The Role of Neurotransmitter Receptors in Mental and Behavioral Disorders: a Biomedical Text Mining Approach

Aliyu Kabir Musa

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the Degree of

Master of Science
in
Computer Engineering

Eastern Mediterranean University
June 2012
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Computer Engineering.

Assoc. Prof. Dr. Muhammed Salamah
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Computer Engineering.

Assoc. Prof. Dr. Bahar Taneri
Co-Supervisor

Assoc. Prof. Dr. EkremVaroğlu
Supervisor

Examining Committee

1. Prof. Dr. Hakan Altınçay
2. Assoc. Prof. Dr. Bahar Taneri
3. Assoc. Prof. Dr. Ekrem Varoğlu
4. Asst. Prof. Dr. Nazife Dimililer
5. Dr. İlmiye Özreis

ABSTRACT

Genetic variation in neurotransmitter receptors have been shown to be implicated both in behavioral variations across individuals in a given population and in various behavioral disorders. There are two aspects of synaptic neurotransmission in terms of its implications in behavioral disorders, both of which are important in healthcare management for such conditions. Firstly, particular allelic variations lead to an increased susceptibility to certain behavioral disorders. Secondly, specific allelic variations determine the response of affected individuals to available drug treatment options. Studies linking genetic variation and behavioral disorders are in general done on a single gene level and are focused on a single or a few disorders at a time. In this study, we aim to approach the relationship of neurotransmitter receptors to behavioral disorders from a different, more global perspective. We employ state-of-the-art text mining methods to put together a comprehensive database linking neurotransmitter receptors with specific mental and behavioral disorders. This study is unique in the sense that it provides this specific subset of gene-disease data. In addition, this tool is publicly accessible and enables researchers and healthcare professionals in the field to have easy access to a large amount of neurotransmission and disease data. This would facilitate analysis of the molecular bases of these conditions within a larger scope.

Keywords: biomedical text mining, machine learning, neurotransmitter receptors, mental disorders, behavioral disorders, gene-disease association, support vector machines.

ÖZ

Nörotransmiter reseptörlerindeki genetik varyasyonların kişilerin davranışlarına olan etkileri ve çeşitli davranışsal bozukluklarla bağlantıları gösterilmiştir. Sinaptik nörotransmisyonun davranışsal bozukluklara olan etkisi ve bu bozuklukların tedavisine yönelik önemi iki açıdan vurgulanabilir. Birincisi, çeşitli allelerdeki varyasyonlar, bazı davranışsal bozukluklar için risk faktörü oluşturabilir. İkincisi ise allelerdeki özgül varyasyonlar kişinin ilaç tedavisine nasıl yanıt vereceği konusunda belirleyicidir. Bu çalışmalar genelde tek bir gen düzeyinde olup, tek veya birkaç hastalığa odaklı olarak rapor edilmektedir. Bu çalışmada, nörotransmitter reseptörleri ve onların davranışsal bozukluklarla olan bağlantılarına farklı ve daha global bir açıdan yaklaştık. En güncel metin madenciliği yöntemlerini kullanarak, geniş kapsamlı bir veritabanı oluşturduk. Bu veritabanı reseptörleri, zihinsel ve davranışsal bozukluklarla birleştirmektedir. Kamuya açık olan bir veritabanı, araştırmacıların yüksek miktardaki nörotransmiter ve hastalık verisine kolayca ulaşıp, geniş kapsamlı analizler yapmalarını sağlamaktadır.

Anahtar kelimeler: biomedikal metin madenciliği, makineye dayalı öğrenme, vector destek makinaleri, nörotransmiter reseptör, zihinsel davranış bozuklukları

To My family

ACKNOWLEDGMENTS

I am heartily thankful to my supervisors, Assoc. Prof. Dr. Ekrem Varoğlu and Assoc. Prof. Dr. Bahar Taneri for their encouragement, help and guidance throughout this study. They devoted their time for helping me to explore knowledge in molecular biology and biomedical text mining with many motivation and supervision.

I would like to extend my gratitude to my monitoring jury members; Prof. Dr. Hakan Altınçay, Asst. Prof. Dr. Nazife Dimililer and Dr. İlmiye Özreis, I have to thank them for the time they take to critically reviewed my work and provide me with useful suggestions.

I would like to thank my staff and colleagues from the Computer Engineering Department of Eastern Mediterranean University, for their help during my studies and publication of this thesis.

I owe a lot to my friends, Uzairu Umar Saleh, Hassan Hamisu Dankaka, Yusuf Yahya, Auwal Yahya, Abdulaziz Musa and Alaa Ali Hamid for their friendship, concern and moral support during my thesis work.

I owe my deepest gratitude to my parents, brothers and sisters who undoubtedly have given me the support no one can ever give me, they have a special place in my heart and indeed I will always be proud them. I will also use this opportunity to thank my uncle, Alh. Yahya Lawal for his generous support and help throughout my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OFABBREVIATIONS

BOW	Bag-of-words
C	SVM Regularization Parameter
CBioC	Collaborative Bio Curation
Cl-	Chloride
CRFs	Conditional Random Fields
DSMIV-TR	Diagnostic and Statistical Manual of Mental Disorders Fourth Edition Text Revision
ENT	Entity
EPSP	Excitatory Postsynaptic Potential
GN	Gene Normalization
GO	Gene Ontology
HMM	Hidden Markov Model
IAT	Interactive Task
IDF	Inverse Document Frequency
IE	Information Extraction
IR	Information Retrieval
K+	Potassium
KDD	Knowledge Discovery
MINT	Molecular INTeraction Database
ML	Machine Learning
Na+	Sodium
NCBI	National Center for Biotechnology Information
NER	Named Entity Recognition
NTreceptorDB	Neurotransmitter Receptor Database
OMIM	Online Mendelian Inheritance in Man
OPHID	Online Predicted Human Interaction Database
POS	Part of Speech
PPI	Protein–Protein Interaction
PSD	Post-Synaptic Density
SVM	Support Vector Machine
TF	Term Frequency
x^2	Chi-Square

Chapter 1

INTRODUCTION

1.1 Background

Text mining has recently gained popularity as a method of knowledge discovery from textual database sources [1], which covers the task of mining interesting non-trivial and new knowledge or patterns of information from unstructured text documents.This domain has shown potential to be useful in different applications and to mine knowledge in the literature within biological databases. Rapid accumulation of high-throughput biomedical data presents oppotunities and at the same time challenges for data integration and interpretation.

The main goal of the post genome era is to further elucidate the role of genetics in human health and diseases [2]. The current amount of biomedical literature regarding the identification of disease genes is rapidly increasing. One of the main challenges researchers in this domain face is that, most of the relevant information are buried in the articles, in the form of unstructured text. It is clear that text mining models are essential for handling large amount of information that is available only in unstructured textual form. For example, databases such as PubMed with 21.7 million records to date [2] contain large amount of data, which can be transformed into a representable way in order to facilitate and improve biological knowledge transfer and data analysis. In particular, text mining involves the analysis of large collection of documents, with the aim of extracting specific information such as relationships

and patterns hidden in text collections. Newly growing research areas, which use machine learning approaches to extract new knowledge from complex databases during the last decade, have drawn valid scientific attention [2]. In the recent years, mining relations between genes and diseases in the text have become a major aim for researchers [3]. Therefore, efforts are underway to use several machine learning techniques in order to extract gene-disease relationships from free text and link these entries into databases.

1.2 Thesis Contribution

In this thesis, we investigate the genetic variations in neurotransmitter receptors that are associated with certain behavioral disorders. This study focuses on finding the relationships between neurotransmitter receptors and behavioral disorders from text data, using a set of neurotransmitter receptors. We review here the available methodologies for the classification of neurotransmitter receptors that are associated with mental or behavioral disorders. A large set of search terms for all known neurotransmitter receptor families is generated and used to retrieve a large number of relevant articles from PubMed, in order to perform text mining on the abstracts of relevant articles and identify the mental and behavioral disorder associations of given neurotransmitter receptors.

We generate a comprehensive search term set for each neurotransmitter receptor listed in the 2008 article by Iwama and Gojobori [4], since a gene may appear in an abstract by its symbol, name or even by its description. In order to populate this list a pipeline is used to access the Gene DB of NCBI [5] using the Entrez Programming Utilities. We use the Diagnostic and Statistical Manual of Mental Disorders (DSMIV-TR) [8] as the input list for our mental and behavioral disorder set.

We downloaded 835691 abstracts from PubMed corresponding to 1337 neurotransmitter receptor symbols, including names, description, other names and aliases. Since no annotated data on the neurotransmitter receptor-disorders is available we manually annotated a train data set of 570 sentences from the retrieved abstracts. To the best of our knowledge this is an original and first of its kind dataset specifically annotated in this domain. We believe that the dataset can be used in by others in the future for machine learning based systems. We train a Support Vector Machine (SVM) [9] using the generated train data set and test the classifier model on 5143 sentences. We identify 1517 unique association between the neurotransmitter receptors and mental disorders under consideration.

Using the association sentences, we have constructed a database containing neurotransmitter receptor-disorder association data, based on biomedical literature using the text mining approach. This database and the associated user friendly web interface enables storage and access of the relevant neurotransmitter receptor–disorder data. End users such as bioinformaticians, biologists, pharmacologists and biomedical researchers will be able to view annotations, search for biological data, validate linked resources, and create new information to apprehend new concepts as they arise. Furthermore, external sources such as ontologies, databases and dictionaries can be curated based on the database presented here.

To the best of our knowledge, to date majority of the proposed biomedical systems does not focus particularly on the gene-disease relationship associated with neurotransmitter receptors and behavioral or mental disorders. Therefore, our database is unique in itself. For the first time, it provides a specific source on neurotransmitter receptors and their associated disease conditions. Furthermore, the

newly created database which classifies the articles retrieved and links them based on the association of these two entities, presents a novel platform for the analysis of these particular diseases on a large scale.

1.3 Thesis Outline

This thesis is organized in six chapters. In Chapter 1, we provide an introduction to the thesis topic. In Chapter 2, we cover basic biomedical text mining concepts, new progress in the field, state-of-the-art work and we present a literature review. In Chapter 3, we discuss the genetic variation in neurotransmitter receptors, behavioral disorders and their relationships. We also point out the genetic variation in neurotransmitter receptors that leads to a specific mental or behavioral disorder. In Chapter 4, we present the data used and describe the methodologies used to populate our neurotransmitter receptors and disorder lists, and the classification scheme used to generate the associations between neurotransmitter receptors and mental disorders. In Chapter 5 results are given and discussed. Lastly, in Chapter 6 a summary of the discussion on the results and future work are given. The details of the resources used are presented in the appendices.

Chapter 2

LITERATURE REVIEW

The growing demand for information from text have made researchers to render more computable forms of data and to crosslink the information with related biological databases [6]. This linkage has the potential to increase the connection between the annotations in biological databases and the supporting evidence in the literature [6] [10]. The biological databases available heavily rely on expert human curation, which requires that biomedical researchers read the relevant literature carefully, extract specific information and encode the extracted information into an entry in a database using ontology or a precise vocabulary [10]. It is important to note that biological text mining is continuously gaining the interest of researchers in academia as well as in many web based consumer applications [10].

2.1 Overview of Text Mining

Considering the large amount of information available in unstructured textual form, computers can no longer process and handle this vast amount of information, because a computer usually processes text as a simple string of characters. Therefore, to extract meaningful information, careful pre-processing methods and procedures are required [6]. Text mining can be seen as a process of extracting generally new interesting information and knowledge from unstructured text. It is one of the new research topics that are related to many other fields of study. For example, text mining has a relation with many research areas such as machine learning,

information retrieval, statistics, computational biology and more especially data mining [6]. However, text mining is different from data mining, which discovers information from structured data.

Knowledge discovery in databases (KDD) is the process of extracting useful patterns in databases. Many processing steps have to be applied iteratively in order to mine the information from the databases under consideration. Applications of numerous steps mostly need interactive feedback from the user [11] [12]. Data preparation is one of the most difficult and time consuming tasks in the initial problem of analysis and understanding a KDD task. This is also known as data pre-processing task. Therefore, text mining algorithms require data preparation which needs special processing methods to convert data into a suitable text format [12].

Moreover, in closing the process of the data mining algorithms cycle, the data preparation step is the major interest of text mining approaches. It usually involves the creation of new methods or alteration of the existing algorithms for estimating the model obtained as well as the implementation of the new application [11] [12] [13]. In KDD, the data analysis aims at discovering hidden patterns and connection in data sets. Using this approach we can find the available facts present in the database. The same is also true for a simple text file [12]. We can use comprehensive measures to find the quality of the patterns discovered in the data in order to know the validity in the context of using statistical measures, originality and the accuracy of the patterns found [13]. Furthermore, there are different methods used to discover new patterns as well as coming up with comprehensive models that represent the relationships discovered in the data [13]. The discovered evidences found for an application is

used for the benefit of the user. Therefore, exact the meaning of KDD is associated with the specific application under concern. Figure 2.1 shows an overview of a text mining concepts in biomedical domain.

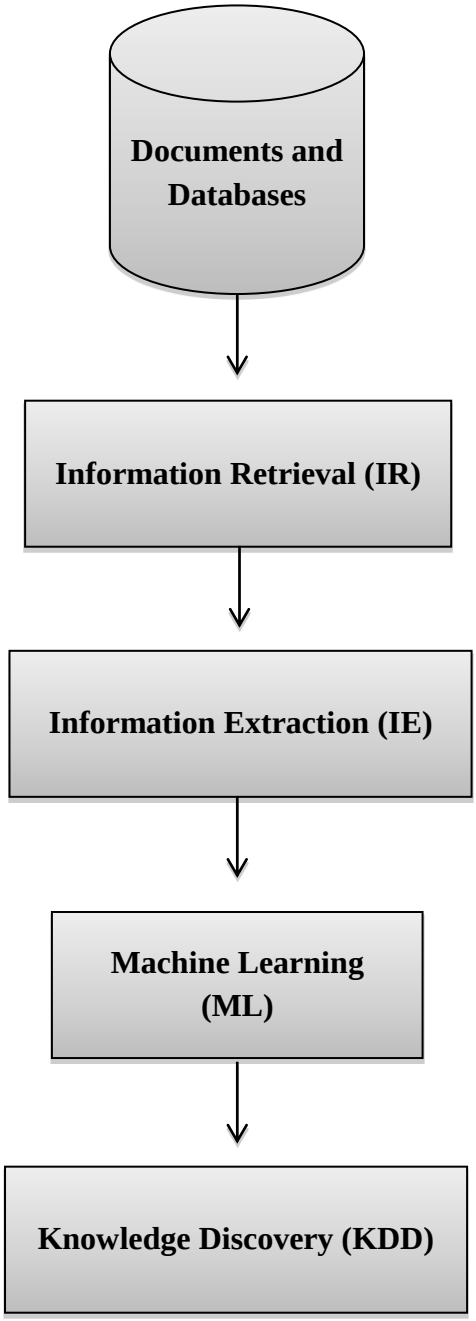


Figure 2.1: An overview of text mining concepts used in biomedical domain.

