. Therefore, in our proposed models, we rely on the silhouette scoring method for automatically finding the value of *k*. In this context, a silhouette scoring method is utilized to calculate the silhouette coefficient for each data point using Equation 1 (Kaoungku et al., 2018):

$$s(o) = \frac{b(o) - a(o)}{Max\{a(o), b(o)\}} \quad \textbf{(1)}$$

Where:

- s(o) is the silhouette coefficient of the data point o.
- a(o) the internal average distance between the point o and others belongs to the same cluster.
- b(o) the minimum average distance between the point o and others clusters which o does not belong to.

The value is in the range of [-1, 1]. Where:

- 1 score indicate that the data points is close within the same cluster, far away from others clusters.
- 0 score indicate overlapped clusters.
- -1 score indicates the data points assigned into the wrong cluster.

We explicitly assigned k clusters range, for each k clusters we calculate the silhouette score and compute the mean value, the largest mean value between all means indicates the optimal number of clusters (Shahapure & Nicholas, 2020).

In our proposed research project, we have implemented the below clustering models for re-categorizing ADHD questions:

- **Knowledge Graph based Clustering**

In this research, we exploited the medical knowledge bases in order to construct knowledge graphs to represent the n-gram medical terms from the UMLS metathesaurus. Our goal in this context is to extend the original n-gram tokens with additional lexical semantic related terms that can be

acquired from the used knowledge resources. We mainly focused on hypernyms and meronyms of each n-gram token that are located in the hierarchical structure of the exploited resources. To construct the knowledge graphs, we utilized the NetworkX library to link entities and their relationships.

Initially, the centroids of the clusters are the average of nodes divided by the number of nodes in the graph and each node clusters goes with the closer centroid. We have applied term reweighting technique to represent the original and expanded nodes based on the degree of concept similarity in the vector space. Following feature extraction, we applied graph-based clustering techniques, where each graph and it subsequent subgraphs are grouped based on the top broad term in the semantic hierarchy. More details are described in chapter 4.

- **LDA-based Clustering**

LDA on the other part chooses the most suitable topic for the data, therefore it acts as the intermediate cluster for the data. Based on algorithms, LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Topics describes words cluster based on the probability estimate. The general purpose of the LDA technique is to develop a model based on probability so as to clusters discrete data. The probability in Equation 2 for LDA is given as follows according to (Blei et al., 2003).

$$p(\theta, w, z|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta\right)d\theta \quad \textbf{(2)}$$

N-the number of topics, θ is the set of topic mixture in z and w topics. α - the value of Dirichlet based on word clustering while β is the Dirichlet value based on topic clustering. LDA uses probabilities to determine the value. The number of documents clustered into different topics depend largely on the words that fit the topic; therefore, some clusters may have higher number of documents compared to other. The is no limit in clustering of topics (Liu, 2013).

- **K-means Clustering**

K-mean is focused on mapping out data that have the same characteristic and using the centroids. Number of clusters are first estimated for the algorithm to work on them. The algorithm creates centroids of the probable clusters. The section has no particular pattern predetermined, the algorithm has discretion to gather and cluster content. Equilibrium is achieved by ensuring the centroid is in the middle of a cluster. This is achieved through the computer algorithm. Developing algorithm for the k-mean requires the use of abstract space which is applicable in the various functions for creating distance between the space. The commonly used is the Euclidean distance (Alok Malik, 2019). Distance measurement in the data cluster so as to locate the centroid may differ based on the method used. Equation 3 for K-mean is given according to (Li, 2015).

$$S = \sum_{s=1}^{k} \sum_{i=1}^{n} (x_i - y_s)^2 \quad \textbf{(3)}$$

y_j-the centroid of the cluster and $(x_i - y_s)$ represent the distance, the represent the cluster number while n represents the occurrences in the data. The k value in this case is determined in the input therefore the computer operates with a known number of clusters (Li, 2015).

- **Gaussian Mixture Clustering**

The Gaussian mixture model (GMM) use the Expectation-Maximization (EM) algorithm which extends the K-means approach to perform clustering in two different ways: (i) Instead of assigning examples to clusters to maximize the differences in means for continues variables, the GMM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The aim of this clustering algorithm is to maximize the overall probability or likelihood of the data, given the (final) clusters. (ii) The boundaries are flexible in elliptical shape but require more iterations to reach the convergence.

Unlike the classic implementation of K-means clustering, the general GMM algorithm can be applied to both continuous and categorical variables which provides a wider range of input type support. Equation 4 (Reynolds,2008) provides the GMM model.

$$p(x|\lambda) = \sum_{i=1}^{M} wi\ g(x|\mu i, \Sigma i) \quad \textbf{(4)}$$

Where:

- x: The dimensional continuous valued data vectors.
- wi: The mixture of weights.
- $g(x|\mu i, \Sigma i)$: The component Gaussian densities.

3.3 Summary

In this chapter, our goal is to provide a general overview of our system including the NLP pipeline, the term weighting approaches and SNAP-IV Questions clustering techniques. Five phases for the NLP pipeline: (i) Normalization where all questions text are transformed into small letters and change numbers into text, (ii) Punctuation Removal to remove all commas parentheses, (iii) Tokenization, here we submit the n-gram tokens into medical resources(cTAKES, MetaMap, ScispaCy, MeSH) then check for the their existence in the UMLS metathesaurus, each identify terms expanded to Parent/Child relationships terms, (iv) manual exclude for the stopwords, (v) lemmatization for the terms. Three different term weighting techniques: (i) term re-weighting techniques this is based on the medical expansion terms, (ii) using pre-trained language models, (iii) using traditional weighting methods. Four clustering techniques proposed: (i) defined knowledge graph from medical and expanded medical terms, (ii) LDA topic modeling, (iii) K-means clustering, (iv) Gaussian Mixture clustering. We clarified the role of each model in producing the new SNAP-IV categories.

Chapter 4 – System Implementation Details

In this chapter, we present the implementation details of the proposed system. This includes the NLP pipeline, feature extraction techniques, term re-weighting technique, the algorithmic steps for each in sections 4.1, 4.2, 4.3 and clustering approaches algorithmic steps in section 4.4. Finally, we summarize this chapter in section 4.5.

4.1 Development Details of the Proposed NLP pipeline

The pre-processing includes text cleaning, normalization, tokenization, stopword removal and lemmatization. Following are the implementation steps:

1. Inputting the Data (Text from SNAP IV Questions). The text is inputted in an excel format and read using the pandas[19] Python library to structure our data for analysis.
2. Cleaning The Data. We convert all question text to lower case and remove punctuations.
3. Tokenization. This entails n-gram tokenization from medical knowledge-based resources. We submit unigram/bigram/ and trigram tokens to the medical resources.
4. Medical Term Expansion. In this step, we submit each of the recognized n-gram tokens exploited from the knowledge-based medical resources into the UMLS Metathesaurus.
5. Manual exclusion of common English stopwords, then lemmatization.

After submitting and cleaning the data, we started the tokenization process. From 242209 MeSH term 123 matched texts were recognized. In ScispaCy medical resource 266 unique

[19] https://pandas.pydata.org/

matched text were recognized (canonical name, aliases, semantic type and definition) for each concept from ~785k vocabulary and 600k word vectors (For more details on the ScispaCy recognition, we refer the reader to Appendix B). Below is an example of ScispaCy recognition of a UMLS entity:

```
Name:  difficulty
CUI: C1299586, Name: Has difficulty doing (qualifier value)
Definition: Something not easily done, accomplished, comprehended, or
solved.
TUI(s): T033
Aliases: (total: 10):
      Difficult, Difficult, Able with difficulty, Able with
difficulty (qualifier value), Difficulty, Difficulty, Difficulty,
difficulty, difficulties, Has difficulty doing


Name:  seated
CUI: C0277814, Name: Sitting position
Definition: The state or act of one who sits; the posture of one who
occupies a seat.
TUI(s): T033
Aliases (abbreviated, total: 18):
      Sitting position, Position, Sitting, sitting position, Sitting
Position, Sitting Positions, Sitting, Sitting, Sitting, Sitting,
Sitting


Name:  classroom rules
CUI: C0870287, Name: Classroom
Definition: A room or space, especially in a school, where classes
are held.
TUI(s): T073
Aliases: (total: 4):
      classroom, classroom, classrooms, Schoolroom
```

In the cTAKES (under SNOMEDCT_US coding scheme) 92 matched texts were recognized. Below (Table 3) provides the number of recognized phrases on the medical semantic category (Type) and below is an example of a cTAKES bigram recognized output:


```
leisure activities{
DiseaseDisorderMention: [
{
begin: 2086,
end: 2095,
text: "impulsive",
polarity: 1,
conceptAttributes: [
{
code: "286756000",
cui: "C0564567",
codingScheme: "SNOMEDCT_US",
tui: "T048"
}
],
type: "DiseaseDisorderMention",
}
```


Table 3: cTAKES-based Recognition of Term Semantics

Disease Disorder	Anatomical Site	Sign Symptom	Procedure
16	5	70	1

Furthermore, in the MetaMap medical resource, 193 unique recognized tokens matched from different sources like MeSH, NCI, HI7, MTH, ICD9. Below is an example for the MetaMap unigram (motor) recognition output which contains in addition to the medical sources, a preferred name and semantic type.

```
conceptinfo: {
conceptstring: "Motor",
sources: [
"MTH",
"NCI_FDA",
"NCI"
],
cui: "C1705994",
preferredname: "Device Motor",
semantictypes: [
"mnob"
]
```


Table 4 provides the number of unique recognized tokens exploited from each one of the four medical resources (MetaMap, cTAKES, ScispaCy and MeSH) recognized from the n-gram SNAP-IV questions text tokens.

Table 4: Number of Unique Recognized Tokens from Medical Resources

MetaMap	cTAKES	ScispaCy	MeSH
193 unique matched text	92 unique matched text	266 unique matched text	123 unique matched text

Then, we submit each of the recognized n-gram tokens exploited from the knowledge-based medical resources into the UMLS Metathesaurus. The number of tokens Identified by UMLS Metathesaurus is 530. For each Identified token, we conducted an additional extraction of semantically related concepts (hypernym/meronym) referred to as enrichment terms. We used the following expansion algorithm where we get from UMLS metathesaurus the (Broader, Narrower) Concepts for each recognized UMLS token and add them into a new lists:

Algorithm: Expansion of medical terms into their semantically related (hypernym/meronym) terms

Input: Med_list [t1,t2, …tn]
Output: list of semantically related hypernyms and meronyms terms

```
1: Parent_list ← 〈 〉 ;
2: Child_list ← 〈 〉 ;
3: Enriched_list ← 〈 〉 ;
4: for i¬0; i< Med_list.length; i++
5:      if Identified_UMLS(Med_list[i] ) then
6:              ADD(Parent_list,GET_Parent_FROM_UMLS(Med_list[i]));
7:              ADD(Child_list,GET_Child_FROM_UMLS(Med_list[i]));
8:      end if
9:  end for
10: Enriched_list = Parent_list.extend(child_list);
11: Return Enriched_list;
```

For purposes of illustrating the outputs of step #4, below (Table 5) we provide an example from SNAP IV question 66 after the term expansion based on the hypernym/meronym relationship.

Table 5: Enriched Terms Example from SNAP-IV

Question	Medical Term	Narrow Terms	Broad Terms
sometimes for at least two weeks has depressed mood sad hopeless discouraged	depressed mood	C/O - feeling depressed	assessment of mood
		Decreased Depressed Mood	EMOTION, AFFECT AND MOOD
		Depressed mood in Alzheimer's disease	Emotional problems
		Depressed mood with postpartum onset	Emotional state observations
		depressed, but unlike previous grieving for	

	a death or loss	Emotions
	Depressive episode, unspecified	Level of depression
	Depressive symptoms due to primary psychotic disorder	Level of fatigue

In the final step, we manually removed common stopwords. Below (Table 6) is an example that demonstrates the outcomes of this step using four questions from the SNAP-IV dataset.

Table 6: Lemmatized SNAP-IV Questions sets

Question	Clean question
Often fails to give close attention to details or makes careless mistakes in schoolwork or tasks	often fail give close attention detail make careless mistake schoolwork task
Often has difficulty sustaining attention in tasks or play activities	often difficulty sustain attention in task play activity
Often does not seem to listen when spoken to directly	often not seem listen speak directly
Often does not follow through on instructions and fails to finish schoolwork, chores, or duties	often not follow through instruction fail finish schoolwork chore duty

4.3 Development Details for Knowledge Graph Clustering

We have exploited a number of medical resources in order to expand questions and accordingly construct knowledge graphs that represent each question. The goal of this step in this context is to

link question terms by different types of semantic relations that are encoded in the used medical knowledge resources. Such relations are later incorporated for the purpose of clustering of ADHD questions. In each knowledge graph, nodes are the n-gram medical entities that are recognized from the medical resources and identified by the UMLS metathesaurus with their enrichment of hypernym/meronym entities. Edges are the links between the nodes representing the hierarchical semantic relationship. Figure 4 depicts a knowledge graph that was automatically constructed for SPAN-IV question number 19:

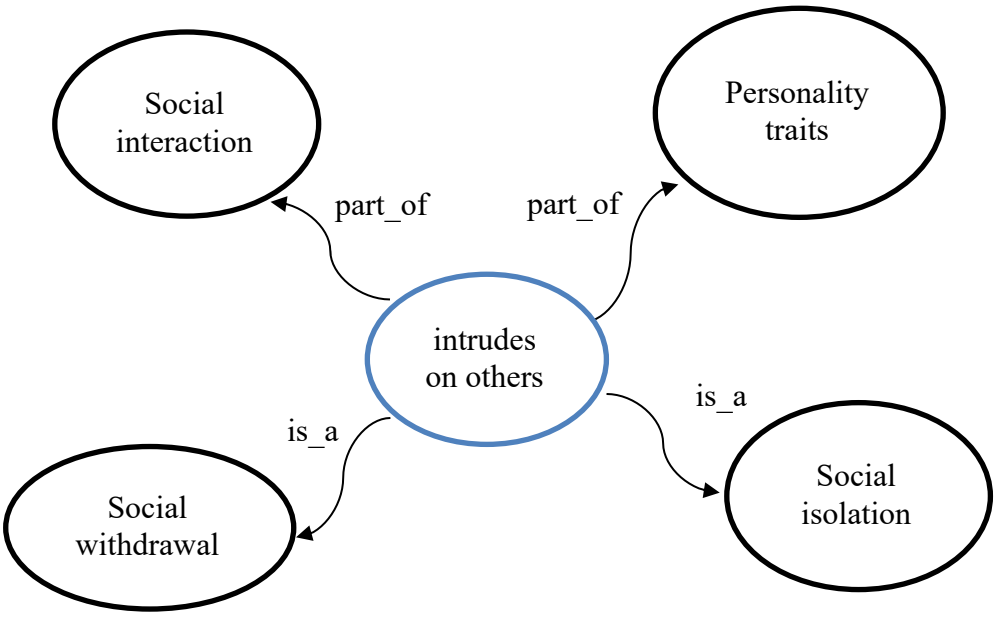


Figure 4: Sample of the Knowledge Graph for SNAP-IV

Each subgraph is clustered and labeled based on the medical knowledge-based resources. The question in Figure-4 was categorized under hypernem/meronym “ability to interact with others”). For the group in Figure 4, the n-gram medical tokens exploited from the medical knowledge-based resources are (interrupts, intrudes on others, annoy other people, blames others, misbehavior,

easily annoyed by others, response to stress, depressed mood) all shared the same enrichment medical concept (ability to interact with others) where it assigned as this category label. The medical terms which are distributed among different questions with the same hypernym relationship are reassessed using the term re-weighting algorithm:

Algorithm: Medical Term re-weighting for classification purposes

Input: Med_list [t1,t2, …tn]
parent_list[h1,h2,...hn]
Output: list_terms_concept[ ]

```
For every term  in Med_list do
    calculate the semantic weight wt of term ti
    For every enriched hj in parent_list do
        if term ti is a parent of bj then
            wt += ti +1
            Add ti to a new list_terms_concept
        end if

    End for
End for
```

Table 7 provides an example of two subgraph clusters of symptoms/questions based on the new medical term re-weighting classification:

Table 7: Knowledge Graph based Enhancement on the Original Questions

Ability to interact with others	Disturbance in the content of thought
Often **interrupts** or **intrudes on others** butts into conversations games	Has **obsessions** (persistent and intrusive inappropriate ideas, thoughts, or **impulses**)
Often deliberately does things that **annoy other people**	Often has need for undue admiration, **grandiose behavior**, or lack of empathy
Often **blames others** for his or her mistakes or **misbehavior**	Often **blames others** for his or her mistakes or **misbehavior**
Often touchy or **easily annoyed** by **others**	Sometimes for at least a week has **flight of ideas** or says that thoughts are racing

Currently has an emotional nervous, worried, hopeless, tearful **response to stress**	Has episodes of failure to resist **aggressive impulses** (to **assault others** or to destroy property).
Sometimes for at least two weeks has **depressed mood, sad**, hopeless, discouraged.	Often changes **mood** quickly and drastically

As observed from the example above, the question "Often blames others for his or her mistakes or misbehavior" is categorized under both groups. For this, we measure the semantic distance from the medical term node which represents the number of edges in the graph structure. Ultimately, the question is assigned to the closer distance node in the hierarchy. Table 8 provides the new categorization and number of questions per category based on knowledge-gram term re-weighting method:

Table 8: Produced Categories from Knowledge Graph Clustering

Topic	Topic medical enrichment	Number of questions
Topic 1	Hyperactivity Level	6
Topic 2	Level of Agitation	7
Topic 3	Ability to interact with others	8
Topic 4	Abnormal behavior	12
Topic 5	Emotions	5
Topic 6	Disturbance in the content of thought	6
Topic 7	Cognitive Response	2
Topic 8	Ability to participate in classroom activities	3

Each group of questions could be considered as a group of similarities in the vector space based on the term re-weighting technique. We used the cosine similarity algorithm (below) to measure similarity score between questions.

Algorithm: Similarity Measure: Cosine Similarity Algorithm

```
1: def cosine_similarity(weight_vector):
2:   cos_df = pd.DataFrame()
3:   counter = 0
4: for x in weight_vector:
5:   cosine_dict = {}
6:   cos_vector = []
7:   for y in weight_vector:
8:     cos_sim = dot(weight_vector[x], weight_vector[y])/(norm(weight_vector[x])*norm(weight_vector[y]))
9:     cos_vector.append(cos_sim)
10:   cosine_dict[x] = cos_vector
11:  if(x == 0):
12:        cos_df = pd.DataFrame.from_dict(cosine_dict)
13:  else:
14:         new_df = pd.DataFrame.from_dict(cosine_dict)
15:         cos_df = cos_df.join(new_df)
16:  return cos_df
```

Table 9 show the difference between finding the similarity scores from traditional *tf.idf* weighting and the term re-weighting algorithm.

Table 9: Similarity Scores based on Cosine Similarity Algorithm

Question	Questions within the same group	Similarity Score based on Term re-weighting technique	Similarity Score based on TF.IDF weighting
Often interrupts or intrudes on others butts into conversations games	Often deliberately does things that annoy other people	0.8257674	0
	Often blames others for his or her mistakes or misbehavior	0.7624129	0
	Often touchy or easily annoyed by	0.8435933	0

	others		
	Currently has an emotional nervous, worried, hopeless, tearful response to stress	0.60958455	0.104939780147944
	Sometimes for at least two weeks has depressed mood, sad, hopeless, discouraged.	0.50172452	0.201550747713151

4.4 Development Details for LDA Topic Modeling

The aim of employing this model is to automatically assign the topic/topics that can describe each group of questions that may fall under the same category. To find the hidden topics related to SNAP-IV questions, we applied LDA topic modeling where every question is a probability distribution of topics and every topic is a probability distribution of words. We applied two feature extraction techniques, namely (1) Bag-of-Words, and (2) GloveWord2Vec.

First, we applied LDA Topic Modeling with the Bag-of-Words (BoW) weighting model:

1. Assigned Optimal topic number using silhouette score when running K-means
2. Converted the words into bag-of-words model.
3. Applied LDA modeling.
4. Extracted the top 10 words and their probability for each topic.

Table 10 provides the resulting topics from LDA modeling. Each topic includes the top 10 frequent keywords

Table 10: Top 10 keywords produced based on LDA (BoW) Model

Topic	Topic keywords
Topic 1	classroom, week, difficulty, mood, motor, impulse, verbal, excessively, period, concentrate
Topic 2	week, behavior, currently, activity, self-esteem, pick, difficulty, request, repetitive
Topic 3	difficulty, activity, classroom, seat, task, easily, problem, year, time, chronic
Topic 4	hyper-vigilant, self-esteem, careless mistake, anger outbursts, muscle tension, Chronic feelings, leaves seat, depressed mood, apprehensive, attention seeking behavior

Second, we applied LDA topic modeling using the GloveWord2Vec weighting model. The main difference here is that the word weights in the vector space were based on the Stanford GloVe[20] pre-trained model. Following are the additional steps implemented to find the topics:

1. Built the embedding dictionary using Glove weights.
2. Built Word2Vec model from GloveWord2Vec based on cosine similarity algorithm.
3. Applied LDA topic modeling to label each predicted cluster.

Table 11 summarizes the predicted topics. Each topic includes the top 10 frequent keywords

Table 11: Top 10 keywords produced based on LDA (GloveWord2Vec) Model

Topic	Topic keywords

[20] https://nlp.stanford.edu/projects/glove/

Topic 1	rarely, frequently, typically, usually, occasionally, generally, especially, involve, problem, turn
Topic 2	think, difficulty, sit, room, lack, require, pick, difficulty, frequently, repetitive
Topic 3	violate, people, follow, run, task, change, year, fail, usually, frequently
Topic 4	week, classroom, careless mistake, problem, increase, chronic feelings, leaves seat, depressed mood, self-esteem, frequently

4.4 Development Details for the K-Means Clustering Technique

The implementation for this clustering technique with BERT embedding was conducted for two main reasons:

- To predict new non-overlap semantically-related clusters in SNAP-IV.
- To compare the contextual weights results with the term reweighting algorithm in the graph clustering. Following are the implementation steps:

    1. Tokenizing using the BERT tokenizer.
    2. Found the optimal number of clusters using silhouette analysis. Table 12 below shows the silhouette score for each of the resulting four clusters.

Table 12:Silhouette Scores Based on K-Means Clustering

Number of Clusters	Silhouette Score
2	0.25
3	0.19
4	0.17

5	0.18
6	0.17

We plotted silhouette graphical representation for each k. When comparing multiple representations. All clusters are crossing mean value and distribution is also uniform when k =4, Then this value of 'k' is used for clustering.

3. Applied LDA topic modeling for labeling each predicted cluster result. Table 13 provides the produced topics based on K-means clustering.

Table 13: Top 10 Keywords Produced from K-Mean Clustering Results with BERT embedding

Topic	Topic keywords
Topic 1	drive by a motor, activity, task, schoolwork, attention, remaining seated, maintaining alertness, orient, careless mistake, distract
Topic 2	repetitive motor behavior, reduce anxiety, inappropriate, verbal activity, violate, destructive, societal norm, aggressive, behavior, impulsive
Topic 3	week, grandiose behavior, talkative, euphoric mood, relationships, concentrating, irritable, sleepiness, risky, activities
Topic 4	classroom, assignments, psychomotor retardation, psychomotor agitation, staying seated, hopelessness, year, depressed mood, sometimes, period

Similarly, the same procedure applied based on BioBERT embedding (Table 14) provides the produced topics:

Table 14: Top 10 Keywords Produced from K-Mean Clustering Results with BioBERT embedding

Topic	Topic keywords
Topic 1	drive by a motor, activity, task, schoolwork, attention, remaining seated, maintaining alertness, orient, careless mistake, distract
Topic 2	repetitive motor behavior, reduce anxiety, inappropriate, verbal activity, violate, destructive, societal norm, aggressive, behavior, impulsive
Topic 3	week, grandiose behavior, talkative, euphoric mood, relationships, concentrating, irritable, sleepiness, risky, activities
Topic 4	classroom, assignments, psychomotor retardation, psychomotor agitation, staying seated, hopelessness, year, depressed mood, sometimes, period

4.6 Development Details for the Gaussian Mixture Clustering Technique

We used Gaussian mixture clustering algorithm with BERT embedding since our dataset has significant overlapping. K-means clustering does not perform well under overlapping conditions given its reliance on Euclidean distance measures without data points probability. In the Gaussian mixture modeling, each data point assigned a probability in a normal distribution, which in turn, Increases the clustering efficiency.

We followed the implementation steps below to apply Gaussian mixture modeling:

1. Identified the number of clusters using silhouette analysis. Table-15 below shows the silhouette scores for each cluster number.

Table 15: Silhouette Score Based on Gaussian Mixture Model Clustering

Number of Clusters	Silhouette Score
2	0.20
3	0.15
4	0.15
5	0.13
6	0.15

We plotted silhouette graphical representation for each k. When comparing multiple representations. All clusters are crossing mean value and distribution is also uniform when k =4 Then this value of k is used for clustering.

2. Applied the BERT embedding using transformers.
3. Applied Gaussian Mixture clustering
4. Applied LDA topic modeling for labeling each predicted cluster. Table 16 provides the produced topics based on Gaussian Mixture clustering.

Table 16: Top 10 keywords Produced from Gaussian Mixture Clustering Results

Topic	Topic keywords
Topic 1	verbal motor, exaggerated startle response, alert, attention-seeking behavior, euphoric mood, muscle tension, week, motor, excessive, impulsive

Topic 2	irritable mood, aggressive, restless, school absent, pressure, stress, ruining, fighting, behavior, children
Topic 3	week, grandiose behavior, talkative, euphoric mood, relationships, concentrating, irritable, sleepiness, risky, activities
Topic 4	classroom rules, interactions, period, psychomotor retardation, psychomotor agitation, staying seated, hopelessness, year, depressed mood, sometimes

4.6 Summary

The aim of this chapter was to present the methods and techniques we used in our proposed system. We described the implementation steps for the proposed system components. The first component is the NLP pipeline with its various steps, including pre-processing, medical term recognition from medical resources, and medical term expansion using UMLS. Further, we illustrated the term re-weighting technique based on the semantically related concepts. Finally, we expanded on topic modeling and clustering techniques, including: knowledge graph clustering, LDA topic modeling, K-means clustering and Gaussian mixture model clustering).

Chapter 5 – Proposed Prototype Details and Experimental Evaluation

5. 1 Dataset

- The SNAP-IV questions

The SNAP-IV consists of 90 questions (For more details on the used questions and scoring criteria, we refer the reader to Appendix A.1: SNAP-IV-C Teacher and Parent Scale). Table-17 provides details about the n-gram tokens extracted from the 90 SNAP-IV questions based on exploiting medical knowledge-based resources.