The detail of each stage is given below:

1) **Documents and Databases:** Databases and documents are sources used to store text that is used as input for most text mining tasks.
2) **Information Retrieval (IR):** This is a field that includes searching and collecting relevant documents from large amount of document collections based on user the user's information need.
3) **Information Extraction (IE):** This involves extraction of relevant information from unstructured sources in order to come up with new approaches for analysis, querying, and organization of data.
4) **Machine Learning (ML):** This is generally seen as design of computer programs to be able to find patterns, regularities, or rules from past experiences for classification, clustering, regression or prediction tasks.
5) **Knowledge Discovery (KDD):** This is the process of learning new and useful knowledge from a collection of data using computational tools. The discovered knowledge is often used for curation of databases.

2.2 Text Mining Preliminaries

Before a textual document can be mined for new knowledge, it needs to go through a number of preliminary steps. For most applications, all or many of these steps are required. Some of the methods used include techniques such as pre-processing, tokenization, filtering, lemmatization, stemming, index term selection and vector space representation of data.

2.2.1 Text Pre-processing

Working on a large amount of document collection is not suitable for further processing and mining a textual data. Therefore, it is essential to process the data and save it in a data structure. There are many methods that try to achieve the syntactic

and general structure of text for classification purpose, but text mining methods mostly are created based on the idea of representing a text document with a set of words [14]. The set of words in the documents are contained in a data structure, which is known as the bag-of-words (BOW). A vector representation is used in order to define the significance of a specific word inside a given document. For this purpose, a numerical value is given for each word for the representation. The probabilistic model [14], vector space model [15] and the logical model [16] methods are widely used methods.

2.2.2 Tokenization

The tokenization process is required for extracting all words in a given text document. A document is tokenized into a stream of words by taking out all punctuation characters and replacing tab characters and other non-textual characters with a white space character. The token representation is again used for further processing of the document in order to collect a set of unique tokens. The set of different tokens found in all the documents is called a dictionary of the corresponding document collection.

2.2.3 Filtering, Lemmatization and Stemming

In order to decrease the dimensionality and the size of a dictionary, filtering, lemmatization and stemming methods are used on the set of tokens extracted from the documents [17]. Filtering methods can be used to get rid of the words in the dictionary and the documents that occur excessively in the document. The most commonly used filtering method is the stop word removal, which uses the idea of removing and filtering of stop words that endure little or do not have any content information, prepositions, articles and conjunctions are examples of such words. Moreover, words that occur too frequently in the entire document are assumed to

have less information content in differentiating between the documents. Hence, such words can be removed from the dictionary [18] [19].

The main method used in lemmatization is to or try to put together the infinite tense to the singular form and verb forms to the nouns found in the document. However, the form in which the word is represented has to be known and the part-of-speech (POS) of each word in a document has to be given. Stemming approaches try to find the initial root of the words by taking out the plural form such as `s` that appear in nouns, the `ing` form in verbs, or other attached affixes from the rest of the POS. A stem is a root of word with equal meaning. After stemming is done, all the words in a document are changed to their root stem. The most popular stemming algorithm was originally introduced by Porter [19]. This algorithm defines a set of rules of products to repeatedly convert English words in a document into their original root stems.

2.2.4 Index Term Selection

For minimizing the number of words that can be used in documents, term selection algorithms or indexing approaches are widely used in text mining [19] [20]. Here, only the selected word terms can be used to describe the documents. Keyword selection can be done extracting keywords based on their entropy, which is the measure that quantifies the expected value of information contained in word collection.

2.2.5 Vector Space Representation

Vector space representation was initially presented for collecting information and indexing purposes [21]. However, vector space representation is applied in several text mining applications as well as in most of the presently available document retrieval systems. It has a less sophisticated data structure that does not use any

explicit semantic form of information. It enables very effective analysis of vast amount document collections.

The vector space model is used in transforming documents into numerical vectors of *m* dimensional space. Therefore, a document *d* is described by feature vector represented by numerical values of *m* dimension. For example, in IR applications documents can be processed by the use of vector operations and user queries that can easily be executed by encoding the query terms similar to the documents in a query vector. The query vector will then be compared to each document, and then a result list can be obtained by ordering the documents according to their computed similarity [22]. The main principle of the vector space representation for a given document is to compute an appropriate encoding of the feature vector. That is, deciding on the weights that should be used as the elements of the vector.

Every element of the vector usually shows the specific area of a word in the collected documents. The easiest way of transforming a document is to use binary vector of terms, that is, a vector element is set to value ‘one’ if the word corresponding it is used in the document and to ‘zero’ if the word is not in the given document. The vector dimension is defined by the number of words in the whole document collection [23]. This resulting encoding will be obtained in simple Boolean evaluation or even as a search term if a query is encoded. By using Boolean representation the significance of all terms for a precise query or comparison is considered as comparable terms. In order to take into account the frequencies of terms, usually certain term weighting methods are used, where each weight reflects the importance of a given word in a particular document of that collection. Several

weighting schemes can be used such as Term Frequency (TF), Inverse-document Frequency (IDF), TF-IDF, and Chi-Square (x^2) [24] methods.

2.2.6 Linguistic Preprocessing

Sometimes to further increase the performance of classification, linguistic preprocessing [25] may be used to improve the available information in the terms. Further preprocessing can be valuable in text mining methods, for this, the following approaches are frequently useful for linguistic preprocessing:

- **Part-of-speech tagging (POS):** This defines the part of speech tag in sentences, for example, noun, verb, adjective, etc. of the terms [25].
- **Text chunking:** It focuses on grouping neighboring terms in a given sentence, as noun phrases, verb phrases etc. [26].
- **Parsing:** This creates a full parse tree of an input sentence. As a result, we can discover the association of every term in a sentence, as well as its meaning in the sentence. For example, subject, object and so on [27].

Linguistic preprocessing typically uses a lexicon or different handcrafted rules as resources on the terms in text data. However, when a set of examples for training is presented, machine learning approaches such as Support Vector Machine (SVM) [9], Conditional Random Fields (CRFs) [28], Hidden Markov Model (HMM) [29] and Maximum Entropy Markov Model (MEMM) [30] can be used. It has been demonstrated that, for most text mining tasks linguistic preprocessing has proven to be less valuable compared to the simple BOW approach with basic preprocessing. One reason is that the co-occurrence of terms in a given vector model can be seen as an automatic disambiguation in classification methods [30]. Hence, enhancing the bag of words method has shown to improve the linguistic feature for text in clustering and classification base on recent studies [30] [31].

2.3 Overview of Classification Applied to Text Mining Tasks

The main reason why data mining techniques are applied to text documents is to organize and structure the content of the data for simplicity and easy access of the documents to the user. Prominent structures are book catalogs or library. The major difficulty of manually created indexes is the amount of time it is required to maintain them. Thus, they are very often not updated and are not usable for frequently changing sources of information such as the information on the Internet or modern publications. The current methods for organizing document collections either try to categorize documents or other entities based on the set of keywords (classification or categorization methods) or automatically structure the document entity collections to find groups of related documents (clustering methods). Figure 2.2 shows the diagram representation of classification methods applied to text mining tasks.

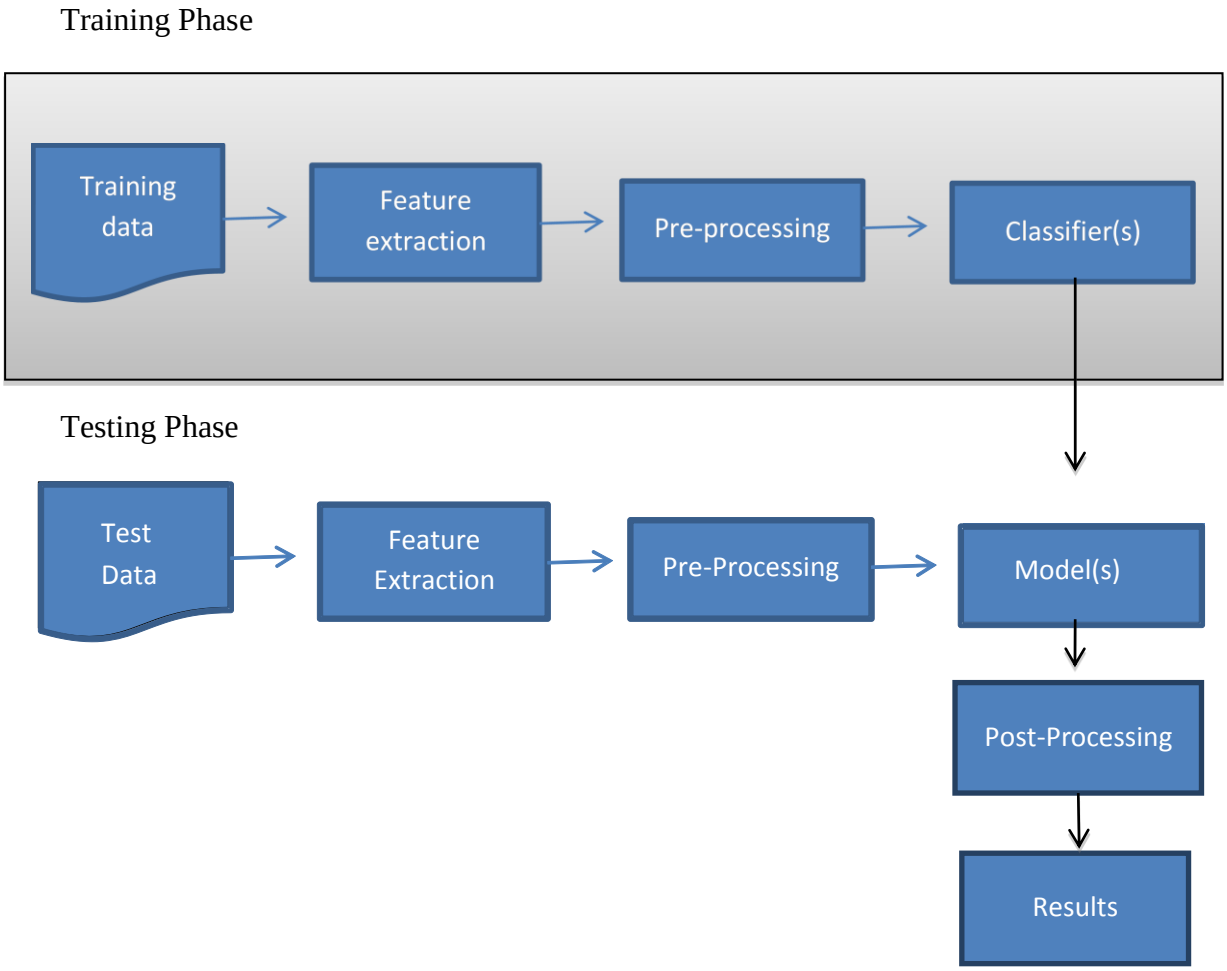


Figure 2.2: Main steps used in classification.

For example, text classification is generally used for assigning pre-defined classes to text documents [32]. It can be used to instinctively label every incoming news story in a newswire with a topic based on the categories such as “sports”, “politics”, or “art”. Regardless of the specific task, a text classification task begins with a training set of documents that are labeled with predefined class labels, in order to come up with a classifier model. The generated model is then used to label a new set of documents accordingly. To know the success of a classification task, an arbitrary part of the labeled documents is kept separately for testing the performance. This set is known as the test data set and it is not used in the training stage. We can classify the documents of the test data set with the classification model generated. Then we

compare the estimated labels with the true labels of the test data set to measure the success of the classification. An appropriately classified fraction of documents corresponding to the total number of documents is referred to as *accuracy* of the system and usually used as the first performance measure [33][34]. Accuracy can be computed using the confusion matrix in Table 2.1 below.

Table 2.1 Confusion matrix for measuring classifier performance

		Predicted Label	
		Positive	**Negative**
Train Label	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Eq. 2.1}$$

In some classification task, the target class uses only a small fraction of all available train samples, which result in a high accuracy since the target class is small and the other class is larger. Therefore, the performance may be 'misleadingly' high due to success in the negative class. Hence, different methods of measuring classification performance are used. For example, *Precision* computes the fraction of discovered documents that are assumed to be relevant which belong to the target class [34].

$$Precision = \frac{TP}{TP + FP} \quad \text{Eq. 2.2}$$

Recall on the other hand shows which fraction of the related documents is retrieved [33].

$$Recall = \frac{TP}{TP + FN} \quad \text{Eq. 2.3}$$

Almost all classifiers internally define some "degree of membership" in the resulting class. Often there is a tradeoff between precision and recall. Usually documents that have high score are marked as the selected class label if the precision is high. But, many significant documents can be ignored in the process, which results in a very low recall performance if the number of the document left out is high, sometimes the reverse may be true. However, if the search is in-depth during the measurement, the recall increases and the precision decreases. The F-score, which is the harmonic mean of precision and recall, is often used for measuring the overall performance of classifiers [34]. F-Score can be formulated as:

$$F - score = \frac{2 \text{ x } Precision \text{ x } Recall}{Precision + Recall} \quad \text{Eq. 2.4}$$

2.4 Related Work in Biomedical Text Mining Research

Biomedical articles in literature are growing rapidly. Now, over 21 million submitted articles are available in PubMed [35]. Hence, researchers find it hard to distinguish and curate this large amount of information available in the vast amount of biomedical text. Thus, the databases contently accessible may contain small portion of all the available information. Therefore, mining the available information from the vast amount of biomedical text has become a main task recently. Several biomedical text mining tasks have recently drawn the interests of researchers. Some of these tasks will be discussed next.

2.4.1 Biomedical Named Entity Recognition (NER)

Biomedical entity recognition can be considered as the first stage of every biomedical text mining task. This stage serves as a means of identifying and classifying meaningful keywords in the subject of molecular biology. For instance these entities contain the names of genes, proteins and their sites of action such as cells or organism names, drugs, diseases and chemical components. Named entity recognition has become increasingly essential with the very large growth in related results due to high-throughput experimental methods used. Hence, these methods can be used in several biomedical text mining tasks [36].

2.4.2 Gene Normalization (GN)

The gene normalization task supports a direct connection between genes and proteins mentioned in the text and available databases, by the use of a unique database identifier, to arrange information in the given databases. Thus, for that reason gene normalization tools are important to link existing online literature sources to meaningful databases. The main concern of Semantic Web is to integrate data and recent technologies for the biomedical field [36]. The task of gene and protein normalization is the crucial phase on the way to extract textual annotations for these entities. The major limitation of this task so far has been due to the fact that normalization was performed on abstracts that only focus on human genes [36]. Moreover, this limitation resulted in an artificial consequence, because in practice the disambiguation and how to link genes from different species is required for real applications. This is particularly important for human and mouse genes. Nevertheless, a controlled set up can be used to possible tear apart important aspects of gene normalization.

2.4.3 Protein-Protein Interactions (PPI)

The PPIs play an important role in biological events. Cycle control, cell metabolic, signaling pathways and disease pathways have proven to be vital for researchers lately. These relationships can lead to a complex networks known as PPI networks (PPIN). In such networks, the nodes show the proteins and the edges show the relationships between the pairs of proteins. Most of the recent graph theory based studies of PPIN mine the relationships from curated databases [36]. Recently, studies show that PPIN analyses are also constructed by mining the literature [36] [37].

PPINs have also gained popularity from researchers in predicting gene-disease relationships [37], For example, Chen *et al.* [38] used an Online Mendelian Inheritance in Man (OMIM) database to come up with an initial gene list from Alzheimer's, and built a collaboration network in order to mine the relationships of the matching proteins found on the Online Predicted Human Interaction Database (OPHID). They come up with function measure for the genes based on a graph. When constructing the linkage, only the relations between the initial genes and the relations of the initial genes and their counterparts were taken into account. The relationships between the neighboring genes were not considered. As reviewed by Gonzalez *et al.* [39] initial genes list of the CBioC database obtained were automatically mined and used to form a relationship network by bringing out the interactions of the original genes from the mentioned database [36] and curated databases such as BIND [34] and MINT [35]. In the study discussed in [16], they did not consider the interactions between the non-initial genes list. In order to remove all bias in the support of the initial genes, they developed a sophisticated role that allows just the associations with initial genes and putting together a level to influence the

gene under concern. All information is contained in OMIM or mentioned in the text that is related to the disease under concern.

2.4.4 Gene-Disease Associations

Most of the applications developed that use text mining methods to mine gene-disease relationships in the text use the co-occurrence statistics of genes and diseases. Recently, Adamic *et al.* [40] use a method that determines the presence of a gene in biomedical document that indicates a mentioned disease is statistically significant. Their approach was evaluated using breast cancer and a relevance of human-edited breast cancer gene database [40]. Al-Mubaid and Singh [41] conducted a similar study on gene-disease association. When a disease name is given, documents that have the mentioned disease name (documents marked as positive set) also the arbitrarily selection of document set (documents marked as negative set) are mined. The co-occurrence together with term frequency uses classification models from theories which are generally used to come up with the gene names that are ominously related with a disease mentioned. The researchers found 6 substantial genes related with Alzheimer's disease. The accuracy of the work was verified through articles retrieved from PubMed.

In other to find the genes that related with a particular disease, lab investigation is usually needed over a set of contender genes. Recent text mining methods make the use of databases and forecast gene-disease relationships by the use of similarities on keyword to extract genes disease. A typical illustration is the GeneSeeker [42], which is a system on the web that incorporates positional, and an expression or a phenotypic fact from 9 altered mouse and human databases with a summary of existing contender genes. The researchers reported their approach for 10 syndromes. Typically, 163 candidate genes list were compressed by system down to 22 genes

that still have the precise disease-gene relationship. According to Freudenberg and Propping, the system of grouping diseases according to their phenotypic similarities can be used in disease relationship of the index terms, which can be found in OMIM database [43]. Genes that qualify for a disease in a group are assumed to be selected by their functional similarity genes on the genes related with similar diseases in the group.

Chapter 3

NEUROTRANSMITTER RECEPTORS, MENTAL AND BEHAVIORAL DISORDERS

3.1 Neurotransmission and Synaptic Communication

Accumulating evidence shows the detailed molecular understanding of neurotransmitters, their receptors, and communication between the two [44]. Several investigations by researchers have yield to a draft outline of the overall molecular structure of the mammalian neuronal synapse. The complex nature of the synaptic proteome has over 1000 proteins. Mapping the organization of the synapse leads to a global view of the role of structure of the synapses and disease relations [45] [46].

Neurotransmission (or synaptic transmission) is the way in which neurons communicate by the movement of chemicals and electrical signals through a synapse [44]. In an interneuron, the function is to accept information as input from neighboring neurons across synapses in order to process that information, and to send it back as an output, to other neurons across the synapses since the neurons are connected to one another in a network [44]. A neural network (or network of neurons) is simply a collection of neurons that share information flow between neurons.

Chemical neurotransmission occurs at synapses. Pre-synaptic and post-synaptic neurons are divided by a small opening called the synaptic cleft. This is filled up with extracellular liquid. Even though, it is only a few nanometers, the synaptic cleft

generates a physical blockade for the electrical signal that is carried by one neuron to be relocated across to the other neuron in the network. Neurotransmitter functions in the cleft to overcome this electrical short form, it ensures that by acting as a chemical messenger within the network [47]. Neurotransmission is very important as it enables regions of the brain to interact with one another, in addition it facilitates all functions of the nervous system. Figure 3.1 shows an illustration of a synapse.

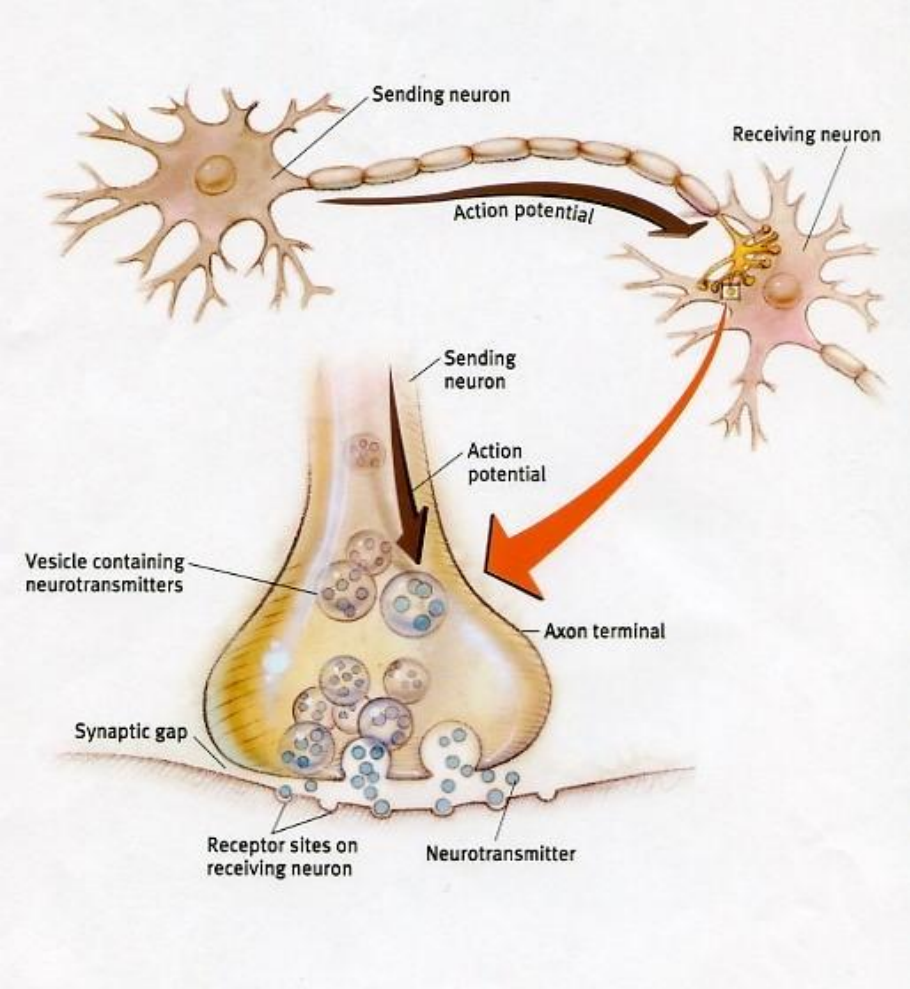


Figure 3.1: A diagram of the axon terminal and synapse adopted from [48].

3.2 Neurotransmitters

One of the interesting questions in science is the way in which the brain reads information from the accumulated senses and returns back with actions such as

thought, emotion or movement. This is done by the use of neurotransmitters in the neuron, which are chemical components that act as signals between neurons in the brain [49]. They usually act very rapidly, connecting up with molecules called neurotransmitter receptors, which reside on the dendritic surface of the neurons, waiting to bind neurotransmitter molecules. Examples of neurotransmitters include serotonin, a neurotransmitter which primarily affects arousal, mood, and sleep, dopamine a neurotransmitter which influences movement, learning, attention and emotion [50], acetylcholine, which is accountable for much of the stimulation of muscles, including the muscles of the gastro-intestinal structure, and GABA a neurotransmitter which influences mood and anxiety control [50].

The neurotransmitter receptors, which bind their neurotransmitters, act to convert chemical signals into electrical signals, which allow the recipient cell to react or not react, to become triggered or to stay silent. In order to interpret the chemical signals that come into the neuron, the neurotransmitter receptors must change shape extremely fast in response to the incoming neurotransmitter [51]. Electrical signaling between neurons is further discussed below.

3.3 Resting and Action Potentials

When a neuron is at rest, there is an electrical charge difference within the neuron, because of the relative concentration of positive and negative ions [51]. The ions discussed are sodium (Na+), potassium (K+) and chloride (Cl-) ions [51]. During the resting state, the neuron's inside is more negatively charged than its outside; therefore, the neuron is considered to be polarized in this state. Moreover, the neuron is ready for an action potential and is always ready for changing its electrical charge [51] [52].

An action potential happens when a neuron transfers information down the axon. The action potential is a discharge of an electrical movement that happens by depolarizing current. When the action potential reaches the end of the axon, it reaches the pre-synaptic terminal, therefore, message passed by the action potential go through the synaptic cleft in order to pass the message carried to the next neuron (or to a cell in the body) [51]. An electrical impulse carried by the action potential activates the release of neurotransmitter into the synaptic cleft to send to the dendrite on the neuron end that the message has to go to across the synapse. When neurotransmitter binds to its receptor, the neurotransmitter makes the neighboring neuron either more possible or less possible to trigger an action potential across its own axon [44].

3.4 Neurotransmitter Receptors

Neurotransmitter receptors are formed on the surface of postsynaptic cells and they bind ligand specific neurotransmitters. In addition, neurotransmitter receptor molecules are expressed on the pre-synaptic cells to deliver feedback process and reduce excessive neurotransmitter secretion [53]. Neurotransmitter receptors are mainly essential membrane proteins with seven membrane domains, commonly tied up to the G-proteins. A ligand binding by a specific neurotransmitter receptor may result in the initiation of a many cell signal transduction pathways [53] [54].

Neurotransmitters can either excite its neighboring neuron by binding to the specific neurotransmitter receptors on the post-synaptic neuronal membrane, which will increase its activity, or inhibit its neighbor neuron, which will suppress its activity [52]. Generally, the activity of a neuron relies on the stability among the number of

excitatory and inhibitory connections affecting it, and these can occur concurrently between neurons [53].

Most commonly known neurotransmitter receptors can be divided into two groups: ligand gated receptors and G-protein linked receptors [53].The incentive of a ligand-gated neurotransmitter receptor enables a passage in the neurotransmitter receptor to open and let the influx of chloride and potassium ions direct into the cell [53]. When positive or negative charges enter the cell it will either excite or inhibit the neuron cell. The non G-protein linked receptors for these neurotransmitter receptors include excitatory neurotransmitters, such as glutamate and, to a slighter extent, aspartate. Coming together of these ligands to the receptor yields an excitatory postsynaptic potential (EPSP) [53].

Alternatively, when the inhibitory neurotransmitter ligands come together, such as glycine and GABA, this yields an inhibitory postsynaptic potential (IPSP) [54]. These ligand-gated neurotransmitter receptors are also called inotropic or fast receptors. G-protein linked receptors are secondarily related to ion channels, via an alternative messenger system involving adenylatecyclase and G-proteins [54]. These receptors are not considered exactly as excitatory or inhibitory, but have moderate tolerance of the typical excitatory and inhibitory, neurotransmitters such as glycine and glutamate. These receptors can be seen as inhibitory if they are linked to the Gi-protein within cell membrane, nevertheless more like excitatory if they are linked to the Gs-protein [53]. These neurotransmitter receptors are known as slow receptors and are also called metabotropic. Examples of them include GABA-B, glutamate, serotonin (5-HT1A, 5-HT1B, 5-HT1D, 5-HT2A and 5-HT2C) and dopamine (D1 and D2) receptors [53].