Table 17: SNAP-IV Terms Description

Number of questions	Number of Medical terms recognized from MeSH ontology	Number of Medical terms recognized from cTAKES medical resource	Number of Medical terms recognized from sciscpaCy medical resource	Number of Medical terms recognized from MetaMap medical resource	Number of Medical terms Identified in the UMLS
90	123	92	266	193	530

- ADHD questions human expert selection

Clinical experts in the domain select 37 questions from the SNAP-IV as ADHD related questions. This selection was used for our experimental evaluation.

- ADHD and healthy children dataset for SNAP-IV

For validation purposes, we used a dataset provided by the Palestinian Neuroscience Initiative (PNI), consisting of SNAP-IV results for 30 children from Palestine, aged between 4 -6 years old,

18 subjects were diagnosed with ADHD, and 12 matched healthy controls. This dataset is used for the evaluation with the produced clusters from each of our proposed models.

5.2 Developed Prototypes

- **First Prototype**

The first prototype is based on the new categories based on the knowledge graph clustering technique. For the web interface implementation, we used HTML, CSS, and JavaScript functional programming to build the web interface for the questionnaire and embed it within a Google Site published link[21]. Figure 5 provides a screenshot for this prototype. Figure 6 provides the database where scoring for each question is added. Figure 7 provides the automatic average scoring per category.

[21] https://sites.google.com/view/nlp-adhd-umls

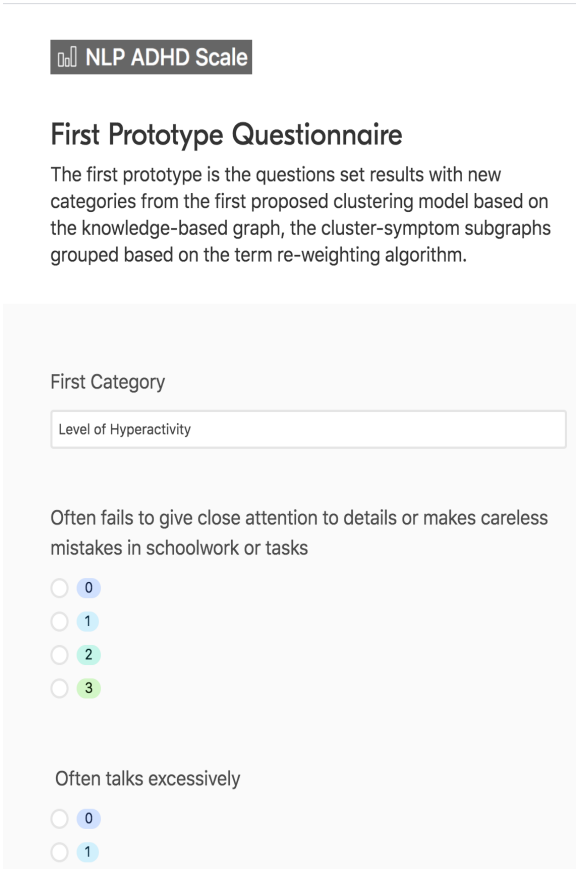
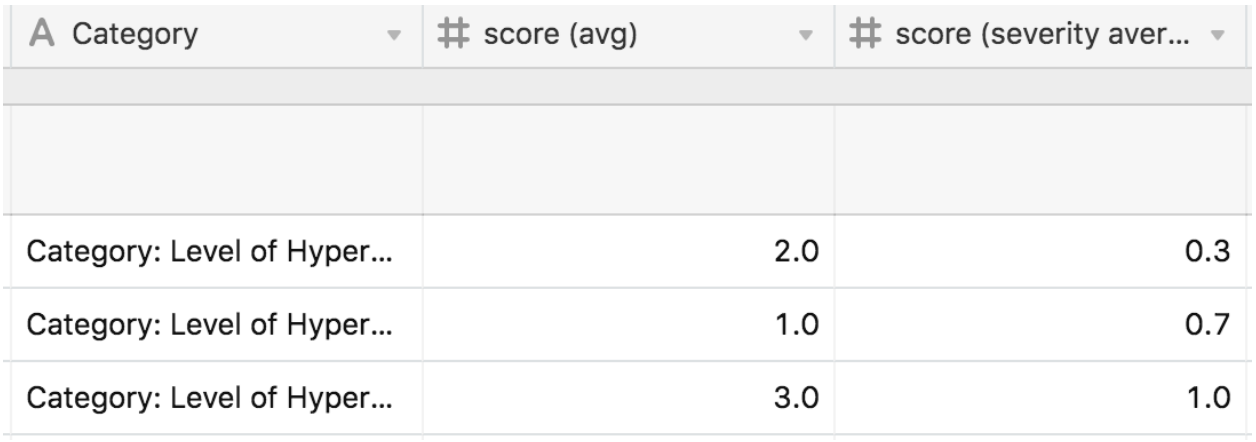


Figure 5: First Prototype Model Web Interface

Often fails to give cl...	Often talks excessi...	Has episodes of fail...	Sometimes for at le...
1	2	2	2
2	2	1	1
1	3	2	1

Figure 6: First Prototype Model Web Scoring database storage



Category	score (avg)	score (severity aver...
Category: Level of Hyper...	2.0	0.3
Category: Level of Hyper...	1.0	0.7
Category: Level of Hyper...	3.0	1.0

Figure 7: First Prototype Model Automatic Average Scoring Calculation

This prototype produces eight new clusters for the 90 SNAP-IV questions and five clusters for the 37 ADHD question set within SNAP-IV. Table-18 provides the produced clusters for both (90 and 37 question Sets).

Table 18: First Prototype Produced Clusters for both (90 and 37 Question Sets)

90 Question Set		**37 Question Set**	
Cluster Number	**Category**	**Cluster Number**	**Category**
1	Hyperactivity Level	1	Hyperactivity Level
2	Abnormal behavior	2	Abnormal behaviors
3	Level of Agitation	3	Level of Agitation
4	Ability to interact with others	4	Ability to interact with others
5	Ability to	5	Ability to

	participate in classroom activities	participate in classroom activities
6	Cognitive Response	
7	Disturbance in the content of thought	
8	Emotions	

We used independent-sample t-tests to compare the averages scores for the answered questions for each category between children with ADHD and matched controls (For more details on the used scoring criteria we refer the reader to Appendix A.2: Scoring Instructions for the SNAP-IV-C Rating Scale) as summarized in Table-19 (re-clustered 90 questions vs. original SNAP-IV clustering). All results showed significant differences between children with ADHD and healthy controls. SEM refer to the standard error of the mean indicates the accurate mean sample compared to the true population mean. The low values indicated that the random sample means are closely distributed to the true population calculated using the standard error.

Table 19: First Model Comparison between the new categories and the baseline from the 90 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P Value)**
Hyperactivity Level	2.25	1.08	0.17	0.17	0.0000
Abnormal behavior	2.29	0.26	0.13	0.09	0.0000

Level of Agitation	2.03	0.42	0.19	0.16	0.0000
Ability to interact with others	1.00	0.17	0.24	0.13	0.0058
Ability to participate in classroom activities	1.89	0.08	0.18	0.05	0.0000
Cognitive Response	0.88	0.17	0.18	0.07	0.0013
Disturbance in the content of thought	0.74	0.22	0.17	0.11	0.0171
Emotions	1.13	0.32	0.20	0.08	0.0013
SNAP-IV categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Overactivity	2.20	0.58	0.15	0.14	0.0000
Aggressive/ Defiance	1.27	0.58	0.17	0.17	0.0085
ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

Similarly, we used independent-sample t-tests to compare the averages scores between children with ADHD and matched controls as summarized in Table-20 (re-clustered 37 questions vs.

original SNAP-IV clustering). All results showed significant differences between children with ADHD and healthy controls.

Table 20: First Model Comparison between the new categories and the baseline from the 37 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
Hyperactivity Level	2.25	1.08	0.17	0.17	0.0000
Abnormal behavior	0.75	0.36	0.06	0.06	0.0001
Level of Agitation	2.03	0.42	0.19	0.16	0.0000
Ability to interact with others	1.00	0.17	0.24	0.13	0.0058
Ability to participate in classroom activities	1.89	0.08	0.18	0.05	0.0000
SNAP-IV categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Over activity	2.20	0.58	0.15	0.14	0.0000
Aggressive/Defi ance	1.27	0.58	0.17	0.17	0.0085

ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

We used logistic regression to assess model performance measurements. Results are summarized in table-21

Table 21: Performance Measurements for the First Model

Measure	Value
Sensitivity	1.0
Specificity	1.0
Precision	1.0
F1 Score	1.0
Accuracy	1.0

The Confusion matrix is [[4 0] [0 5]]. True Positives(TP) = 4, True Negatives(TN) = 5, False Positives(FP) = 0, which is the type 1 error the false positive rate, False Negatives(FN) = 0. Which is type 2 error the false negative rate. Where:

True Positive: The prediction positive (ADHD diagnosis) and its true prediction.

True Negative: The prediction negative (non-ADHD diagnosis) and its true prediction.

False Positive: The prediction positive (ADHD diagnosis) and its false prediction.

False Negative: The prediction negative (non-ADHD diagnosis) and its false prediction.

Sensitivity: calculated based on how many children diagnosed with ADHD disorder using the following formula:

Sensitivity (Recall) = $\frac{number\ of\ true\ positives}{number\ of\ true\ positives+number\ of\ false\ negative}$

Specificity: calculated based on how many children are healthy controls diagnosed with non-ADHD disorder using the following formula:

Specificity = $\frac{number\ of\ true\ negatives}{number\ of\ true\ negatives+number\ of\ false\ positivies}$

Precision: calculated based on the precise positive predictions, how many are actually positive (ADHD diagnosed) using the following formula:

Precision = $\frac{True\ positive}{true\ positive+False\ positive}$

F1 score is a function of precision and recall to calculate the balance using the following formula:

F1-score =2 X $\frac{precision * recall}{precision + recall}$

- **Second Prototype**

The second prototype is the question clustering results using LDA topic modeling. Similar to prototype 1, we used the same technical technologies here to implement the user interface form for this prototype. Figure-8 portrays the user-interface. Table-22 provides the resultant clusters for both the 90 and 37 question sets.

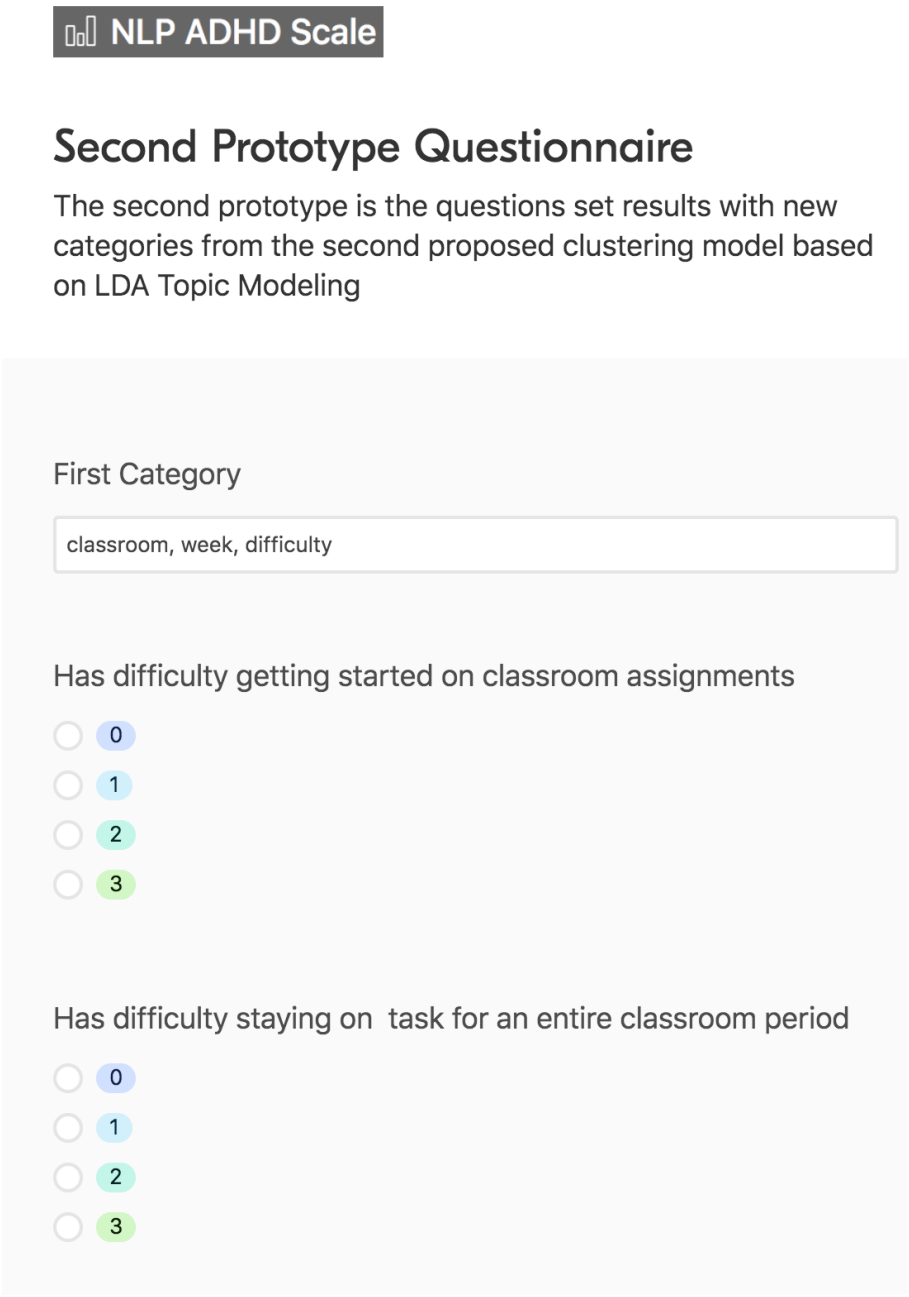


Figure 8: Second Prototype Model Web Interface

Table 22: Second Prototype Produced Clusters for both (90 and 37 Question Sets)

90 Question Set		37 Question Set	
Cluster Number	**Keywords**	**Cluster Number**	**Keywords**
1	classroom, week, difficulty	1	task, activity, difficulty
2	Week, behavior, currently	2	classroom, work, assignment
3	Difficulty, activity, classroom	3	child, overactive, restless
4	Hyper-vigilant, self-esteem, careless mistake	4	classroom, rule, interaction

Using independent-sample t-tests, we compared children with ADHD and matched controls as summarized in Table-23 (re-clustered 90 questions vs. original SNAP-IV clustering). All results showed significant differences between children with ADHD and healthy controls.