3.5 Genetic Variation and Neurotransmitter Receptors

Genetic variation describes the certain genetic differences between individuals of the same species. Such variation allows survival of living things in a population in the framework of changing environmental conditions. Genetic variation in a population results from a wide variety of alleles [54]. The genetic variations in neurotransmitter receptors have been shown to be implicated both in behavioral variations across individuals in a given population and in various behavioral disorders [55]. Imbalances in neurotransmission can result in depression, anxiety and other mood disorders.

There are two aspects of synaptic neurotransmission and its implications in behavioral disorders, both of which are important in healthcare management for such conditions [55]. Firstly, certain allelic variations lead to an increased susceptibility to certain behavioral disorders [56]. Secondly, specific allelic variations determine the response of affected individuals to available drug treatment options [51].

3.6 Neurotransmitter Receptor-Disease Relationship

Allelic variation in synaptic neurotransmission has shown to be implicated in various behavioral and neurological disorders including depression [54], alcoholism [55], drug dependence [55], and bipolar disorder [56]. Abnormalities in the production of or functioning of certain neurotransmitter receptors have been linked with a number of diseases. Mal-functional neurotransmitter receptors particularly glutamate and dopamine neurotransmitter receptors is main cause of major biological underlying brain pathologies [55] [56].

In this thesis, we develop and employ computational tools to detect neurotransmitter receptor-disease association on a large scale, from accumulating biomedical literature data. As shown in Figure 3.2, the association of specific neurotransmitter receptor and various mental and behavioral disorders are mined from existing literature. As explained in detail in Chapter 4, the data uncovered is presented in a new database.

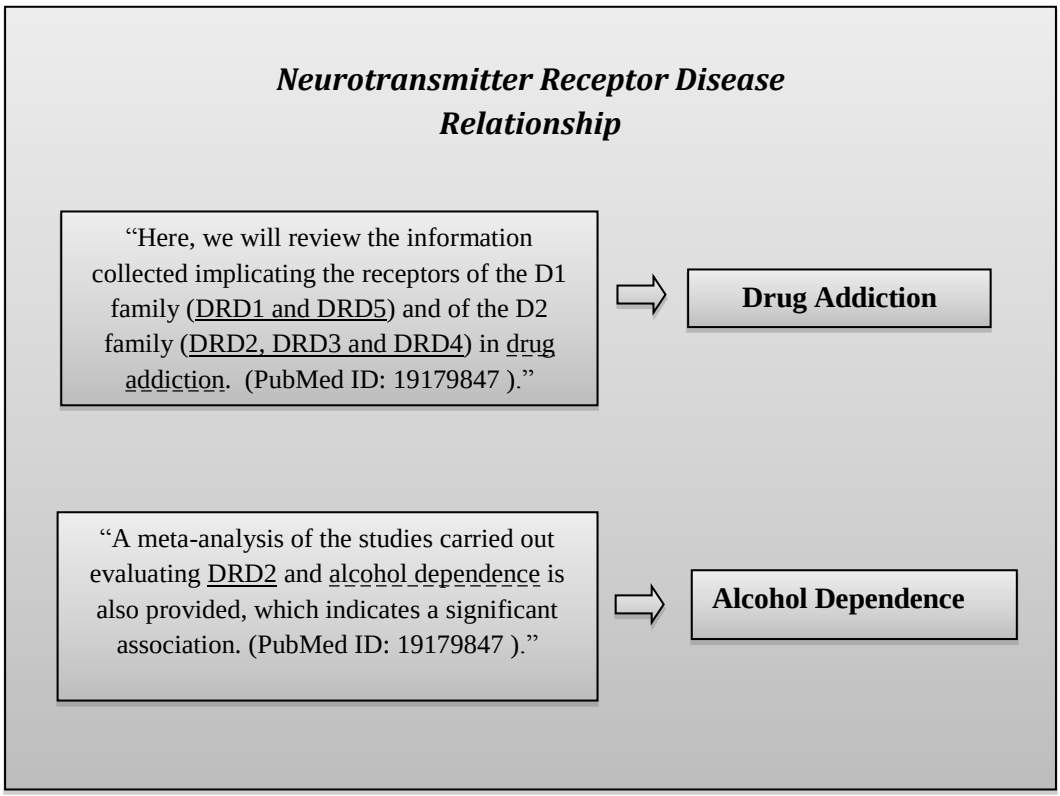


Figure 3.2: Text data highlighting dopamine receptors with associated behavioral disorders in literature.

Chapter 4

DATASET GENERATION AND METHODS USED

4.1 Overview

We use state-of-the-art text mining methodologies in order to extract associations between neurotransmitter receptors and behavioral disorder from biomedical documents indexed in the NCBI's PubMed [35]. The overview of the pipeline used is shown in Figure 4.1. The details of each step are given in the sections that follow.

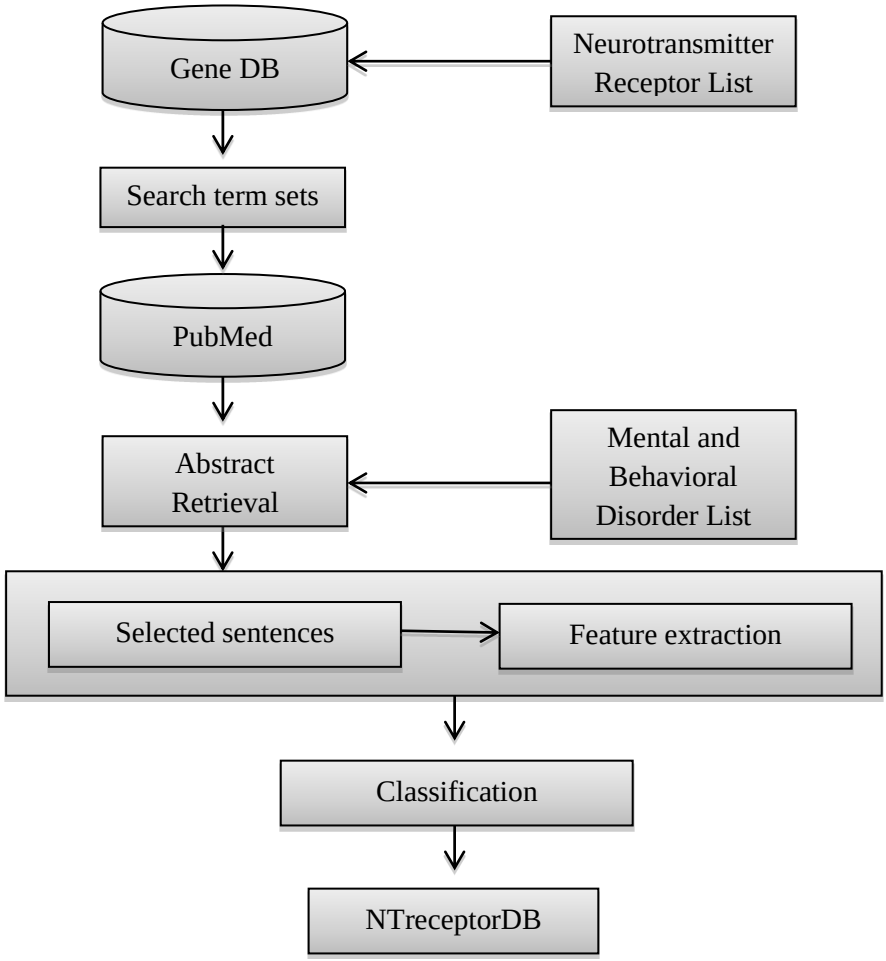


Figure 4.1: An overview of Text Mining Pipeline.

4.2 Neurotransmitter Receptors Search Term Set Generation

In their 2008 article, Iwama and Gojobori published an extensive review on different categories of neurotransmitter receptors [4]. Building upon the original list provided in this text, we generate a comprehensive search term set for neurotransmitter receptors. As shown in Table 4.1 a comprehensive search term set is generated for each neurotransmitter receptor provided in Iwama and Gojobori's list [4] since a gene may appear in an abstract by its symbol, name, alias or even by its description. In order to populate this list, a pipeline is used to access the Gene DB of NCBI [5] using the Entrez Programming Utilities. To construct the literature mined association, we performed keyword oriented searches against the Gene DB with the initial list of keywords.

A neurotransmitter receptor name can be mentioned by its synonyms. For example, DRD1 which denotes the dopamine receptor 1 neurotransmitter receptor, might appear as dopamine D1 receptor, D(1A) dopamine receptor, DADR, DRD1A, dopamine receptor D1, or D1A in biological text [54][58]. To standardize naming system of the genes, from our original list we represented all the neurotransmitter receptor by a single notation in the literature. We used the Gene DB to expand the keywords. We harmonized the tagged neurotransmitter receptor names against the official symbol, name, other names, and descriptions of the Gene DB. We combined each marked neurotransmitter receptor with its matching approved official symbol in the database.

4.2.1 Gene DB Queries

In order to expand the search term set of the related symbols in the original set, we performed a keyword search against the Gene DB using the initial set of official

symbols as an input parameter in formatted queries. The query term for each symbol is sent to the database encoded in a URL in order to retrieve the names, aliases, descriptions and other designations in the return parameters. The URL below shows an example of a query used to search for DRD1 symbol:

"http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=drd1&usehistory=y"

These queries are formed based on the NCBI standard, which provides external access of the data outside of the systematic web query interface. From the set of returned results, we manually selected only neurotransmitter receptor symbols that are already verified by experimental data [57].

Table 4.1: Example of a neurotransmitter family, available keywords and the new list of search term set.

Neurotransmitter Receptor Family	**Official Gene Symbol (initial list of keywords)**	**Official Gene Symbol with names, aliases, descriptions and designators (expanded list of keywords)**
Dopamine Rc	DRD1	DRD1 dopamine D1 receptor D(1A) dopamine receptor DADR DRD1A dopamine receptor D1 dopamine D1A receptor D1A
	DRD2	DRD2 dopamine receptor D2 isoform D2R D(2) dopamine receptor dopamine receptor 2 protein Drd-2 dopamine receptor 2 dopamine D2 receptor D2DR

4.2.2 Symbols and Synonym Generation

We pre-processed the newly generated list to further expand the search term set using the synonym generation method implemented previously by Kafkas *et al.* [59]. According to the criteria implemented by Kafkas *et al*., we generate new symbols that can be used to retrieve information about a neurotransmitter receptor family. We expand the list in such a way as to include all possible spelling differences and word forms like cooperation or co-operation and standardize or standardise. Using this method, keyword with spaces and symbol characters (i.e. "-") are used to generate a synonym by removing these set of character symbols to form a new keyword. We then manually analyzed a subset of the newly generated list and eliminated the possible symbols that are not related to the official gene symbol, in order to obtain an accurate set of keywords, which are associated to each neurotransmitter receptor family in the initial set [4].

4.3 Article Retrieval

Although there are many articles retrieved from PubMed, most of these articles may not be relevant for the problem under investigation. Therefore, the retrieval of articles directly related to the problem under investigation is a crucial step in automatically extracting neurotransmitter-disease relationships.

Using the generated list, each term in the search term set generated for a particular gene is submitted to the PubMed database in order to find the PMIDs of abstracts associated with that gene. Total of 835691 unique relevant abstracts were found for this study. Each abstract was retrieved using a query based on NCBI eUtility tools. The query form used is as follows:

"http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PubMed&term=Query&usehistory=y"

Each keyword was submitted as the Query term to obtain related abstract.

4.4 Disorder Search Term Set Generation

In order to have a comprehensive list of mental and behavioral disorders we refer to the Diagnostic and Statistical Manual of Mental Disorders Text Revision (DSM-IV TR) [8]. This list is used as a standard by researchers and even the legal system to describe and identify the types and thresholds of mental illness. We extract the list of diseases from DSM-IV TR and then we list the mental disorders that have been generally considered to be associated with a neurotransmitter receptor. Finally, we manually filtered the disorder list that we are going to use as our search terms.

4.5 Sentence Pre-Processing and Sentence Filtering

We selected the possible association sentences from the articles retrieved in the earlier stage of this work by firstly parsing the articles using the Stanford Sentence Splitter [27]. After splitting the articles into sentences, we chose those sentences with a co-occurrence of a neurotransmitter receptor and a mental disorder. Our assumption is that a sentence that describes a relationship between a neurotransmitter receptor and mental disorder should contain at least one neurotransmitter receptor and at least one mental disorder. We clean out all the sentences that did not have the co-occurrence of such entities. We found a total of 4642 sentences with at least one neurotransmitter receptor and at least one mental disorder mentioned. An example sentence which contains one neurotransmitter receptor and one mental disorder (PubMed ID: 20732371) is shown below in Figure 4.2.

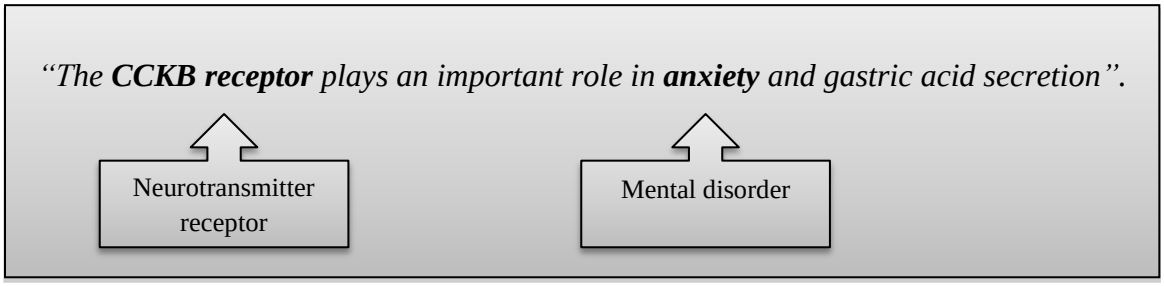


Figure 4.2: Example sentence with neurotransmitter receptor and mental disorder co-occurrence.

Table 4.2 shows the distribution ratio of the pair of neurotransmitter receptors and mental or behavioral disorders found within the sentences.

Table 4.2: Co-occurrence statistics of neurotransmitter receptors and mental or behavioral disorders in sentences.

Observance (Neurotransmitter : Disorder)	**Number of Sentences**
1:1	**3093**
1:2	**483**
2:2	**163**
2:1	**644**
3 or more : 3 or more	**259**

As it can be seen from the table, majority (67%) of the sentences contain only one neurotransmitter receptor and one disorder. A fewer number of sentences which contain more than one neurotransmitter receptor or disorder have been identified.

4.6 Feature Extraction

Feature extraction or selection is a topic that spans a large number of disciplines such as statistics, pattern recognition, text mining, and machine learning [60]. The main purpose of feature selection is to decrease the dimensionality of the feature space and remove noisy, irrelevant or redundant data. Decreasing the size of the feature space

generally effectively speeds up the learning algorithm [60]. Furthermore, careful selection of an optimal subset of features increases the performance or accuracy of the system [61]. The set of features used in this thesis is described in the following sections.

4.6.1 Bag-of-Words Feature

Bag-of-words (BOW) feature extraction is the process of transforming what is essentially a list of words into a feature vector that can be utilized by a classifier. Many classifiers use a dictionary style feature set, so we transform our text into a form of dictionary. The Bag of Words model is the simplest method; it constructs a word presence feature set from all the words of an instance. The idea is to convert a list of words into a dictionary, where each word in the corpus becomes a key with the value true [61]. In other words, the existence of each word from a corpus in the dictionary is marked as a '1' in the feature vector, when the binary representation is used.

In order to apply the supervised harmonic functions or the kernel based SVM methods, we need to define a similarity measure between two sentences. For this purpose, we use the bag-of-words feature representing the sentences. Unlike a syntactic parse, the bag-of-words usually sentence captures the semantic predicate dispute relationships among its words [61]. The idea of using bag-of-words for relation extraction in general was studied by Mooney and Bunescu [62]. To extract the relationship between two entities, they designed a kernel function that uses the 'words between', '3 words preceding the left entity' and '3 words following the right entity' of every sentence and constructed a bag-of-words (dictionary). The motivation is based on the observation that the shortest path between the entities

usually captures the necessary information to identify their relationship. We adapt the idea of Mooney and Bunescu to the task of identifying neurotransmitter receptor-disorder association sentences. Bag-of-words is believed to capture the relationship between two entities (i.e. neurotransmitter receptor and mental or behavioral disorder) in the sentences by including all words which contribute to the association under concern. In this study we assume that the neurotransmitter receptor and mental disorder names have already been mentioned in the sentences and focus instead on the task of extracting the association for a given pair of the entities in the sentences. For example, Figure 4.3 shows the bag-of-words feature we constructed for the sentence with PubMed ID: 20732371,

"*The **CCKB receptor** plays an important role in **anxiety** and gastric acid secretion*".

The words in the sentence between these entities are 'receptor', 'plays', 'an', 'important', 'role', and 'in'. Among these words 'receptor 'and 'in' are not likely to directly suggest an association between neurotransmitter receptor CCKB and anxiety disorder but the phrase 'plays an important role' clearly shows the relationship between them. Thus, the words in the bag-of-words between this pair give sufficient information to identify their relationship. In addition, the left word 'the' and the right words 'and', 'gastric' and 'acid' are used in the BOW representation.

Left: "*The"*

Middle: "*receptor plays an important role in"*

Right: "*and gastric acid"*

Figure 4.3: Example of BOW feature extraction from a sentence.

4.6.2 Association Words Feature

Association words are often used as a domain specific feature in order to extract associations between entities [63]. Here, the assumption is that sentences containing interaction words are more likely to describe an association between the entities. A list of interaction words that consists of 30 verb root words was gathered from the articles retrieved [63]. This list is provided in Appendix C. The presence of any interaction words in a candidate sentence is marked as an entry in the feature vector representation.

4.6.3 Lexical Features

Grammatical functions of the words in the textual data are known as lexicons. The lexical feature used in this thesis is part-of-speech (POS) tags. POS tag of a word describes if it is a noun, adjective, preposition etc. in the sentence. Since, the biomedical names mentioned in literature are lowercase, uppercase and expressive; the use of POS tags is likely to improve recognition performance particularly in identification of word boundaries.

The effect of POS tag feature has shown prominent improvements in biomedical domains by various experiments and different views on its effect have been shown [24][25]. Therefore, we decided to test the effect of this feature on neurotransmitter-disease association performance of our system. Since the training and test data did not include POS tags, we tagged both data sets using the Genia Tagger, which is a POS tagger, trained on both biomedical and newswire domains [24] [64].

4.7 Classification Using SVMs

Support Vector Machines (SVM) is a supervised machine learning approach which has recently been used in many text classification and text mining problems including the biomedical domain [2]. Figure 4.4 illustrate a simple classification system.

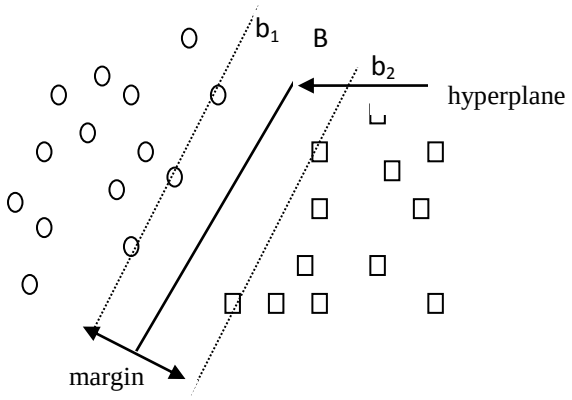


Figure 4.4: Example of Linearly Separable Binary Classification Problem.

SVM is a binary classifier which finds the optimal hyperplane separating the two classes by maximizing the margin between the hyperplane and the subset of training data points nearest to the hyperplane known as support vectors. Figure 4.4 shows the classification task. In the non-linearly separable case, SVMs use a kernel function that transforms the data into a higher where a linear hyperplane is used. In this study, SVM^{light} [9] is used throughout the experiments with a linear kernel setting. The feature vectors are composed of numerical values. The dimension of the feature vector is 12401 corresponding to all the word stems belonging to words represented in the BOWs of different sentences, the association words and POS tags feature. Figure 4.5 shows an example of feature vector representation used by SVM^{light}. The feature representation shows the target value that indicates the class of the example

+1 as the mark value shows a positive example, -1 a negative example respectively, as shown in the example below;

```
+1  73:0.1  86:0.41   129:0.51  271:0.14  321:0.1  597:0.1  650:0.4  672:0.2  912:0.3  945:0.1
-1  1:0.1   76:0.12   101:0.21  115:0.831 133:0.2  431:0.1  446:0.41 597:0.1  833:0.5  965:0.3
```

Figure 4.5: Feature vector representation of SVM^{light} classifier.

From the positive sample sentence for which feature number 73 has the value 0.1, feature number 86 has the value 0.41, feature number 129 has the value 0.51, respectively.

4.8 Training and Test Data Set Used

The train and test data sets used in this thesis are constructed manually by annotating randomly selected sentences from the set of abstracts retrieved. Dr. Bahar Taneri from Department of Biological Sciences Eastern Mediterranean University performed the manual annotation. The training set contains 570 annotated sentences with 479 positive and 91 negative sentences respectively. The test data on the other hand consist of 100 sentences with 55 positive and 45 negative samples. We summarized the data sets in Table 4.3.

Table 4.3: Summarizes the training and testing data obtained.

Data	**Class**	**Number of Sentences**
Train	Positive	**479**
	Negative	**91**
	Total	**570**
Test	Positive	**55**
	Negative	**45**
	Total	**100**

In order to find all possible associations in a sentence, if a sentence contains *n* different neurotransmitter receptors and one mental disorder, there are a number of hypothetical pairs of neurotransmitter receptor-mental disorders [62] [63]. Here, every sentence that contains *n* neurotransmitter receptors and *m* mental disorders is replicated as *n*x*m* pairs of candidate sentences. A sentence may be a relevant sentence for the existence of one neurotransmitter receptor and one mental disorder. For example, Figure 4.6 shows example of a replicated sentence. It contains 2 neurotransmitter receptors and 2 disorders. To reduce data sparseness, we use the entity pair under investigation to select only one occurrence of neurotransmitter receptor-disorder pair in a sentence and replace the rest with ENT symbol. So for our example sentence from article with PubMed ID: 749280, we have the following instances in the training set.

Original sentence:

"These data provide a potential role for **mGluR7** in **anxiety** and suggest that **mGluR8** may not be a therapeutic target for **schizophrenia**."

Replicated sentences:

"These data provide a potential role for **mGluR7** in **anxiety** and suggest that ENT may not be a therapeutic target for ENT."

"These data provide a potential role for **mGluR7** in ENT and suggest that ENT may not be a therapeutic target for **schizophrenia.**"

"These data provide a potential role for ENT in **anxiety** and suggest that **mGluR8** may not be a therapeutic target for ENT. "

"These data provide a potential role for ENT in ENT and suggest that **mGluR8** may not be a therapeutic target for **schizophrenia**."

Figure 4.6: Example of a replicated sentence.

Chapter 5

RESULTS AND DISCUSSION

5.1 Effect of Features Used

The features used for neurotransmitter receptor-disease association are explained in Chapter 4. In what follows, we present the effect of individual features as well as using features in concatenation on the results of the text mining approaches for mining these associations.

5.1.1 Bag of Words Feature (BOW)

As explained in section 4.6.1, the BOW feature sets include the nearest 3 left stem words, the nearest 3 right stem words and all the other stem words in between the two entities (i.e. neurotransmitter receptor and mental disorder entities) as illustrated Figure 4.3 for a given sentence.

Table 5.1 shows the performance of the classifier with the BOW features only. The performance of the proposed method is measured with 3-fold cross validation by using one neurotransmitter receptor and mental disorder occurrence criterion in a sentence. In addition, the regularization parameter C was set to 2 in order to control the hyperplane of the SVM^{light} used for separating the two classes. The value is selected based on the cross validation experiments. The experiment is repeated with different number of sentences in order to decide on the sufficient number of sentences for the train data. This task is necessary since all data is manually

annotated. Although it can be seen that some additional improvement in the classification performance can be obtained by including more sentences, we stopped annotation after no further significant gain in the F-Score value was achieved since the annotation process is a very time consuming event. The annotation of additional data will be dealt with as future work. We stopped annotation after no further significant gain in the F-Score value was achieved. An F-Score of 94.40% was achieved where it was observed that the recall of the system is better than the precision.

Table 5.1: 3-fold cross validation results using BOW feature

Sentences	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)
50	58.89	72.23	64.80	60.42
100	67.37	79.40	72.87	71.32
150	76.49	83.60	79.89	79.17
200	82.94	90.90	86.63	84.36
250	82.92	95.68	88.82	84.40
300	83.74	97.18	89.95	84.67
350	85.91	97.71	91.43	86.29
400	87.51	97.60	92.28	87.27
450	89.13	97.57	93.16	87.93
550	89.98	98.38	93.99	89.02
570	90.60	98.54	94.40	90.18

5.1.2 Association Word Features

We manually extracted 30 association words mentioned from the positive class sentences. For simplicity, we have considered only association verbs. Our aim is to study the effect of incorporating association words on the performance of the classifier. The existence of an association word in a sentence is marked a ‘1’ entry

into the feature vector. 3-fold cross validation experiments are repeated by concatenating the association word feature using all 570 sentences. Table 5.2 shows the results. It can be seen that concatenation of BOW with association words results in a very minor increase of 0.08% in the classification performance.

Table 5.2: Effect of concatenating association words with BOW feature.

Folds	Precision (%)	Recall (%)	F-score (%)	Average (%)
1	90.64	96.88	93.66	89.01
2	90.34	99.38	94.64	90.53
3	91.81	98.74	95.15	91.53
AVG	90.93	98.33	94.48	90.36

5.1.3 Lexical Features

The lexical feature used in this study is POS tags, which involves the single lexical features with the root of the words. The previous experiment is now repeated using a concatenation of BOW features with the POS tags. The results are shown in Table 5.3. It is observed that the F-score performance is decreased by 0.38% compared to using BOW features only.

Table 5.3: Effects of concatenating POS Tag with BOW feature

Folds	Precision (%)	Recall (%)	F-score (%)	Average (%)
1	89.93	98.15	93.86	89.79
2	91.58	96.06	93.77	88.43
3	90.87	98.27	94.43	92.12
AVG	90.79	97.49	94.02	90.11

5.2 Concatenation of all Features Used

It has earlier been shown that a careful combination of features may improve the classification performance [65]. Therefore, although it has been observed in the

previous sections that pair-wise concatenation of the association word and lexical features with the BOW feature slightly degrades or improves the classification performance respectively, a final experiment is conducted using all feature types used in concatenation. The results are presented in the last row of Table 5.4. The performance of the classifier using all 3 features in concatenation achieves a 0.38% improvement over the classifier which uses BOW feature only. Although the improvement may not be regarded as significant, the classifier that uses all 3 features is used to classify the remaining sentences in the test data collection since the classifier has performed the best using all three features.

In addition, experiments were conducted to study the effect of feature combinations of the three features used in this study. The feature combinations were engineered based on the results of the single feature experiment. The strategy adopted was to combine the feature in such a way that they complement each other's strengths and weaknesses in precision and recall values. It was observed that some specific combinations of feature types do not have a significant improvement in performance. However, the results suggest that combining all three features slightly improve the system performance. The results of the experiments shows feature combinations significantly improve performance as given in Table 5.4.

Table 5.4: Results of feature combination.

BOW	**Association Words**	**POS Tag**	**Feature Combination Results**		
			Precision (%)	**Recall (%)**	**F-score (%)**
x			90.60	98.54	94.40
x	x		90.93	98.33	94.48
x		x	90.79	97.49	94.02
x	x	x	91.25	98.59	94.78

5.3 Main Findings

By applying text mining methods, we investigate the association 1337 unique neurotransmitter receptor symbols and 465 unique mental and behavioral disorders. Overview of the general results is shown in Table 5.5.

Table 5.5: Main data retrieved and analyzed

Number of Abstract retrieved	Number of sentences analyzed	Number of unique associated pairs identified
835691	4642	1517

A brief overview of sample the associations from the database are provided in Table 5.6.