Table 23: Second Model Comparison between the new categories and the baseline from the 90 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
classroom, week, difficulty	1.43	0.27	0.16	0.06	0.0000
Week, behavior, currently	1.37	0.22	0.14	0.15	0.0000
Difficulty, activity, classroom	0.20	0.03	0.08	0.01	0.0662
Hyper-vigilant, self-esteem,	1.37	0.25	0.14	0.06	0.0000

careless mistake					
SNAP-IV categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Overactivity	2.20	0.58	0.15	0.14	0.0000
Aggressive/ Defiance	1.27	0.58	0.17	0.17	0.0085
ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

Similarly, we compared children with ADHD and matched controls as summarized in Table-24 (re-clustered 37 questions vs. original SNAP-IV clustering). All results showed significant differences between children with ADHD and healthy controls.

Table 24: Second Model Comparison between the new categories and the baseline from the 37 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
task, activity, difficulty	2.30	0.53	0.11	0.12	0.0000
classroom,	2.23	0.50	0.11	0.11	0.0000

work, assignment					
child, overactive, restless	2.08	0.50	0.19	0.14	0.0000
classroom, rule, interaction	1.89	0.67	0.19	0.15	0.0000
SNAP-IV categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Over activity	2.20	0.58	0.15	0.14	0.0000
Aggressive/Defi ance	1.27	0.58	0.17	0.17	0.0085
ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

To assess the accuracy of the current prototype, we also used logistic regression analysis to assess the sensitivity/specificity of the classification. Results are summarized in Table-25.

Table 25: Performance Measurements for the Second

Measure	Value
Sensitivity	1.0
Specificity	0.8333
Precision	0.75
F1 Score	0.91
Accuracy	0.8889

Confusion matrix is [[3 1] [0 5]]. True Positives(TP) = 3. True Negatives(TN) = 5. False Positives(FP) = 1, False Negatives(FN) = 0.

- **Third Prototype**

The third prototype represents the question categories resulting from K-means clustering based on BERT embedding. Similar to prototypes 1 and 2, we used the same technical technologies here to implement the user interface form for this prototype. Figure-9 portrays the user-interface. Table-26 provides the resultant clusters for both the 90 and 37 question sets.

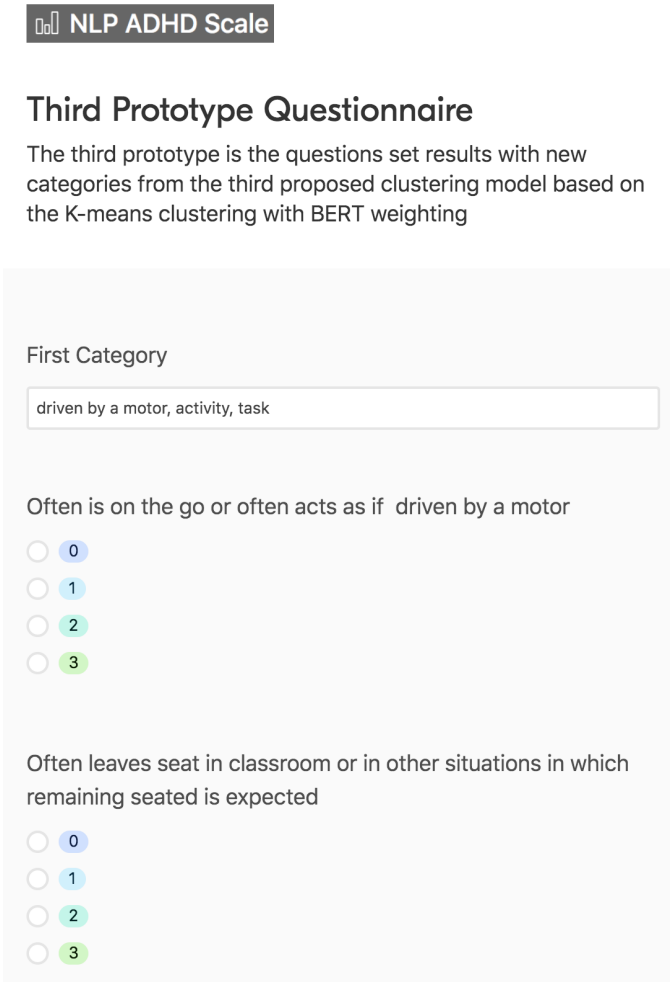


Figure 9: Third Prototype Model Web Interface

Table 26: Third Prototype Produced Clusters for both (90 and 37 Question Sets)

90 Question Set		37 Question Set	
Cluster Number	**Keywords**	**Cluster Number**	**Keywords**
1	driven by a motor, activity, task	1	difficulty, attention, schoolwork
2	repetitive motor behavior, reduce anxiety, inappropriate	2	excessively, home, impulses
3	week, grandiose	3	child, overactive,

	behavior talkative		restless
4	classroom, assignment, psychomotor retardation	4	classroom, rule, interaction

Using independent-sample t-tests, we compared children with ADHD and matched controls as summarized in Table-27 (re-clustered 90 questions vs. original SNAP-IV clustering) and Table-28 (re-clustered 37 questions vs. original SNAP-IV clustering). All results showed significant differences between children with ADHD and healthy controls.

Table 27: Third Model Comparison between the new categories and the baseline from the 90 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P Value)**
driven by a motor, activity, task	2.35	0.48	0.09	0.08	0.0000
repetitive motor behavior, reduce anixety,inappropriate	1.55	0.59	0.16	0.10	0.0000
week, grandiose behavior talkative	0.59	0.06	0.20	0.05	0.0189
classroom, assignment, psychomotor retardation	1.10	0.24	0.13	0.06	0.0000
SNAP-IV categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the**	**HC Standard Error of the**	**T-Test (P Value)**

			Mean	Mean	
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Overactivity	2.20	0.58	0.15	0.14	0.0000
Aggressive/ Defiance	1.27	0.58	0.17	0.17	0.0085
ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

Table 28: Third Model Comparison between the new categories and the baseline from the 37 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P Value)**
difficulty, attention, schoolwork	2.39	0.47	0.09	0.06	0.0000
excessively, home, impulses	2.02	0.39	0.16	0.13	0.0000
child, overactive, restless	1.88	0.56	0.15	0.12	0.0000
classroom, rule, interaction	1.44	0.12	0.18	0.08	0.0000
SNAP-IV categories	**ADHD Average**	**HC Average**	**ADHD Standard**	**HC Standard**	**T-Test (P Value)**

	Score	Score	Error of the Mean	Error of the Mean	
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Over activity	2.20	0.58	0.15	0.14	0.0000
Aggressive/Defi ance	1.27	0.58	0.17	0.17	0.0085
ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

We used logistic regression analysis to assess the sensitivity/specificity of the classification as summarized in Table-29.

Table 29: Performance Measurements for the Third Model

Measure	Value
Sensitivity	1.0
Specificity	1.0
Precision	1.0
F1 Score	1.0
Accuracy	1.0

The Confusion matrix is [[4 0] [0 5]] True Positives(TP) = 4 True Negatives(TN) = 5 False Positives(FP) = 0 False Negatives(FN) = 0.

- **Fourth Prototype**

The third prototype represents the question categories resulting from the Gaussian mixture modeling clustering based on BERT embedding. Similar to prototypes 1, 2, and 3, we used the same technical technologies here to implement the user interface form for this prototype. Figure-10 portrays the user-interface. Table-30 provides the resultant clusters for both the 90 and 37 question sets.

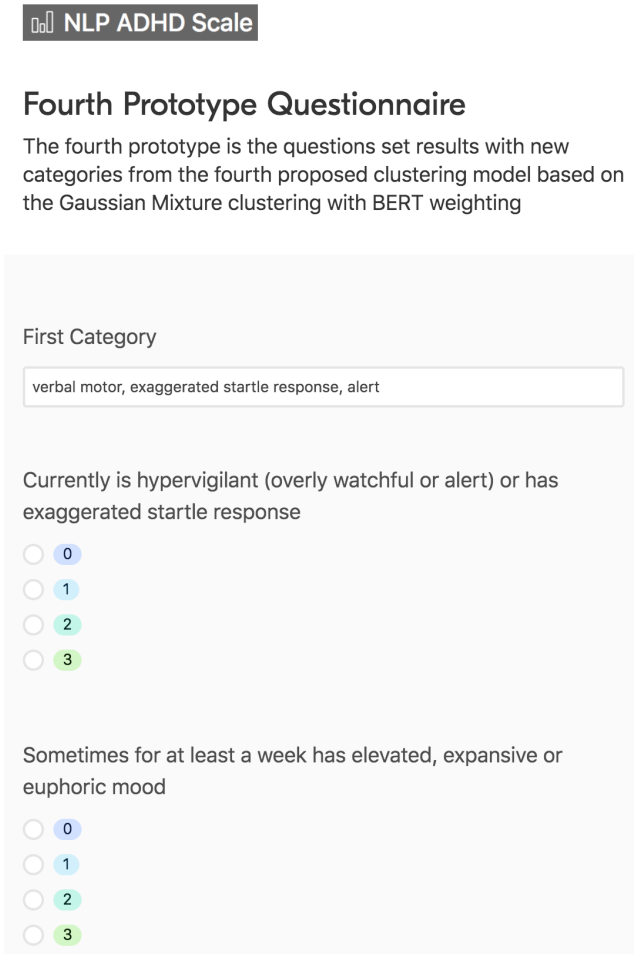


Figure 10: Fourth Prototype Model Web Interface

Table 30: Fourth Prototype Produced Clusters for both (90 and 37 Question Sets)

90 Question Set		37 Question Set	
Cluster Number	**Keywords**	**Cluster Number**	**Keywords**
1	verbal motor, exaggerated startle response, alert	1	difficulty, attention, schoolwork
2	repetitive motor behavior, reduce anxiety, inappropriate	2	excessively, home, impulses
3	week, grandiose behavior talkative	3	child, overactive, restless
4	classroom, assignment, psychomotor retardation	4	classroom, rule, interaction

We used independent-sample t-tests to compare children with ADHD and matched controls as summarized in Table-31 (re-clustered 90 questions vs. original SNAP-IV clustering) and Table-32 (re-clustered 37 questions vs. original SNAP-IV clustering). All results showed significant differences between children with ADHD and healthy controls.

Table 31: Fourth Model Comparison between the new categories and the baseline from the 90 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
verbal motor, exaggerated startle response, alert	1.64	0.38	0.15	0.09	0.0000
Irritable mood,	1.08	0.29	0.17	0.05	0.0003

aggressive, restless					
week, grandiose behavior talkative	1.89	0.19	0.12	0.06	0.0000
classroom, assignment, psychomotor retardation	1.86	0.58	0.11	0.08	0.0000
SNAP-IV categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Over activity	2.20	0.58	0.15	0.14	0.0000
Aggressive/Defiance	1.27	0.58	0.17	0.17	0.0085
ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

Table 32: Fourth Model Comparison between the new categories and the baseline from the 37 questions

The Statistical Measurement					
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
excessively drastically quickly	2.07	0.81	0.18	0.18	0.0000

classroom activity difficulty	2.21	0.35	0.10	0.05	0.0000
child, overactive, restless	1.46	0.32	0.11	0.05	0.0000
classroom, rule, interaction	1.87	0.19	0.14	0.06	0.0000
Produced categories	**ADHD Average Score**	**HC Average Score**	**ADHD Standard Error of the Mean**	**HC Standard Error of the Mean**	**T-Test (P-Value)**
ADHD-IN	2.41	0.45	0.10	0.08	0.0000
ADHD-H/Im	2.07	0.53	0.16	0.15	0.0000
ADHD-C	2.24	0.49	0.10	0.08	0.0000
ODD	1.45	0.75	0.22	0.11	0.0097
Inattention/Over activity	2.20	0.58	0.15	0.14	0.0000
Aggressive/Defiance	1.27	0.58	0.17	0.17	0.0085
ACADEMIC	2.10	0.11	0.17	0.05	0.0000
DEPORTMENT	1.34	0.15	0.18	0.10	0.0000

We used logistic regression analysis to assess the sensitivity/specificity of the classification as provided in Table 33.