Table 5.6: Number of association for specific neurotransmitter receptor-disease pairs

	Schizophrenia	Anxiety	Alcohol Dependence
DRD3 receptor	124	8	2
5-HT1A receptor	93	21	16
MGLU5 receptor	32	55	-
MOR receptor	-	14	-

This is an overview of a subset of information available in the database. It is only feasible to display and analyze all the information in a database format. The NTreceptorDB database covering this data is presented in Section 5.6.

5.4 Manual Assessment and Common Sources of Error

Since a test data set is not available, we annotated an additional 100 randomly selected sentences that contain a neurotransmitter receptor-disease pair from the data set and annotated them manually in order to prepare a test set. The classifier was

trained using all the 570 sentences generated as train data and tested on the previously unseen 100 test sentences.

The results show that the SVM classifier achieves a 53.78% precision and 78.77% recall, resulting in 62.89% F-score. The F-Score achieved on the test set is about 30% lower than the F-score achieved during the 3-fold cross validation. This may be due to several reasons. Firstly, the generalization performance of the classifier might have dropped due to the presence of sentences that can be regarded as 'may be' (or noisy) in the test data set. Such sentences can be regarded as 'may be' sentences since they possess clues to an association between a neurotransmitter receptor and a mental disorder but the indication for an association is not so strong, or very clear. Below is an example of a 'may be' sentence with PubMed ID: 250882.

*"Comparison of transcript levels in **schizophrenia** patients and unaffected siblings found lower patient expression of **GABRA6** and coexpressed genes of **GABRA1**."*

Such sentences were not included in the train data set, whereas they were labeled as positive samples in the test data set. The "different" nature of the train and test datasets may account for this decrease in the classification performance. In the future we may try to train the classifier with noisy data in order to improve the classification performance

Secondly, there exist a number of negation words in the sentences such as 'not', 'excluding' etc. which indicates the absence of an association. Due to the occurrence of negation words in the sentences, our classifier as wrongly classified many such

sentences as positive. In our future work, the classifier performance could be improved by incorporating a negation module to deal with this problem [66].

The high recall-low precision of the classifier can be attributed to the unbalanced training data used. The training data contains 479 positive samples and only 91 negative samples. As is well known, larger number of samples in the positive class always result in higher recall values.

5.5 Conflicting Experimental Evidence

There has been conflicting evidence reported in the biomedical literature relevant to genetic variations in neurotransmitter receptors and their associations as susceptibility genes for certain diseases. Following issues pertinent to the nature of the specific data investigated in this thesis present challenge in correct mining of the neurotransmitter receptor-disease relationships.

5.5.1 Polymorphisms and Difference in Disease Association

Different polymorphisms of a given gene could either be implicated in a disease state or could be irrelevant for that particular disease. This variation is explained with an example sentence in Table 5.7.

Table 5.7: Polymorphisms of neurotransmitter receptor genes and difference in disease association

Sentences	**PubMed ID**
"Dopamine receptor D1 gene -48A/G polymorphism is associated with bipolar illness but not with schizophrenia in a Polish population."	249316

5.5.2 Conflicting Results from Different Studies

There has also been conflicting evidence in the field for a given neurotransmitter receptor and disease association. Several examples of these can be seen below in Table 5.8. Sometimes even in a single sentence conflict can be mentioned, as shown below in Table 5.9.

Table 5.8: Conflicting results from various studies

	Sentences	**PubMed ID**
Positive Evidence	*"Recent studies suggest a possible involvement of 5-HT2A receptors in the pathophysiology and treatment of schizophrenia."*	643657
Negative Evidence	*"Our results suggest that an abnormality in the 5-HT2A receptor gene in schizophrenia is unlikely."*	643468
Positive Evidence	*"Our genetic dissection of the CCK system thus far suggests that the CCK-B receptor gene variation may contribute to the neurobiology of panic disorder."*	98161
Negative Evidence	*"However, no evidence of allelic association was found between the polymorphic repeat of the CCKBR gene and either panic disorder or schizophrenia (P = 0.186 and 0.987, respectively)."*	220905

Table 5.9: Single sentence conflicting evidence

Sentences	**PubMed ID**
"In the logistic regression analysis, the long form variants of the DRD4 polymorphism did predict schizophrenia after the contributions of the age and gender of the subjects were included (p = 0.036, OR = 2.319), but the CC and GG genotypes of the codon 72 polymorphism of TP53 did not."	239118

5.5.3 Indirect Evidence

Some evidence of association could be indirectly reported, for example by reporting gene expression analysis details. Detection of such cases requires additional advanced computational methods. For such sentences manual analysis is needed. Examples of these sentences can be seen below in Table 5.10.

Table 5.10: Indirect evidence of association

Sentences	**PubMed ID**
"The expression of NR1 and NR2C subunit transcripts is decreased in the thalamus in schizophrenia."	671634
"Decreased NR1, NR2A, and SAP102 transcript expression in the hippocampus in bipolar disorder."	693369
"There was a significant decrease in the expression of transcripts for NR1 and NR2A subunits and SAP102 in bipolar disorder."	693369

5.5.4 Allelic Variation in Different Human Populations

Another level of complexity is added to the data based on allelic variations across different human populations. Results could be different, even contradicting, in different human population. Examples are given in Table 5.11.

Table 5.11: Sentences with different evidence in allelic variation in different human populations

	Sentences	**PubMed ID**
Positive Evidence	*"DRD4 and COMT genes were observed to be the most important candidates in North Indian schizophrenia subjects."*	202218
Negative Evidence	*"The present results do not support a major role for DRD4 in the etiology of schizophrenia among Caucasians from Sweden."*	241623

5.5.5 Animal Studies

Lastly, it is important to note that a large group of studies that indicate neurotransmitter receptor-disease association are in fact animal studies. In this study, we did not limit the evidence to human data only. However, we have not checked cross-species experimental validation. Table 5.12 shows results from animal studies.

Table 5.12: Sentences that involve animal studies

Sentences	**PubMed ID**
"The findings in this study indicate that the 5-HT2A receptor is involved in the pathophysiology of anxiety disorders in dogs."	643016
"Activation of the serotonin 5-HT2C receptor is involved in the enhanced anxiety in rats after single-prolonged stress."	650976
"Taken together, these data demonstrate a selective and robust reduction in anxiety- and depression-related behavior in NMDA receptor NR2A subunit KO mice."	693154
"Altered NR2A subunit expression in the medial prefrontal cortex of rats reared in isolation suggests that NMDA receptor dysfunction may contribute to the underlying pathophysiology of this preclinical model of aspects of schizophrenia."	684494

5.6 NTreceptorDB Web Interface

A web-interface that enables users to analyze association between neurotransmitter receptor and mental disorder data is developed and is named as NTreceptorDB. Abstracts available in NTreceptorDB show biomedical evidence for neurotransmitter receptor-disease association and are linked to the PubMed database. Hence, NTreceptorDB serves as a public tool for analysis of the relationship between neurotransmitter receptors and mental or behavioral disorders. NTreceptorDB is accessible via http://NTreceptorDB.emu.edu.tr. Figure 5.1 shows the web-interface for NTreceptorDB.

Search for an association of a set neurotransmitter receptor and/or a mental disorder could be made by a query or by list provided in the web-interface as shown in Figure 5.2. Submitting the neurotransmitter receptor's official symbol as listed in Entrez Gene DB can use the user query interface. The system allows input of several synonyms of a symbol. For instance, the system is not case sensitive to user queries, and also spaces of letters and digits are usually ignored by the system. Any user can access the NTreceptorDB to search for a relationship mined from a particular PubMed record by giving its PubMed ID in the web-interface.

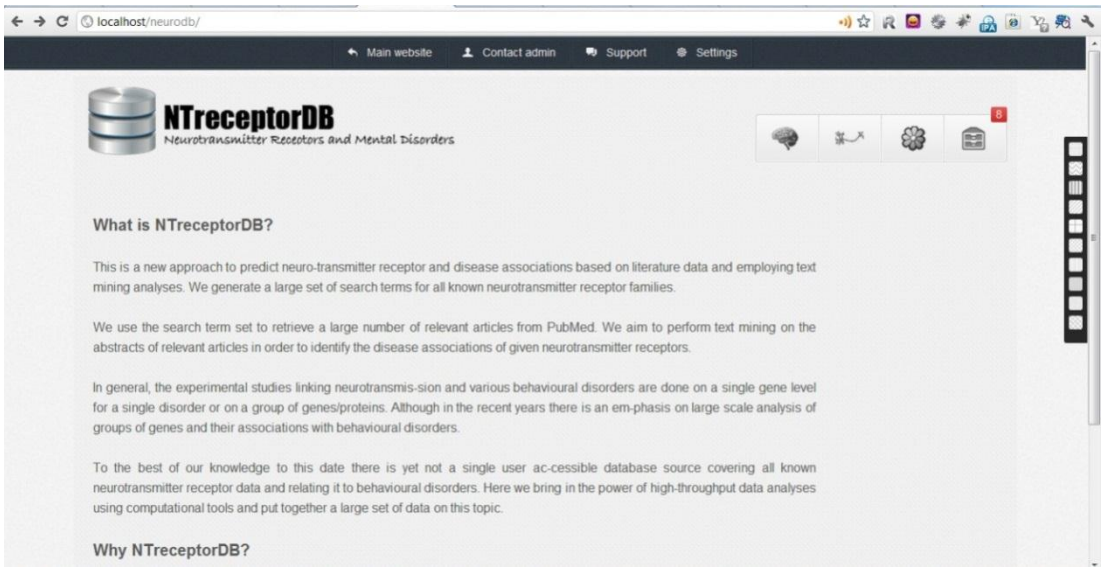


Figure 5.1: NTreceptorDB web interface.

The benefits and usage of the web-interface in NTreceptorDB is demonstrated below in Figure 5.3. This figure is a snapshot from NTreceptorDB showing the retrieval of the association of DRD2 (dopamine receptor) along with its associated disease data. The legend in the figure shows the relationship that contains information between the two entities that is linked to the PubMed database. The information is also linked to the NCBI's Entrez Gene DB by using the Gene DB IDs in the search query.

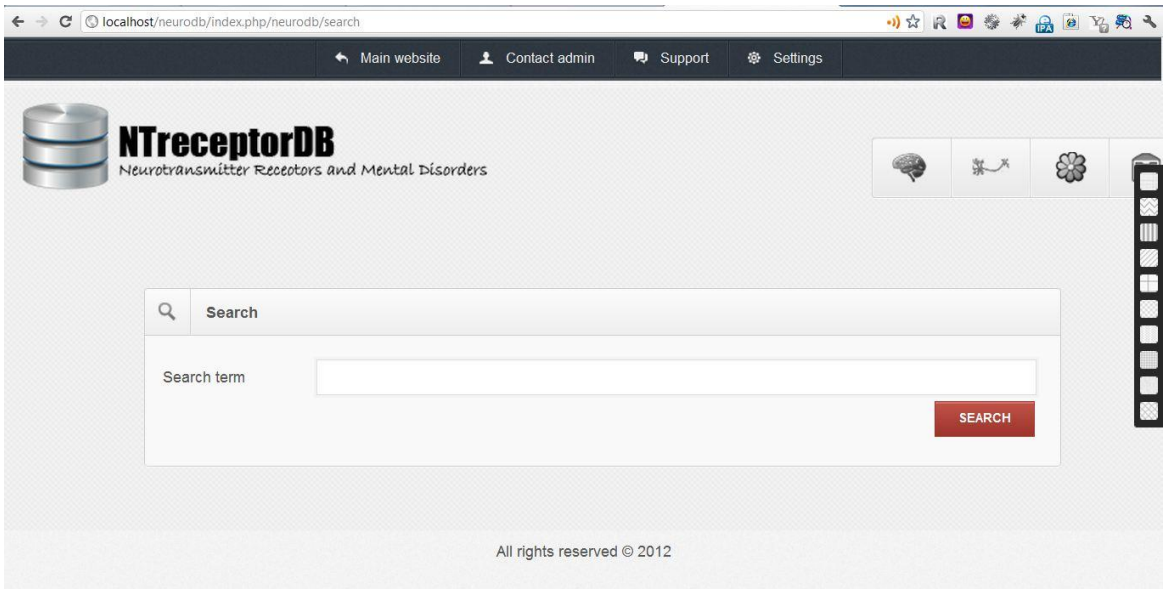
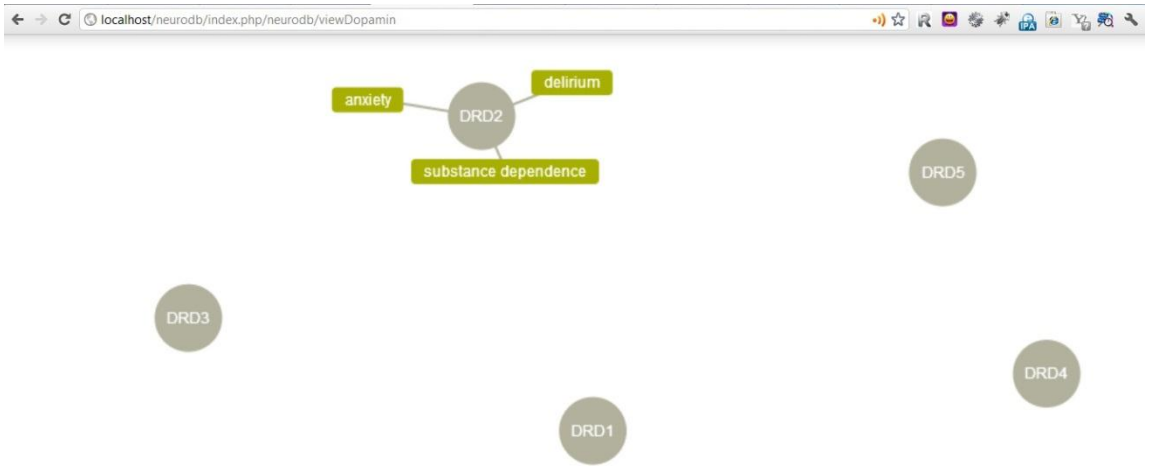


Figure 5.2 NTreceptorDB search query interface

The Entrez Gene DB gives additional information on gene names such as functions based on GO concepts and also with metabolic ways that they are involved in. This tool provides a quick visualization of an enormous amount of data, which generates an easy platform for understanding neurotransmitter receptor-disease association.