Table 33: Performance Measurements for the Fourth Model

Measure	Value
Sensitivity	1.0
Specificity	1.0
Precision	1.0
F1 Score	1.0
Accuracy	1.0

The Confusion matrix is [[2 0] [0 4]] True Positives(TP) = 2 True Negatives(TN) = 4 False Positives(FP) = 0 False Negatives(FN) = 0.

5.3 Summary

In this chapter, we presented the model prototypes produced from our proposed system. The first prototype used the UMLS expansion terms approach helped in information and relationship discovery among terms of the questions, the original and expanded relationship terms formulate a knowledge graph provides a shared concept categorization for each cluster. Helped in generating meaningful clusters and helped in providing labels to the clusters according to the content of the clusters which is the shared medical parent concept, also in this prototype 90 questions set reduced into 40 questions and the 37 questions set reduced into 12 questions without harming the assessment procedure provides more precise clusters symptoms. The second prototype used Latent Dirichlet Allocation (LDA) Topic Modeling to extract topics for our corpus, we have tried different codes runs the best one is based on medical terms expansion and find out new extracted topics rely on the probabilistic distribution, the topics in second prototype was not strong like the first one because the topics here extracted based on the occurrence rather than the concept hierarchical relationships which provides less precise results. The third and fourth prototype used the pre-trained language module for embedding, the expanded medical terms features were not included in the pre-processing steps the predicted results was based on the bidirectional context term weighting from the language model rather than concept relationship, the predicted clusters was less concise due to our small corpus, the pre-trained

modules provided better results on a larger corpus. We used statistical analysis to evaluate the performance of our prototypes using data collected from children with ADHD and healthy controls. We used independent-samples t-tests to compare the groups, binary logistic regression analysis to evaluate the ability of our prototypes to correctly classify ADHD.

Chapter 6 – Conclusions and Future Work

This chapter summarizes our proposed approach, discusses its findings and contributions that we achieved by building the proposed system. In addition, it outlines the future extensions for the current version of our proposed system. The chapter is divided into two sections. Section 6.1 presents a discussion of the contribution of our research work and highlights the techniques/approaches that we utilize in the proposed system. Section 6.2 discusses future work and other challenges that we plan to tackle in future system updates.

6.1 Conclusions

Healthcare professional recommends answering questionnaire for symptoms rating scale to diagnose the ADHD disorder. one of the current questions set tool used to assess the screening is SNAP-IV rating scale. The problem is in the length and accuracy for such questionnaire diagnostic tools, there is significant repetition amongst the questions asked to patients and their parents in structured clinical interviews. The strength of our research was that the work on semantics-based clustering of ADHD questions is new and we Aim to investigate the impact of utilizing knowledge captured by medical knowledge resources in this context. For this we proposed Topic Modeling and ADHD Questions Clustering Techniques, we have used unsupervised clustering techniques in obtaining new patterns, to cluster the SNAP-IV questions into more concise and semantically related groups we proposed four models:

First, we have constructed a knowledge graph that links n-gram tokens both extracted from

question texts and the medical knowledge resources using the parenty/meronymy relationships. In this context, each sub graph represents a group of questions with a shared parent concept in a semantics-based hierarchical structure. Here we have used a term reweighting algorithm for the terms express the same medical concept we have re-assigned their weights to be grouped under the same class. Second, we employed LDA topic modeling technique. The goal of this step is to identify the various topics under which each question term can be classified. Third and Fourth models, we utilized two different models, namely K-means and Gaussian Mixture clustering. The goal of utilizing both techniques is to group similar questions under different categories, without the help of any knowledge resource. Our goal here was to demonstrate how the exploitation of medical knowledge resources can assist in improving the quality of such clustering approaches. We have implemented four prototypes and test each one of them with a real word completed ADHD dataset, we used independent-sample t-tests to compare the averages between children with ADHD and matched controls. All new produced categories for all prototypes showed significant differences between children with ADHD and healthy controls. We used logistic regression to assess model performance measurements. The newly proposed categories based on the semantically-enhanced medical terms re-weighting technique which is based on medical knowledge graphs has proved to produce a new categorization of questions that are close in their semantic distance and have less overlap questions to assist healthcare professionals in their diagnostic procedures.

6.2 Challenges and Future work

Although the conducted experiments showed promising results, there are potential improvements to the techniques presented in this research work. This can be summarized in the following points:

A. In the proposed NLP pipeline, for the Identified UMLS terms we can expand to add the derivations and synonyms from the UMLS specialist lexicon and classify the lexically similar questions to further discover the lexical patterns between questions set.
B. Enhance the clustering models based on pre-trained language models such as BERT/BioBERT to add the expanded medical enrichment relationships in the pre-processing step and to compare the results with the graph based term reweighting technique.
C. To use different distance measurements for the clustering techniques based on that to compare the results to find out the best distance measurement for the context.
D. In the new produced categories, new question text formulation based on severity, frequency and past history and the enrichment medical new produced class name to be included. So that the Healthcare professional can change the scoring criteria, knowing all these variables earlier, further analysis and predictions could be used to aid the screening measures.
E. Evaluate the new prototypes with large dataset and ask healthcare professionals to test and evaluate the prototypes.
F. Build a transformer pipeline that apply the clustering techniques into a multiple questionnaires at the same time.

G. Build neuroinformatics computational model based on the new categorization. Terms where have the most contribution in the produced categories is an input to feed cognitive data to build personalized treatment recommender system.

References

Ahmadi, A., Kashefi, M., Shahrokhi, H., & Nazari, M. A. 2021. Computer aided diagnosis system using deep convolutional neural networks for ADHD subtypes. Biomedical Signal Processing and Control, 63, 102227.

Aizawa, A. 2003. An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45–65.

Allahyari, M., & Kochut, K. 2016. Discovering Coherent Topics with Entity Topic Models. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 26–33.

Abdi, A., Idris, N., Alguliyev, R. M., & Aliguliyev, R. M. 2017. Query-based multi-documents summarization using linguistic knowledge and content word expansion. Soft Computing - A Fusion of Foundations, Methodologies and Applications, 21(7), 1785–1801.

Afsharizadeh, M., Ebrahimpour-Komleh, H., & Bagheri, A. 2018. Query-oriented text summarization using sentence extraction technique. *2018 4th International Conference on Web Research (ICWR),* 128–132.

American Psychiatric Association. 1980. Diagnostic and Statistical Manual of Mental Disorders (Third Edition).

Basile, P., Caputo, A., & Semeraro, G. 2014. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. 1591–1600.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267–D270.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.

Bussing, R., Fernandez, M., Harwood, M., Hou, W., Garvan, C. W., Eyberg, S. M., & Swanson, J. M. 2008. Parent and Teacher SNAP-IV Ratings of Attention Deficit Hyperactivity Disorder Symptoms: Psychometric Properties and Normative Ratings From a School District Sample. Assessment, 15(3), 317–328.

Batet, M., Sánchez, D., & Valls, A. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, 44(1), 118–125.

Colby, J. B., Rudie, J. D., Brown, J. A., Douglas, P. K., Cohen, M. S., & Shehzad, Z. 2012. Insights into multimodal imaging classification of ADHD. Frontiers in Systems Neuroscience, 6, 59.

Chiaramello, E., Pinciroli, F., Bonalumi, A., Caroli, A., & Tognola, G. 2016. Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *Journal of Biomedical Informatics*, 63, 22–32.

Díaz-Galiano, M. C., García-Cumbreras, M. Á., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. 2008. Integrating MeSH Ontology to Improve Medical Information Retrieval. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, & D. Santos (Eds.), *Advances in Multilingual and Multimodal Information Retrieval* (pp. 601–606). Springer.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs].

Doernberg, E., & Hollander, E. 2016. Neurodevelopmental Disorders (ASD and ADHD): DSM-5, ICD-10, and ICD-11. CNS Spectrums, 21(4), 295–299.

Dobrean, A., Păsărelu, C. R., Balazsi, R., & Predescu, E. 2021. Measurement invariance of the ADHD Rating Scale–IV home and school versions across age, gender, clinical status, and informant. Assessment, 28(1), 86-99.

Faries, D. E., Yalcin, I., Harder, D., & Heiligenstein, J. H. 2001. Validation of the ADHD Rating Scale as a clinician administered and scored instrument. *Journal of Attention Disorders*, 5(2), 107–115.

Ghiasvand, O. 2014. Disease Name Extraction from Clinical Text Using Conditional Random Fields. University of Wisconsin Milwaukee, Ph.D. Thesis.

Hall, C. L., Guo, B., Valentine, A. Z., Groom, M. J., Daley, D., Sayal, K., & Hollis, C. 2020. The Validity of the SNAP-IV in Children Displaying ADHD Symptoms. Assessment, 27(6), 1258–1271.

Harrison, A. G., Nay, S., & Armstrong, I. T. 2019. Diagnostic accuracy of the Conners' adult ADHD rating scale in a postsecondary population. *Journal of attention disorders*, 23(14), 1829-1837.

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 2333–2338.

Ibrahim, Z. A. A., Haidar, S., & Sbeity, I. 2019. Large-scale Text-based Video Classification using Contextual Features. *European Journal of Electrical Engineering and Computer Science*, 3(2), Article 2.

Kuang, D., & He, L. 2014. Classification on ADHD with Deep Learning. In *2014 International Conference on Cloud Computing and Big Data*, 27–32. IEEE.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., & Lawley, M. 2016. Information retrieval as semantic inference: A Graph Inference model applied to medical search. Information Retrieval Journal, 19(1), 6–37.

Kim, J.-W., Sharma, V., & Ryan, N. D. (2015). Predicting Methylphenidate Response in ADHD Using Machine Learning Approaches. International *Journal of Neuropsychopharmacology*, 18(11).

Kalra, M., Lal, N., & Qamar, S. 2018. K-Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data. In D. K. Mishra, M. K. Nayak, & A. Joshi (Eds.), Information and Communication Technology for Sustainable Development (pp. 61–70). Springer.

Killeen, P. R., Tannock, R., & Sagvolden, T. 2012. The Four Causes of ADHD: A Framework. In C. Stanford & R. Tannock (Eds.), Behavioral Neuroscience of Attention Deficit Hyperactivity Disorder and Its Treatment (pp. 391–425).

Kaoungku, N., Suksut, K., Chanklan, R., Kerdprasop, K., & Kerdprasop, N. 2018. The silhouette width criterion for clustering and association mining to select image features. *International Journal of Machine Learning and Computing*, 8, 69–73.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, btz682.

Li, S. (2015). K-groups: A Generalization of K-means by Energy Distance [Bowling Green State University].

Liu, Z. (2013). High performance latent dirichlet allocation for text mining (Doctoral dissertation, Brunel University School of Engineering and Design PhD Theses).

Logeswari, S., & Premalatha, K. 2013. Biomedical document clustering using ontology based concept weight. *2013 International Conference on Computer Communication and Informatics*, 1–4.

Liu, Y., Gu, Y., Nguyen, J. C., Li, H., Zhang, J., Gao, Y., & Huang, Y. 2017. Symptom Severity Classification with Gradient Tree Boosting. *Journal of Biomedical Informatics*, *75 Suppl*, S105–S111.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*.

Malik, A., & Tuckfield, B. (2019). Applied unsupervised learning with R: Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA. Packt Publishing Ltd.

Naseem, U., Khan, S. K., Razzak, I., & Hameed, I. A. 2019. Hybrid Words Representation for Airlines Sentiment Analysis. In J. Liu & J. Bailey (Eds.), AI 2019: Advances in Artificial Intelligence (pp. 381–392).

Nigg, J. T., Lewis, K., Edinger, T., & Falk, M. 2012. Meta-analysis of attention-deficit/hyperactivity disorder or attention-deficit/hyperactivity disorder symptoms, restriction diet, and synthetic food color additives. *Journal of the American Academy of Child and Adolescent Psychiatry*, 51(1), 86-97.e8.

Nylander, E., Floros, O., Sparding, T., Rydén, E., Hansen, S., & Landén, M. 2021. Five-year outcomes of ADHD diagnosed in adulthood. Scandinavian *Journal of Psychology*, 62(1), 13–24.

Neumann, M., King, D., Beltagy, I., & Ammar, W. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task, 319–327.

Núñez-Jaramillo, L., Herrera-Solís, A., & Herrera-Morales, W. V. 2021. ADHD: Reviewing the Causes and Evaluating Solutions. *Journal of Personalized Medicine*, 11(3), 166.

Pham, T., Tao, X., Zhang, J., Yong, J., Zhou, X., & Gururajan, R. 2019. MeKG: Building a Medical Knowledge Graph by Data Mining from MEDLINE. In P. Liang, V. Goel, & C. Shan (Eds.), Brain Informatics (pp. 159–168). Springer International Publishing.

Rashid, J., Adnan Shah, S. M., Irtaza, A., Mahmood, T., Nisar, M. W., Shafiq, M., & Gardezi, A. 2019. Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering. IEEE Access, 7, 146070–146080.

Reátegui, R., & Ratté, S. 2018. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Medical Informatics and Decision Making, 18(3), 74.

Reynolds, D. 2009. Gaussian Mixture Models. In S. Z. Li & A. Jain (Eds.), Encyclopedia of Biometrics (pp. 659–663). Springer US.

Shahapure, K. R., & Nicholas, C. 2020. Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748.

Schuyler, P. L., Hole, W. T., Tuttle, M. S., & Sherertz, D. D. 1993. The UMLS Metathesaurus: Representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, *81*(2), 217–222.