DRD1	DRD2	DRD3	DRD4	DRD5
schizophrenia	delirium	anorexia nervosa	mood disorder	ADHD
alcohol dependence	substance dependence	dementia	delusional disorder	conduct disorder
bipolar disorder	anxiety	schizoaffective disorder	-	-

Figure 5.3: NTreceptorDB description of a retrieve neurotransmitter receptor.

Chapter 6

CONCLUSION

6.1 Conclusion

In this thesis, we present a new approach to predict neurotransmitter receptor and mental disease associations based on literature data by employing text mining analysis. We retrieve a large number of relevant articles from PubMed and use an SVM classifier to identify the disease associations of given neurotransmitter receptor. The thesis is unique in the sense that it focuses particularly on the neurotransmitter receptor-mental and behavioral disorder association as opposed to the general gene-disease associations. This is first dataset of its kind, specifically focusing on this group of genes and disorders.

In general, the experimental studies linking neurotransmission and various behavioral disorders are done on a single gene level for a single disorder or on a group of genes and proteins. Hence, biomedical data specific to this field is not present in a comprehensive, publicly accessible database. Our results of biomedical literature text mining presented in this thesis provide a centralized, comprehensive source documenting neurotransmitter receptors implicated in several diseases states. It is evident from the data that a given neurotransmitter receptor is implicated in several different diseases as illustrated in Chapter 5 Section 5.4.

We make our results publicly accessible via a new database. NTreceptorDB database was constructed containing neurotransmitter receptor-disease interaction data based on biomedical literature. NTreceptorDB and the associated user friendly web-interface would enable storage of and access to the relevant neurotransmitter receptor–disease data, which is validated using the PubMed database. End users such as biomedical researchers would be able to view annotations, search for biological data, validate links across resources, and create new information resources to capture new concepts as they arise. Main benefit of this work relies in the originality of the dataset and its comprehensive presentation in a publicly accessible platform. These features would facilitate further research in the field on a large scale, in addition would provide healthcare professionals with a valuable biomedical source.

6.2 Future Direction

Our future directions include the expansion of the current study with the following specific aims. Firstly, a manual annotation of large volume of train data set using multiple annotators will be performed, in order to come up with efficient train data and increase the performance of the system. Secondly, the classifier used in this study will be improved in terms of classification, in order to cope with the negations mentioned in the text that increase the false positives. In addition, we will focus on the indirect evidence mining task to extract more data from the abstracts retrieved. As discussed in Chapter 5, Section 5.4.3, there are a lot of neurotransmitter receptor-disease associations revealed as indirect evidence.

Furthermore, the study presented here may be extended to concentrate on polymorphisms specific data mining and documenting. As noted in chapter 5, section

5.4.1, different polymorphisms of the same gene could present different susceptibilities to various disorders. Also our future aims include integration of relevant environmental data. Gene-environment interplay (GxE) plays a critical role in the onset of mental and behavioral disorders [56]. As reviewed by Taneri *et al*. [56], behavioral disorders arise as results of specific GxEs. For example, the specific GxE implication in depression refers to the interaction of early-life stress with certain genetic variation in the serotonin transporter gene. Carriers of a particular allele (specifically referred to as the short allele for 5-HTTLPR) for the serotonin transporter gene, who experience early-life stressful events, are more susceptible to developing depression [56]. Lastly, we will investigate the drug responsiveness of individuals based on their genetic variations. All issues indicated above will be addressed and NTreceptorDB will be updated accordingly.

REFERENCES

[1] Tan. A.H. (1999). Text mining: The state of the art and the challenges. In Proc of the Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, pages 65–70.

[2] Ozgur A., Vu T., Erkan G., and Radev D.R. (2008). Identifying gene-disease associations using centrality on a literature mined gene interaction network. Bioinformatics, Volume 24, Number 13, pp. i277-i285.

[3] Chun H. W. *et al*. (2006). "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning", in *Proc. the Pacific Symposium on Biocomputing,* vol. 11, pp. 4-15.

[4] Iwama H., Gojobori T. (2008). "Identification of Neurotransmitter Receptor Genes Under Significantly Relaxed Selective Constraint By Orthologous Gene Comparisons Between Humans And Rodents" *Mol. Biol. E., vol*. 19pp. 1891-1901.

[5] NCBI's GeneDB: http://www.ncbi.nlm.nih.gov/gene/

[6] Tsai *et al* R. T. H. (2009). HypertenGene: Extracting key hypertension genes from biomedical literature with position and automatically-generated template features. *BMC Bioinformatics*, vol. 10, pp. (Suppl. 15): S9.

[7] George B.C. (2009). Neurotransmitters http://webspace.ship.edu/cgboer/genpsyneurotransmitters.html.

[8] American Psychiatric Association. (2000). Appendix I: Outline for cultural formulation and glossary of culturebound syndromes. In Diagnostic and statistical manual of mental disorders (4th ed., text rev.). doi:10.1176/appi.books. 9780890423349.7060.

[9] Joachims T. (1999). Advances in Kernel Methods-Support Vector Learning. Cambridge, MA, USA: MIT-Press; Making Large-Scale SVM Learning Practical.

[10] Altman, R. C.M. Bergman, J. Blake, Blaschke C., Cohen A., Gannon F., Grivell L., Hahn U., Hersh W., Hirschman L., Jensen L.J., Krallinger M., Mons B., O'Donoghue S.I., Peitsch M., Rebholz-Schumann D., Shatkay H. and Valencia A. (2008). Text Mining for Biology: The Way Forward. Genome Biology 9:S7.

[11] Andreas H., Andreas N., and Gerhard P. (2005). A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, (20)1:19-62.

[12] Allan J., (ed). (2002). Topic Detection and Tracking. Kluwer Academic Publishers, Norwell, MA.

[13] M. B. and Hand D. J. (eds.). (1999). Intelligent data analysis. Springer-Verlag New York, Inc.

[14] Bezdek J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

[15] Bloehdorn S. and Hotho A. (2004). Text classification by boosting weak learners based on terms and concepts. In Proc. IEEE Int. Conf. on Data Mining (ICDM 04), pages 331–334. IEEE Computer Society Press.

[16] Deerwester S., Dumais S.T., G.W. Furnas, and T.K. Landauer. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Sciences, 41:391–407.

[17] Fox K. L., Frieder O., Knepper M. M., and Snowberg E. J. (1999) Sentinel: A multiple engine information retrieval and visualization system. Journal of the American Society of Information Science, 50(7):616–625.

[18] Donner A, Klar N (2004). Pitfalls of and Controversies in Cluster Randomization Trial. Am J Public Health, 94:416-422.

[19] Forgy E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classification. Biometrics, 21(3):768–769.

[20] Fayyad U. M. Piatetsky-Shapiro G., and Smith P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In Knowledge Discovery and Data Mining, pages 82–88.

[21] Gaizauskas R. (2003). An information extraction perspective on text mining: Tasks, technologies and prototype applications. http://www.itri.bton.ac.uk/projects/euromap/TextMiningEvent/Rob_Gaizauskas.pdf.

[22] Gersho A. and Gray R. M. (1992). Vector quantization and signal compression. Kluwer Academic Publishers.

[23] Good I. J. (1965). The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Press, Cambridge, MA.

[24] Lan M, C L Tan and H B Low (2006). Proposing a new term weighting scheme for text categorization, 21st National Conference on Artificial Intelligence, AAAI-2006, 16-20, Boston, Massachusetts, USA.

[25] Kristina T. and Christopher D.M. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[26] Erik F., Tjong K.S., and Sabine B. (2000). Introduction to the CoNLL-2000 Shared Task: Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal.

[27] Dan K. and Christopher D.M.. (2003). Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

[28] Agarwal S, Yu H. (2010). Detecting Hedge Cues and their Scope in Biomedical Literature with Conditional Random Fields. Journal of Biomedical Informatics-Elsevier, 43(6):953-961.

[29] Nielsen H., Krogh A. (1998). Prediction of signal peptides and signal anchors by a Hidden Markov Model J Glasgow, T Littlejohn, F Major, R Lathrop, D Sankoff, C Sensen (Eds.), Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, pp. 122–130.

[30] Chen, S., and Rosenfeld, R. (1999). Efficient sampling and feature selection in whole sentence maximum entropy language models. In Proceedings of ICASSP'99.IEEE.

[31] Zhao Le, Callan J. (2010). Term neces sity prediction, Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto, ON, Canada.

[32] Ronan C., Colm O. (2005). Evolving General Term-Weighting Schemes for Information Retrieval: Tests on Larger Collections, Artificial Intelligence Review, v.24 n.3-4, p.277-299.

[33] Hearst M. (1999). Untangling text data mining.In Proc. of ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics.

[34] Junker M. and Hoch R. and Dengel A. (1999). On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy. Proceedings ICDAR 99, Fifth Intl. Conference on Document Analysis and Recognition. Bangelore, India.

[35] NCBIs PubMed: http://www.ncbi.nlm.nih.gov/pubmed.

[36] Hirschman L., Yeh A., Blaschke C., Valencia A. (2005) Overview of BioCre-AtIvE: critical assessment of information extraction for biology. BMC Bioinformatics, 6:S1.

[37] Valencia A. (2008). "Text mining in genomics and systems biology", Proceeding of the 2nd international workshop on Data and text mining in bioinformatics (DTMBIO 08), ACM, pp.3-4.

[38] Chen H., Sharp B.M.. (2004). Content-rich biological network constructed by mining pubmed abstracts. BMC Bioinformatics;5:147-159.

[39] Gonzalez, et al. (2007). Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. Pac. Symp., p:12:28-39.

[40] Adamic, et al. LA. (2002). A literature based method for identifying gene-disease connections. Proceedings of the IEEE Computer Society Conference on Bioinformatics; Stanford, CA. p. 109-117.

[41] Al-Mubaid H. , Singh RK. (2005). A new text mining approach for finding protein-to-disease associations. Am J Biochem, Biotechnol;1:145-152.

[42] Chen M.S., Han J., and Yu P.S. (1996). Data mining: an overview from a database perspective. IEEE Transaction on Knowledge and Data Engineering, 8(6):866–883.

[43] Freudenberg J., Propping P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics; p. 18 (Suppl. 2):S110-S115.

[44] Robert S.B. (2008). "Neurons, Synapses, Action Potentials, and Neurotransmission". http://www.mind.ilstu.edu/curriculum/neurons_intro/ neurons_ intro.php

[45] Grant S.G., Marshall M.C., Page K.L., Cumiskey M.A., Armstrong J.D. (2005). "Synapse proteomics of multiprotein complexes: en route from genes to nervous system diseases." Hum Mol Genet 14 Spec No. 2:R225-34.

[46] Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature. 444(7119):580-6.

[47] Jonathan W. (2003). "Weill CU researchers find key link in process of neurotransmission". http://www.news.cornell.edu/chronicle/03/2.6.03/Weill_ neurotr ans. html.

[48] Shroomery Forum http://www.shroomery.org/forums/showflat.php/Number/ 10492259.

[49] Jokela M., Keltikangas-Jarvinen L., (2007) "Serotonin Receptor 2a Gene and the Influence of Childhood Maternal Nurturance on Adulthood Depressive Symptoms - Archives of General", Am Med Assoc, vol. 64, pp. 3.

[50] Silvia H.C., (2001). “Communication Between Nerve Cells”, http://www.cerebromente.org.br/n12/fundamentos/neurotransmissores/neurotransmitters2.html.

[51] Neurosciust, (2009). “The Resting Potential and The Action Potential”, http://www.slideshare.net/neurosciust/the-resting-potential-and-the-action-potential.

[52] R and D Systems, (2010). “Neurotransmitter Receptors, Transporters, and Ion Channels”, http://www.rndsystems.com/molecule_group.aspx?g=682&r=5http://www.rndsystems.com/molecule_group.aspx?g=682&r=5.

[53] Lundback Institute (2011). “Neurological Control: Neurotransmitters” ,http://www.brainexplorer.org/neurological_control/ Neurological_Neurotransmitters.shtml.

[54] Dalley J.W. (2009). “Dopamine Receptors in the Learning, Memory and Drug Reward Circuitry: Seminars in Cell and Developmental Biology”, ELSEVIER, vol. 20, pp. 403-410.

[55] Scitable N.E. (2011). “The Genetic Variation in a Population Is Caused by Multiple Factors The Genetic Variation in a Population Is Caused by Multiple Factors”, http://www.nature.com/scitable/topicpage/the-genetic-variation -in-a-population-is-6526354.

[56] Taneri B., Ambrosino E., van Os J., Brand A. (2012). A new public health genomics model for common complex diseases, with an application to common behavioral disorders Personalized Medicine 9: 1. 29-38.

[57] Musa A.K., Varoğlu E., Taneri B. (2011). "Role of Neurotransmitter Receptors in Behavioral Disorders- A high-Throughput Analysis using Text Mining", International Symposium on Health Informatics and Bioinformatics (HIBIT 2011) , Izmir, Turkey, May 2-5.

[58] Le Fool B., Gallo A., Le Strat Y., Lu L., Gorword L. (2009). "Genetics of Dopamine Receptors and Drug Addiction: A Comprehensive Review", Behavioural Pharmacology, vol. 20, pp. 1-17.

[59] Kafkas Ş., Varoğlu E., Taneri B. (2008). "Methods for Abstract Retrieval from Pubmed Database for Alternatively Spliced Genes", International Symposium on Health Informatics and Bioinformatics (HIBIT 2008). Istanbul, Turkey, May 18-20.

[60] Guyon I, Weston J, Barnhill S, Vapnik V. (2002). Gene Selection for Cancer Classification using Support Vector Machines Machine Learning; p. 46:389-422.

[61] Provost F, Fawcett T, Kohavi R. (1998). The Case Against Accuracy Estimation for Comparing Induction Algorithms ICML-98 (15th International Conference on Machine Learning).