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., Clevenger, W., Davies, M., Elliott, G. R., Greenhill, L. L., Hechtman, L., Hoza, B., Jensen, P. S., March, J. S., Newcorn, J. H., Owens, E. B., Pelham, W. E., Schiller, E., Severe, J. B., … Wu, M. 2001. Clinical relevance of the primary findings of the MTA: Success rates based on severity of ADHD and ODD symptoms at the end of treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(2), 168–179.

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., ... & Wigal, T. 2012. Categorical and dimensional definitions and evaluations of symptoms of ADHD: history of the SNAP and the SWAN rating scales. *The International journal of educational and psychological assessment*, 10(1), 51.

Saiyad, N. Y., Prajapati, H. B., & Dabhi, V. K. (2016). A survey of document clustering using semantic approach. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2555–2562.

Salton, G., & Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

Tar, H.H., & Nyaunt, T.T. 2011. Ontology-based Concept Weighting for Text Documents. World Academy of Science, Engineering and Technology, *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 5, 991-995.

Vora, P., & Oza, B. 2013. A Survey on K-mean Clustering and Particle Swarm Optimization. *International Journal of Science and Modern Engineering*, 1(3), 24-26.

Weiss, M., Worling, D., & Wasdell, M. 2003. A chart review study of the inattentive and combined types of ADHD. *Journal of Attention Disorders*, 7(1), 1–9.

Weißer, T., Saßmannshausen, T., Ohrndorf, D., Burggräf, P., & Wagner, J. 2020. A clustering approach for topic filtering within systematic literature reviews. MethodsX, 7, 100831.

Wan, X. 2010. Towards a Unified Approach to Simultaneous Single-Document and Multi-Document Summarizations. *In Proceedings of the 23rd international conference on computational linguistics* (Coling 2010) (pp. 1137-1145).

Zhang, Y., Zhang, O., Wu, Y., Lee, H.-J., Xu, J., Xu, H., & Roberts, K. 2017. Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge. *Journal of Biomedical Informatics*, 75S, S129–S137.

Zhang, X., Jing, L., Hu, X., Ng, M., Jiangxi, J. X., & Zhou, X. 2008. Medical Document Clustering Using Ontology-Based Term Similarity Measures. *International Journal of Data Warehousing and Mining (IJDWM)*, 4(1), 62–73.

Appendix A: SNAP-IV Rating Scale

1. SNAP-IV Teacher and Parents Rating Scale:

For each item, select the box that best describes this child. Put only one tick per item.		NOT AT ALL (0)	JUST A LITTLE (1)	QUITE A BIT (2)	VERY MUCH (3)
1.	Often fails to give close attention to details or makes careless mistakes in schoolwork, work or other activities				
2.	Often has difficulty sustaining attention in tasks or play activities.				
3.	Often does not seem to listen when spoken to directly.				
4.	Often does not follow through on instructions and fails to finish schoolwork, chores, or duties.				
5.	Often has difficulty organizing tasks and activities.				
6.	Often avoids, dislikes, or is reluctant to engage in tasks that require sustained mental effort (e.g. schoolwork or homework).				
7.	Often loses things necessary for tasks or activities (e.g. toys, school assignments, pencils, books or tools).				
8.	Often is distracted by extraneous stimuli.				
9.	Often is forgetful in daily activities.				
10.	Often has difficulty maintaining alertness, orienting to requests, or executing directions				
11.	Often fidgets with hands or feet or squirms in seat				
12.	Often leaves seat in classroom or in other situations in which remaining seated is expected				
13.	Often runs about or climbs excessively in situations in which it is inappropriate				
14.	Often has difficulty playing or engaging in leisure activities quietly				
15.	Often is “on the go” or often acts as if “driven by a motor”				
16.	Often talks excessively				
17.	Often blurts out answers before questions have been completed				

18.	Often has difficulty awaiting turn				
19.	Often interrupts or intrudes on others (e.g., butts into conversations/games)				
20.	Often has difficulty sitting still, being quiet, or inhibiting impulses in the classroom or at home				`
21.	Often loses temper				
22.	Often argues with adults				
23.	Often actively defies or refuses adult requests or rules				
24.	Often deliberately does things that annoy other people				
25.	Often blames others for his or her mistakes or misbehavior				
26.	Often touchy or easily annoyed by others				
27.	Often is angry and resentful				
28.	Often is spiteful or vindictive				
29.	Often is quarrelsome				
30.	Often is negative, defiant, disobedient, or hostile toward authority figures				
31.	Often makes noises (e.g., humming or odd sounds)				
32.	Often is excitable, impulsive				
33.	Often cries easily				
34.	Often is uncooperative				
35.	Often acts “smart”				
36.	Often is restless or overactive				
37.	Often disturbs other children				
38.	Often changes mood quickly and drastically				
39.	Often easily frustrated if demand are not met immediately				
40.	Often teases other children and interferes with their				

	activities				

For each item, select the box that best describes this child. Put only one tick per item.		**NOT AT ALL (0)**	**JUST A LITTLE (1)**	**QUITE A BIT (2)**	**VERY MUCH (3)**
41.	Often is aggressive to other children (e.g. picks fights or bullies).				
42.	Often is destructive with property of others (e.g. vandalism).				
43.	Often is deceitful (e.g. steals, lies, copies the work of others or "cons" others).				
44.	Often and seriously violates rules (e.g. is truants, runs away or completely ignores class rules).				
45.	Has persistent pattern of violating the basic rights of others or major societal norms.				
46.	Has episodes of failure to resist aggressive impulses (to assault others or to destroy property)				
47.	Has motor or verbal tics (sudden, rapid, recurrent, nonrhythmic motor or verbal activity)				
48.	Has repetitive motor behavior (e.g., hand waving, body rocking, or picking at skin)				
49.	Has obsessions (persistent and intrusive inappropriate ideas, thoughts, or impulses)				
50.	Has compulsions (repetitive behaviors or mental acts to reduce anxiety or distress)				
51.	Often is restless or seems keyed up or on edge				
52.	Often is easily fatigued				
53.	Often has difficulty concentrating (mind goes blank)				
54.	Often is irritable				
55.	Often has muscle tension				
56.	Often has excessive anxiety and worry (e.g., apprehensive expectation)				
57.	Often has daytime sleepiness (unintended sleeping in inappropriate situations)				
58.	Often has excessive emotionality and attention-seeking behavior				
59.	Often has need for undue admiration, grandiose behavior, or lack of empathy				
60.	Often has instability in relationships with others, reactive mood, and impulsivity				
61.	Sometimes for at least a week has inflated self esteem or grandiosity				

62.	Sometimes for at least a week is more talkative than usual or seems pressured to keep talking				
63.	Sometimes for at least a week has flight of ideas or says that thoughts are racin				
64.	Sometimes for at least a week has elevated, expansive or euphoric mood				
65.	Sometimes for at least a week is excessively involved in pleasurable but risky activities				
66.	Sometimes for at least 2 weeks has depressed mood (sad, hopeless, discouraged)				
67.	Sometimes for at least 2 weeks has irritable or cranky mood (not just when frustrated)				
68.	Sometimes for at least 2 weeks has markedly diminished interest or pleasure in most activities				
69.	Sometimes for at least 2 weeks has psychomotor agitation (even more active than usual)				
70.	Sometimes for at least 2 weeks has psychomotor retardation (slowed down in most activities)				
71.	Sometimes for at least 2 weeks is fatigued or has loss of energy				
72.	Sometimes for at least 2 weeks has feelings of worthlessness or excessive, inappropriate guilt				
73.	Sometimes for at least 2 weeks has diminished ability to think or concentrate				
74.	Chronic low self-esteem most of the time for at least a year				
75.	Chronic poor concentration or difficulty making decisions most of the time for at least a year				
76.	Chronic feelings of hopelessness most of the time for at least a year				
77.	Currently is hyper-vigilant (overly watchful or alert) or has exaggerated startle response				
78.	Currently is irritable, has anger outbursts, or has difficulty concentrating				
79.	Currently has an emotional (e.g., nervous, worried, hopeless, tearful) response to stress				

80.	Currently has a behavioral (e.g., fighting, vandalism, truancy) response to stress				
81.	Has difficulty getting started on classroom assignments				
82.	Has difficulty staying on task for an entire classroom period				
83.	Has problems in completion of work on classroom assignments				
84.	Has problems in accuracy or neatness of written work in the classroom				
85.	Has difficulty attending to a group classroom activity or discussion				
86.	Has difficulty making transitions to the next topic or classroom period				
87.	Has problems in interactions with peers in the classroom				
88.	Has problems in interactions with staff (teacher or aide)				
89.	Has difficulty remaining quiet according to classroom rules				
90.	Has difficulty staying seated according to classroom rules				

2. Scoring Instructions for the SNAP-IV-C Rating Scale

The SNAP-IV Rating Scale is a revision of the Swanson, Nolan and Pelham (SNAP) Questionnaire (Swanson et al, 1983). The items from the DSM-IV (1994) criteria for Attention-Deficit/Hyperactivity Disorder (ADHD) are included for the two subsets of symptoms: inattention (items #1-#9) and hyperactivity/ impulsivity (items #11-#19). Also, items are included from the DSM-IV criteria for Oppositional Defiant Disorder (items #21-#28) since it often is present in children with ADHD. Items have been added to summarize the Inattention domain (#10) and the Hyperactivity/Impulsivity domain (#20) of ADHD. Two other items were added: an item from DSM-III-R (#29) that was not included in the DSM-IV list for ODD, and an item to summarize the ODD domain (#30).

In addition to the DSM-IV items for ADHD and ODD, the SNAP-IV contains items from the Conners Index Questionnaire (Conners, 1968) and the IOWA Conners Questionnaire (Loney and Milich, 1985). The IOWA was developed using divergent validity to separate items which measure inattention/overactivity (I/ O — items #4, #8, #11, #31, #32) from those items which measure aggression/defiance (A/D — items #21, #23, #29, #34, #35). The Conners Index (items #4, #8, #11, #21, #32, #33, #36, #37, #38, #39) was developed by selecting the items which loaded highest on the multiple factors of the Conners Questionnaire, and thus represents a general index of childhood problems.

The SNAP-IV is based on a 0 to 3 rating scale: Not at All = 0, Just A Little = 1, Quite A Bit = 2, and Very Much = 3. Subscale scores on the SNAP-IV are calculated by summing the scores on the

items in the subset and dividing by the number of items in the subset. The score for any subset is expressed as the Average Rating-Per-Item, as shown for ratings on the ADHD-Inattentive (ADHD-I) subset:

Question	Not At All	Just A Little	Quite A Bit	Very Much	Item Score
Makes careless mistakes			X		2
Can't pay attention				X	3
Doesn't listen				X	3
Fails to finish work			X		2
disorganized		X			1
Can't concentrate				X	3
Loses things		X			1
Distractible				X	3
Forgetful	X				0

ADHD-In Total = 18, Average = 18/9 = 2.0

A scoring template for DSM-IV subtypes of ADHD (In and H/Im), for ODD for the dimensions of the of the CLAM (I/O and A/D); and for the Conners Index are presented below:

ADHD-In	**ADHD-H/Im**	**ODD**	**I/O**	**A/D**	**Conners Index**
Q#1 **– Q#9**	Q#11-Q-19	Q#21-Q28	Q#4, Q#8, Q#11, Q#31, Q#32	Q#21, Q#23, Q#29, Q#34,Q#35	Q#4, Q#11, Q#21, Q#32 Q#33, Q#36, Q#37, Q#38, Q#39
Total In _____	Total H/Im ____	Total ODD ____	Total I/O ____	Total A/D ____	Total CI ____
Average ___	Average ___	Average ___	Average ___	Average ___	Average ___

Tentative 5% Cutoffs:

	Teacher	**Parent**
ADHD-In	2.56	1.78
ADHD-H/Im	1.78	1.44
ADHD-C	2.00	1.67
ODD	1.38	1.88

The items from Q#41 of the SNAP-IV Rating Scale are from other DSM-IV disorders which may overlap with or masquerade as symptoms of ADHD. In some cases, these may be comorbid disorders, but in other cases the presence of one or more of these disorders may be sufficient to exclude a diagnosis of ADHD. The SNAP-IV is not designed to be used in the formal process of diagnosing these non-ADHD disorders, but if symptoms on page 2 of the SNAP-IV receive a high ("Quite A Bit" or "Very Much") rating, then an assessment of the implicated non-ADHD disorders may be warranted. The DSM-IV Manual should be consulted to follow-up with an evaluation of these non-ADHD disorders. The DSM Codes and the page numbers in the DSM Manual are

specified below to help in the assessment of possible conditions which may exclude or qualify a diagnosis of ADHD. A referral to a psychiatrist or a clinical psychologist may be required.

Questions	Disorder	DSM
#41-#45	Conduct Disorder	(DSM 312.8, P. 85)
#46	Intermittent Explosive Disorder	(DSM 312.34, p. 609)
#47	Tourette's Disorder	(DSM 307.23, p. 103)
#48	Stereotypic Movement Disorder	(DSM 307.3, p. 121)
#49-#50	Obsessive-Compulsive Disorder	(DSM 300.3, p. 417)
#51-#56	Generalized Anxiety Disorder	(DSM 300.02, p. 432)
#57	Narcolepsy	(DSM 347, p. 562)
#58	Histrionic Personality Disorder	(DSM 301.50, p. 655)
#59	Narcissistic Personality Disorder	(DSM 301.81, p. 658)

#60	Borderline Personality Disorder	(DSM 301.83, p. 650)
#61-#65	Manic Episode	(DSM 296.00, p. 328)
#66-#73	Major Depressive Episode	(DSM 296.2, p. 320)
#74-#76	Dysthymic Disorder	(DSM 300.4, p.345)
#77-#78	Posttraumatic Stress Disorder	(DSM 309.81, p. 424)
#79-#80	Adjustment Disorder	(DSM 309, p. 623)

Finally, the SNAP-IV includes the 10 items of the Swanson, Kotkin, Agler, Mylnn, and Pelham (SKAMP) Rating Scale. These items are classroom manifestations of inattention, hyperactivity, and impulsivity (i.e., getting started, staying on task, interactions with others, completing work, and shifting activities). The SKAMP may be used to estimate severity of impairment in the classroom. It is important to note that many disorders may produce impairment in the classroom setting, not just ADHD. Therefore, this rating scale is presented last to the possible exclusion conditions (on page 2 of the SNAP-IV) will be considered in addition to the inclusion criteria for ADHD (on page 1 of the SNAP-IV). Both should be considered before interpreting the SKAMP measure of classroom impairment or attributing high ratings on the SKAMP to ADHD.

ACADEMIC	DEPORTMENT	Orienting	Maintaining	Directing	Attention to Other	Attention to Rules
Q#81 – Q#86	Q#87-Q#90	Q#81, Q#86	Q#82, Q#83	Q#48,Q#85	Q#87, Q#88	Q#89, Q#90
Total = _____	Total = _____	Total = _____	Total = _____	Total = _____	Total = _____	Total = _____
Avg. = _____	Avg. = _____	Avg. = _____	Avg. = _____	Avg. = _____	Avg. = _____	Avg. = _____

Appendix B: ScispaCy Recognition Details

ScispaCy Recognized multiple medical concepts per term. Each with confidence property score, and we took the highest probability medical concept if we have different scores and if all matched with 1 probability as shown in the Example below we took all the recognized medical concepts, all the terms belonged to SNAP-IV questions recognized, SciSpacy behaved non-deterministically**,** where it provides minor differences in the medical terms identification.

term <anxiety>:	Medical concepts:
	>('C0003467', 1.0)
	>('C0003469', 1.0)
	>('C1963064', 1.0)
	>('C4050613', 1.0)
	>('C4552735', 1.0)
term <distress>:	Medical concepts:
	>('C0231303', 0.9999998807907104)
	>('C4085211', 0.8169634938240051)
	>('C4321508', 0.7640883326530457)
	>('C0038435', 0.7615172863006592)
	>('C0723460', 0.7615172863006592)

Each medical concept/concepts from UMLS Metathesaurus belong into semantic group as shown below part of the semantic groups:

ACTI|Activities & Behaviors|T052|Activity
ACTI|Activities & Behaviors|T053|Behavior
ACTI|Activities & Behaviors|T056|Daily or Recreational Activity
ACTI|Activities & Behaviors|T051|Event
ACTI|Activities & Behaviors|T064|Governmental or Regulatory Activity
ACTI|Activities & Behaviors|T055|Individual Behavior
ACTI|Activities & Behaviors|T066|Machine Activity
ACTI|Activities & Behaviors|T057|Occupational Activity
ACTI|Activities & Behaviors|T054|Social Behavior
ANAT|Anatomy|T017|Anatomical Structure
ANAT|Anatomy|T029|Body Location or Region
ANAT|Anatomy|T023|Body Part, Organ, or Organ Component
ANAT|Anatomy|T030|Body Space or Junction
ANAT|Anatomy|T031|Body Substance
ANAT|Anatomy|T022|Body System
ANAT|Anatomy|T025|Cell
ANAT|Anatomy|T026|Cell Component
ANAT|Anatomy|T018|Embryonic Structure
ANAT|Anatomy|T021|Fully Formed Anatomical Structure
ANAT|Anatomy|T024|Tissue
CONC|Concepts & Ideas|T185|Classification
CONC|Concepts & Ideas|T077|Conceptual Entity
CONC|Concepts & Ideas|T169|Functional Concept
CONC|Concepts & Ideas|T102|Group Attribute
CONC|Concepts & Ideas|T078|Idea or Concept
CONC|Concepts & Ideas|T170|Intellectual Product
CONC|Concepts & Ideas|T171|Language
CONC|Concepts & Ideas|T080|Qualitative Concept
CONC|Concepts & Ideas|T081|Quantitative Concept
CONC|Concepts & Ideas|T089|Regulation or Law
CONC|Concepts & Ideas|T082|Spatial Concept
CONC|Concepts & Ideas|T079|Temporal Concept

The following terms from SNAP-IV questions are belong into |**T079**|Temporal Concept semantic group:

Temporal Concepts
often
sometimes
week
year
currently

These Terms are important for the diagnosis assessment. We classified and grouped the terms based on the semantic groups and studied the clustering outputs before and after the semantic grouping for an example with and without the temporal concepts.

Appendix C: UMLS Expansion Details

We have used Web API UMLS Metathesaurus for terms Identification. And the Local files UMLS Metathesaurus for expansion Identification. More than 10k which include all (Parents, Child's) terms for each medical concept retrieved from SNAP-IV questions are retrieved from the local UMLS database. For every term in the final pool, we retrieve semantic groups from the UMLS API.

We have developed the following expansion rules:

- Core (unique UMLS Terms)– without Expansion
- Core with Descendants (Childs medical concepts) expansion
- Core + Ancestors (parents medical terms) expansion
- Core (unique UMLS Terms) with Ancestors and Descendants (parents and Childs medical terms) expansion.

Below is an example of Medical Term expansion:


```
{
 "C1367457": {
   "name": "C1367457",
```

```
  "label": "",
  "ancestors": [

  ],
  "descendants": [

  ]
},
"C1154415": {
  "name": "C1154415",
  "label": "",
  "ancestors": [

  ],
  "descendants": [

  ]
},
"C0237236": {
  "name": "C0237236",
  "label": "",
  "ancestors": [
    "C0004063",
```

```
    "C0337267",
    "C0337266",
    "C0337241",
    "C1640801",
    "C0441471",
    "C2720507",
    "C4751223"
  ],
  "descendants": [
    "C0233957",
    "C0418502",
    "C0418503",
    "C0418504",
    "C0418505",
    "C0418506",
    "C0418507",
    "C0418508",
    "C0418509",
    "C0418510",
    "C0418511",
    "C0418512",
    "C0418513",
    "C0418514",
```

```
    "C0418515",
    "C0418516",
    "C0418517",
    "C0418518",
    "C0418519",
    "C0418520",
    "C0034668",
    "C0034668",
    "C0418501",
    "C0424412",
    "C0233954",
    "C5190679"
  ]
},
```


Figure C1 below is the feature matrix without medical terms without expansion:

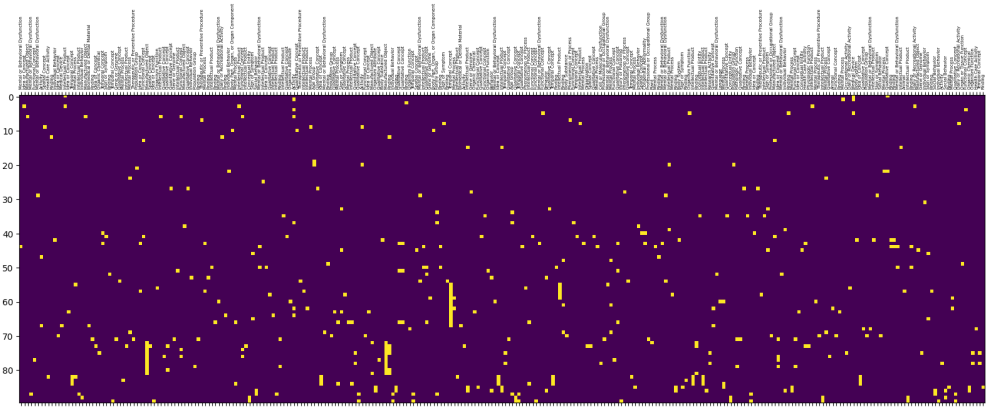


Figure C1: Features Matrix without medical terms expansion

In the x-axis the identified medical terms. Core (unique UMLS Terms)– without Expansion and in the y-axis the SNAP-IV questions. As shown we could not recognize much patterns such as shared medical concepts with questions, on the other hand in the below in the Figure C2 is the feature matrix contains the core (unique UMLS Terms) with Ancestors and Descendants (parents and Childs medical terms) expansion as observed we have additional features in the x-axis and more matched patterns with questions in the y-axis.

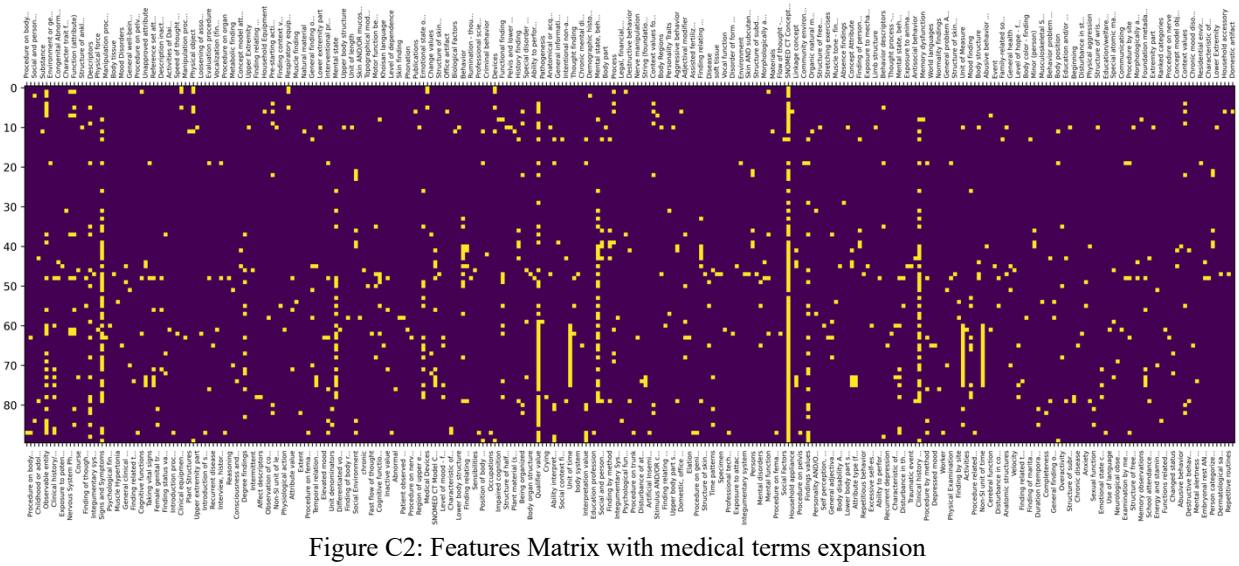


Figure C2: Features Matrix with medical terms expansion

Appendix D: Networkx Visualization Details

In order to draw the knowledge graph, we checked each pair of concepts (nodes) we checked if they belonged to the same medical class, the edge score increases if both belong, nodes grouped multiple times with common parent's medical concepts are having the highest scores and highlighted with viridis color scheme for edges in the graph, at the end we divide the score with the number of clusters, the output from silhouette analysis. We have added a threshold value to visualize the nodes within a certain threshold. The increase of threshold yields to have smaller graph representation. Below is the edge score algorithm:

```
# Calculate edge scores;
edgelist = []
for i, icui in enumerate(medical_concept_names):
   for j, jcui in enumerate(medical_concept_names):
      if j > i:
         score = 0.0
         for single_cluster_terms in group_terms:
            if icui in single_class_terms and jcui in single_cluster_terms:
               score += 1.0

         score = score / len(group_terms)
         if score >= threshold:
            edgelist.append((icui, jcui, score))

return edgelist
```

Figure D1 provides the graph representation without medical expansion terms.

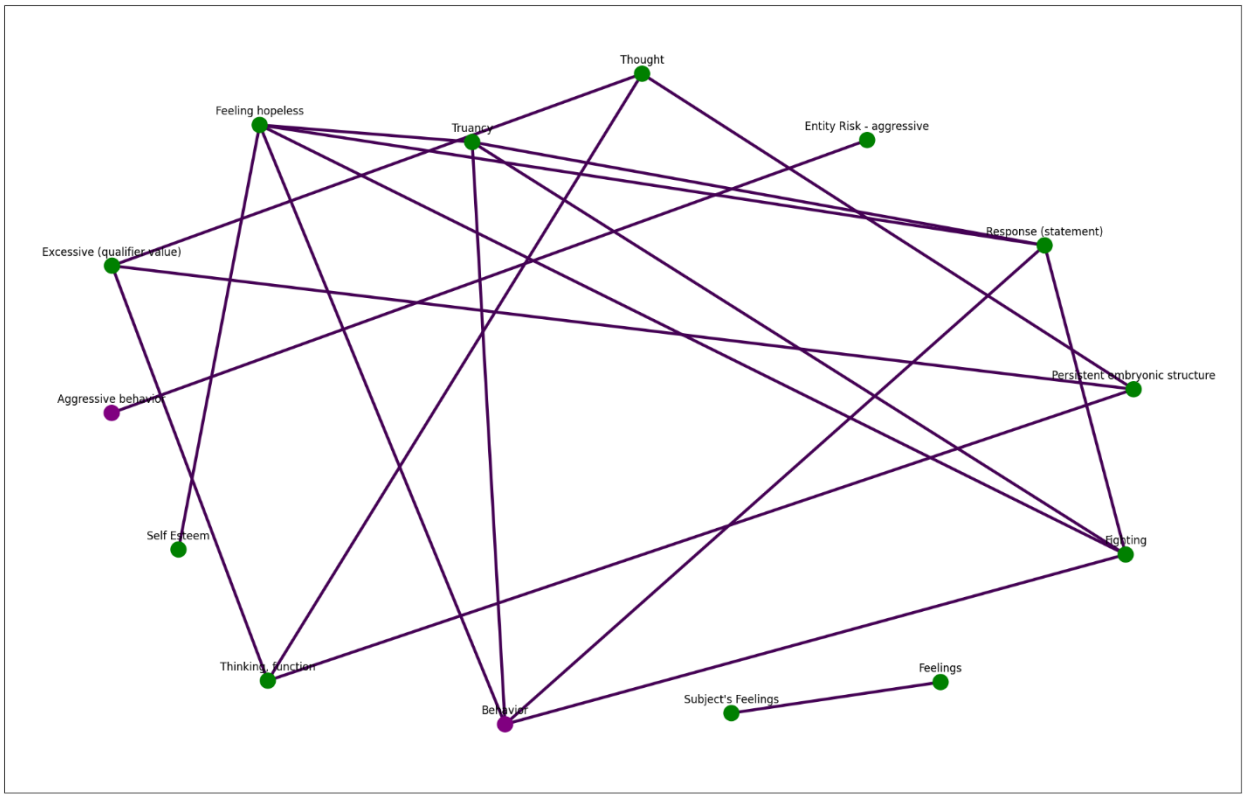


Figure D 1: Knowledge Graph without medical terms expansion

The graph visualization followed the spring layout type powered by the library. Figure D2 provides the graph representation with medical expansion (parents) terms and the nodes (terms) with higher scores (shared the same common parents) are highlighted with green and yellow colors which is an explicit representation for the predictable cluster-symptoms with new produced category class label, the common medical parent concept which our software implemented using python programming language automatically extracts them as cluster feature matrix and use distance

based clustering method, also we have distance matrix to represent clusters, but the clusters are not always clearly seen.

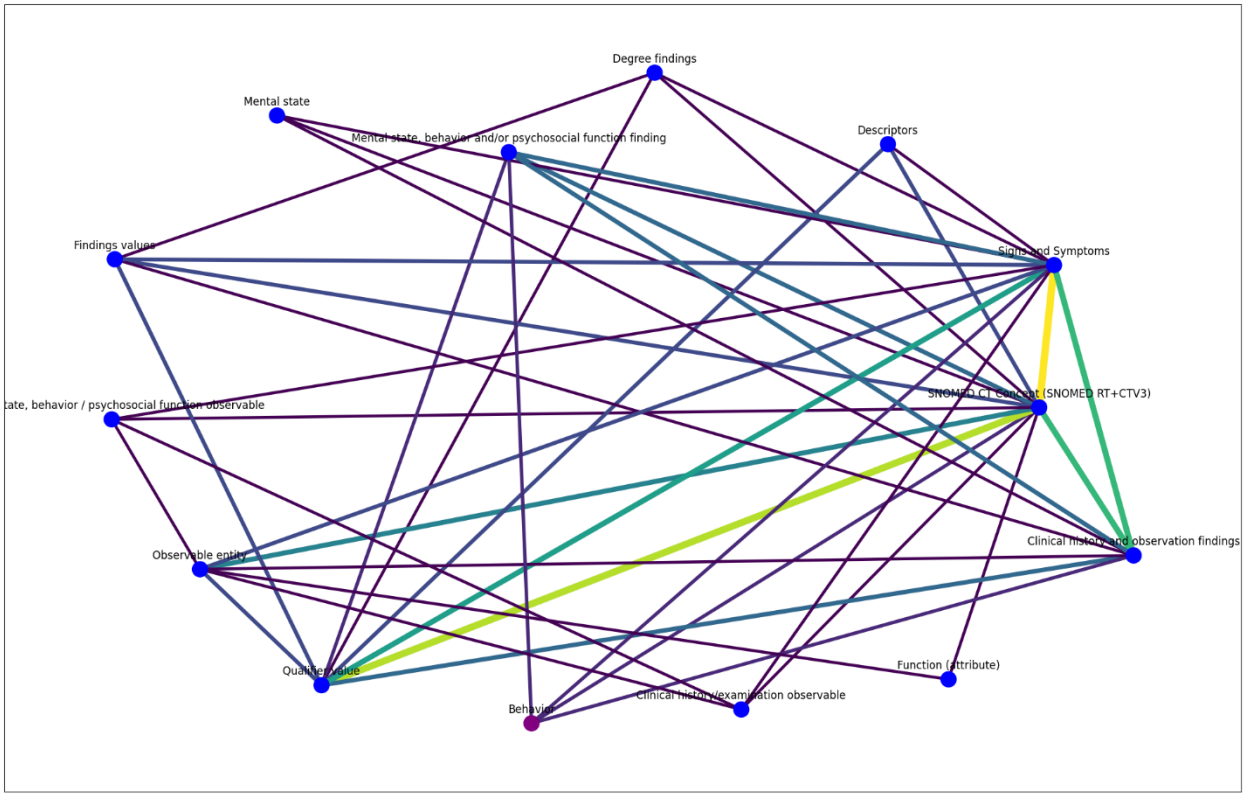


Figure D2: Knowledge Graph with parents medical terms expansion

Appendix E: t-SNE Visualization Details

The number of features after expansion is high, we have used a 2D t-SNE (T-Distributed Stochastic Neighboring Entities) visualization which lower the embedding dimensions to visualize clearly the data points (clustered terms). Figure E1 provides the data points visualization terms with expanded (parents medical concepts terms) and there are cleared grouping terms.

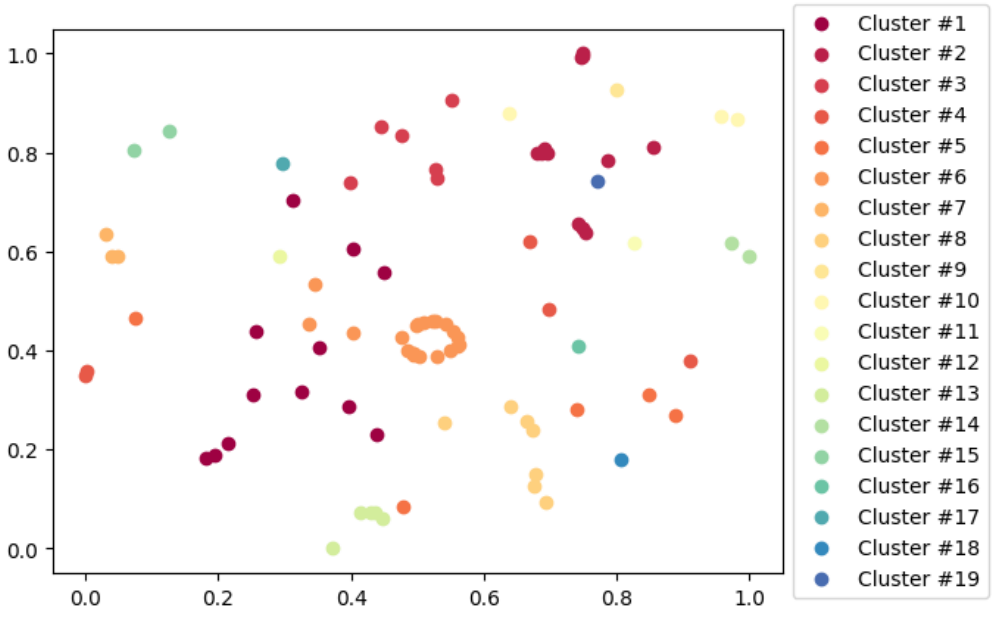


Figure E 1: t-SNE visualization with expanded terms

Figure E2 provides the data points visualization terms without terms and there are cleared overlapping between the data points. Both Figures before assigning the K-clusters from Silhouette analysis

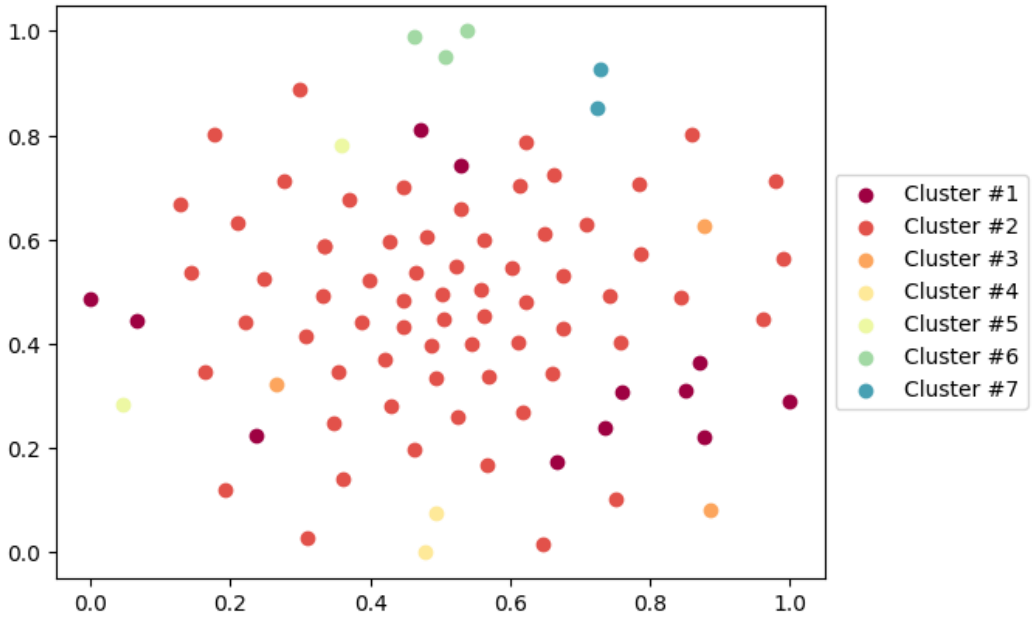


Figure E 2: t-SNE visualization without expanded terms

الملخص

إعادة تصنيف اسئلة تشخيص اضطراب فرط الحركة وقلة الانتباه "سناب ٤" الى مجموعات محددة ومختصرة عن طريق معالجة اللغة الطبيعية .

يعتمد تشخيص اضطراب نقص الانتباه وفرط النشاط في المراكز الصحية من خلال أداة مقياس للأعراض، هذه الأداة هي عبارة عن الاجابة على اسئلة استبيان لتحديد وقياس نسبة الاعراض، أحد هذه الاستبيانات المهمة في القياس يسمى مقياس "سناب الرابع" وهو نسخة محدثة عن "سناب" لمقياس الاعراض يتطلب الاجابة على ٩٠ سؤالاً، حيث يوجد تداخل في معاني الكلمات وتكرار كبير في الاسئلة، التصنيف الحالي للأسئلة تعتمد على مقاييس لغوية غير مستقلة وغير دقيقة. الهدف من هذه الدراسة هو تحسين مقياس "سناب الرابع" من خلال استخراج الدلالات من الكلمات المكونة في الاسئلة من خلال استخدام قواعد بيانات طبية من أجل ١) تقليل عدد الاسئلة ٢) زيادة الدقة عن طريق تجميع الاسئلة التي تتشارك بالمفاهيم الطبية ٣) تمكين المتخصصين في الرعاية الصحية من اتخاذ قرار بشأن الاحتفاظ بالأسئلة المتداخلة او ازالتها. من اجل القيام بعمل تصنيفات جديدة للأداة قمنا باستخدام اربعة انواع مختلفة من تقنيات التجميع: ١) لقد قمنا بإنشاء شبكة هذه الشبكة عبارة عن الكلمات المتواجدة في الاسئلة التي صنفت بالمصطلحات الطبية من خلال قاعدة بيانات طبية واضافة مصطلحات جديدة من قاعدة البيانات مرتبطة به ومن ثم تقسيم هذه الشبكة الى عدة شبكات (عناقيد) من خلال تجميع المصطلحات التابعة لنفس المسمى الطبي في الشبكة. ٢) لقد قمنا باستخدام تقنية استخراج مواضيع من الاسئلة هذه التقنية تعتمد على استخراج الكلمات الاكثر تواجدا من بين مجموعة من الاسئلة ٣) لقد قمنا باستخدام تقنيات التجميع من نوع التجزئة واستخدمنا تضمين للكلمات عن طريق شبكات لغوية مدربة مسبقاً لكي نقوم بالمقارنة مع استخدامنا للشبكة المبنية على المصطلحات الطبية المستخرجة والمضافة لقد قمنا ببناء اربعة نماذج بناء على الاربع نتائج وتطبيقها على بيانات مجابة للأسئلة وتم من خلالها تشخيص الاضطراب ومن خلال القيام بتحليل احصائي مبني على التحقق من ايجاد فروقات بين نتائج وجود الاضطراب او عدم وجودة بناء على التصنيفات الجديدة واثبتت النتائج بأنه يوجد فرق كبير حيث ان التصنيفات الجديدة قادرة على مساعدة

المختصين في عملية الكشف عن الاعراض بناء على تصنيفات جديدة واختصار لعدد الاسئلة كما تظهر نتائج التقنيات التي استخدمت للتجميع أن التجميع باستخدام المصطلحات الطبية المستخرجة من قواعد البيانات أثبت انه ينتج تصنيف موجز وجديد للأداة من خلال مصطلحات طبية متعارف عليها لدى المختصين لمساعدتهم في فهم وتحليل اضطراب فرط الحركة وقلة الانتباه.