[62] Mooney R. J., Bunescu R., (2005). Mining knowledge from text using information extraction, ACM SIGKDD Explorations Newsletter, v.7 n.1, p.3-10, [doi>10.1145/1089815.1089817].

[63] Bhardwaj N. and Lu H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics, 21(11):2730–2738.

[64] Kulick S., Bies A., Liberman M., Mandel M., McDonald R., Palmer M., Schein A. and Ungar L. (2004). Integrated Annotation for Biomedical Information Extraction, HLT/NAACL 2004 Workshop: Biolink, pp. 61-68.

[65] Dimililer N., Varoğlu E., Altinçay H. (2009). Classifier subset selection for biomedical named entity recognition. Appl. Intell. 31(3): 267-282.

[66] Callinan P, Feinberg A. (2006). The emerging science of epigenomics. Hum. Mol. Genetics 15, R95–R101.

APPENDICES

Appendix A: Neurotransmitter Receptor List

No.	Neurotransmitter Receptor
1	Adenosine Receptor
2	Adrenergic Receptor
3	Angiotensin Receptor
4	Bonbesin Like 3 Receptor
5	Bradykinin Receptor
6	Calcitonin Receptor
7	Cholecystokinin Receptor
8	Crh Receptor
9	Dopamine Receptor
10	GABAA Receptor
11	GABAB Receptor
12	Galanin receptor
13	Gastrin-Releasing Peptide Receptor
14	Glucagon Receptor
15	Glycine Receptor
16	Grh Receptor
17	Ghrh Receptor
18	Histamine Receptor
19	Serotonin M Receptor Metabotropic
20	Serotonin I Receptor Ionotropic
21	Glutamate I Receptor Ionotropic
22	Muscarinic Cholinergic Receptor
23	Glutamate M Receptor Metabotropic
24	Nichotinic Cholinergic Receptor
25	Neuromedin B Receptor
26	Neuromedin U Receptor
27	Neuropeptide Y Receptor
28	Neurotensin Receptor
29	Opioid Receptor
30	Purinergic I Receptor Ionotropic
31	Secretin Receptor
32	Somatostain Receptor

33	Tachykinin Receptor
34	Trh Receptor
35	Vip Receptor

Appendix B: Mental Disorder List (DSM-IV TR)

No.	Diagnostic and Statistical Manual of Mental Disorders (DSM-IV TR)
1	Academic Problem
2	Acculturation Problem
3	Acute Stress Disorder
4	Adjustment Disorder
5	Adjustment Disorder With Anxiety
6	Adjustment Disorder With Depressed Mood
7	Adjustment Disorder With Disturbance of Conduct
8	Adjustment Disorder With Mixed Anxiety and Depressed Mood
9	Adjustment Disorder With Mixed Disturbance of Emotions and Conduct
10	Adult Antisocial Behavior
11	Adverse Effects of Medication NOS
12	Age-Related Cognitive Decline
13	Agoraphobia
14	Agoraphobia Without History of Panic Disorder
15	Alcohol Abuse
16	Alcohol Dependence
17	Alcohol Intoxication
18	Alcohol Intoxication Delirium
19	Alcohol Withdrawal
20	Alcohol Withdrawal Delirium
21	Alcohol-Induced Amnestic Disorder
22	Alcohol-Induced Anxiety Disorder
23	Alcohol-Induced Dementia
24	Alcohol-Induced Mood Disorder
25	Alcohol-Induced Persisting Amnestic Disorder
26	Alcohol-Induced Persisting Dementia
27	Alcohol-Induced Psychotic Disorder
28	Alcohol-Induced Psychotic Disorder With Delusions
29	Alcohol-Induced Psychotic Disorder With Hallucinations
30	Alcohol-Induced Sexual Dysfunction
31	Alcohol-Induced Sleep Disorder
32	Alcohol-Related Disorder

33	Amnestic Disorder
34	Amnestic Disorder Without Behavioral Disturbance
35	Amphetamine Abuse
36	Amphetamine Dependence
37	Amphetamine Intoxication
38	Amphetamine Intoxication Delirium
39	Amphetamine Withdrawal
40	Amphetamine-Induced Anxiety Disorder
41	Amphetamine-Induced Mood Disorder
42	Amphetamine-Induced Psychotic Disorder
43	Amphetamine-Induced Psychotic Disorder With Delusions
44	Amphetamine-Induced Psychotic Disorder With Hallucinations
45	Amphetamine-Induced Sexual Dysfunction
46	Amphetamine-Induced Sleep Disorder
47	Amphetamine-Related Disorder
48	Anorexia Nervosa
49	Antisocial Behavior
50	Antisocial Personality
51	Antisocial Personality Disorder
52	Anxiety
53	Anxiety Disorder
54	Anxiolytic Abuse
55	Anxiolytic Dependence
56	Asperger's Disorder
57	Attention-Deficit/Hyperactivity Disorder
58	Attention-Deficit/Hyperactivity Disorder Combined Type
59	Attention-Deficit/Hyperactivity Disorder Predominantly Hyperactive-Impulsive
60	Attention-Deficit/Hyperactivity Disorder Predominantly Inattentive
61	Autistic Disorder
62	Avoidant Personality
63	Avoidant Personality Disorder
64	Bereavement
65	Bipolar Disorder
66	Bipolar I Disorder
67	Bipolar I Disorder Most Recent Episode

68	Bipolar I Disorder Most Recent Episode Depressed
69	Bipolar I Disorder Most Recent Episode Depressed In Full Remission
70	Bipolar I Disorder Most Recent Episode Depressed In Partial Remission
71	Bipolar I Disorder Most Recent Episode Depressed Mild
72	Bipolar I Disorder Most Recent Episode Depressed Moderate
73	Bipolar I Disorder Most Recent Episode Depressed Severe
74	Bipolar I Disorder Most Recent Episode Depressed Severe With Psychotic Features
75	Bipolar I Disorder Most Recent Episode Depressed Severe Without Psychotic Features
76	Bipolar I Disorder Most Recent Episode Hypomanic
77	Bipolar I Disorder Most Recent Episode Manic
78	Bipolar I Disorder Most Recent Episode Manic In Full Remission
79	Bipolar I Disorder Most Recent Episode Manic In Partial Remission
80	Bipolar I Disorder Most Recent Episode Manic Mild
81	Bipolar I Disorder Most Recent Episode Manic Moderate
82	Bipolar I Disorder Most Recent Episode Manic Severe
83	Bipolar I Disorder Most Recent Episode Manic Severe With Psychotic Features
84	Bipolar I Disorder Most Recent Episode Manic Severe Without Psychotic Features
85	Bipolar I Disorder Most Recent Episode Mixed
86	Bipolar I Disorder Most Recent Episode Mixed In Full Remission
87	Bipolar I Disorder Most Recent Episode Mixed In Partial Remission
88	Bipolar I Disorder Most Recent Episode Mixed Mild
89	Bipolar I Disorder Most Recent Episode Mixed Moderate
90	Bipolar I Disorder Most Recent Episode Mixed Severe
91	Bipolar I Disorder Most Recent Episode Mixed Severe With Psychotic Features
92	Bipolar I Disorder Most Recent Episode Mixed Severe Without Psychotic Features
93	Bipolar I Disorder Single Manic Episode
94	Bipolar I Disorder Single Manic Episode In Full Remission
95	Bipolar I Disorder Single Manic Episode In Partial Remission
96	Bipolar I Disorder Single Manic Episode Mild
97	Bipolar I Disorder Single Manic Episode Moderate
98	Bipolar I Disorder Single Manic Episode Severe
99	Bipolar I Disorder Single Manic Episode Severe With Psychotic Features
100	Bipolar I Disorder Single Manic Episode Severe Without Psychotic Features
101	Bipolar II Disorder
102	Body Dysmorphic Disorder

103	Borderline Intellectual Functioning
104	Borderline Personality
105	Borderline Personality Disorder
106	Breathing-Related Sleep Disorder
107	Brief Psychotic Disorder
108	Bulimia Nervosa
109	Caffeine Intoxication
110	Caffeine-Induced Anxiety Disorder
111	Caffeine-Induced Sleep Disorder
112	Caffeine-Related Disorder
113	Cannabis Abuse
114	Cannabis Dependence
115	Cannabis Intoxication
116	Cannabis Intoxication Delirium
117	Cannabis-Induced Anxiety Disorder
118	Cannabis-Induced Psychotic Disorder
119	Cannabis-Induced Psychotic Disorder With Delusions
120	Cannabis-Induced Psychotic Disorder With Hallucinations
121	Cannabis-Related Disorder
122	Catatonic Disorder
123	Child or Adolescent Antisocial Behavior
124	Childhood Disintegrative Disorder
125	Chronic Motor
126	Circadian Rhythm Sleep Disorder
127	Circadian Rhythm Sleep Disorder delayed sleep phase
128	Circadian Rhythm Sleep Disorder jet lag
129	Circadian Rhythm Sleep Disorder shift work
130	Cocaine Abuse
131	Cocaine Dependence
132	Cocaine Intoxication
133	Cocaine Intoxication Delirium
134	Cocaine Withdrawal
135	Cocaine-Induced Anxiety Disorder
136	Cocaine-Induced Mood Disorder
137	Cocaine-Induced Psychotic Disorder

138	Cocaine-Induced Psychotic Disorder With Delusions
139	Cocaine-Induced Psychotic Disorder With Hallucinations
140	Cocaine-Induced Sexual Dysfunction
141	Cocaine-Induced Sleep Disorder
142	Cocaine-Related Disorder
143	Cognitive Disorder
144	Communication Disorder
145	Conduct Disorder
146	Conduct Disorder Adolescent-Onset
147	Conduct Disorder Childhood-Onset
148	Conversion Disorder
149	Cyclothymic Disorder
150	Delirium
151	Delusional Disorder
152	Dementia
153	Dementia Due to Creutzfeldt-Jakob Disease
154	Dementia Due to Head Trauma
155	Dementia Due to HIV Disease
156	Dementia Due to Huntington's Disease
157	Dementia Due to Parkinson's Disease
158	Dementia Due to Pick's Disease
159	Dementia of the Alzheimer's
160	Dementia of the Alzheimer's Type
161	Dementia of the Alzheimer's Type With Early Onset
162	Dementia of the Alzheimer's Type With Early Onset
163	Dementia of the Alzheimer's Type With Early Onset With Behavioral Disturbance
164	Dementia of the Alzheimer's Type With Early Onset Without Behavioral Disturbance
165	Dementia of the Alzheimer's Type With Late Onset
166	Dementia of the Alzheimer's Type With Late Onset
167	Dementia of the Alzheimer's Type With Late Onset With Behavioral Disturbance
168	Dementia of the Alzheimer's Type With Late Onset Without Behavioral Distrubance
169	Dementia With Behavioral Disturbance
170	Dementia Without Behavioral Disturbance
171	Dependent Personality
172	Dependent Personality Disorder

173	Depersonalization Disorder
174	Depressive Disorder
175	Developmental Coordination Disorder
176	Disorder of Infancy Childhood or Adolescence
177	Disorder of Written Expression
178	Disruptive Behavior Disorder
179	Dissociative Amnesia
180	Dissociative Disorder
181	Dissociative Fugue
182	Dissociative Identity
183	Dissociative Identity Disorder
184	Dyspareunia
185	Dyssomnia
186	Dysthymic Disorder
187	Eating Disorder
188	Encopresis
189	Encopresis With Constipation and Overflow Incontinence
190	Encopresis Without Constipation
191	Encopresis Without Overflow Incontinence
192	Enuresis
193	Exhibitionism
194	Expressive Language Disorder
195	Factitious Disorder
196	Factitious Disorder With Combined Psychological and Physical Signs and Symptoms
197	Factitious Disorder With Predominantly Physical Signs and Symptoms
198	Factitious Disorder With Predominantly Psychological Signs and Symptoms
199	Feeding Disorder of Infancy
200	Female Dyspareunia
201	Female Hypoactive Sexual Desire Disorder
202	Female Orgasmic Disorder
203	Female Sexual Arousal
204	Female Sexual Arousal Disorder
205	Female Sexual Dysfunction
206	Fetishism
207	Frotteurism

208	Gender Identity Disorder
209	Gender Identity Disorder in Adolescents
210	Gender Identity Disorder in Adolescents or Adults
211	Gender Identity Disorder in Adults
212	Gender Identity Disorder in Children
213	Generalized Anxiety
214	Generalized Anxiety Disorder
215	Hallucinogen Abuse
216	Hallucinogen Dependence
217	Hallucinogen Intoxication
218	Hallucinogen Intoxication Delirium
219	Hallucinogen Persisting Perception Disorder
220	Hallucinogen-Induced Anxiety Disorder
221	Hallucinogen-Induced Mood Disorder
222	Hallucinogen-Induced Psychotic Disorder
223	Hallucinogen-Induced Psychotic Disorder With Delusions
224	Hallucinogen-Induced Psychotic Disorder With Hallucinations
225	Hallucinogen-Related Disorder
226	Histrionic Personality
227	Histrionic Personality Disorder
228	Hypersomnia
229	Hypnotic Anxiolytic Intoxication
230	Hypnotic Anxiolytic Intoxication Delirium
231	Hypnotic Anxiolytic Withdrawal
232	Hypnotic Anxiolytic Withdrawal Delirium
233	Hypnotic Anxiolytic-Induced Amnestic Disorder
234	Hypnotic Anxiolytic-Induced Anxiety Disorder
235	Hypnotic Anxiolytic-Induced Dementia
236	Hypnotic Anxiolytic-Induced Mood Disorder
237	Hypnotic Anxiolytic-Induced Persisting Amnestic Disorder
238	Hypnotic Anxiolytic-Induced Persisting Dementia
239	Hypnotic Anxiolytic-Induced Psychotic Disorder
240	Hypnotic Anxiolytic-Induced Psychotic Disorder With Delusions
241	Hypnotic Anxiolytic-Induced Psychotic Disorder With Hallucinations
242	Hypnotic Anxiolytic-Induced Sexual Dysfunction

243	Hypnotic Anxiolytic-Induced Sleep Disorder
244	Hypnotic Anxiolytic-Related Disorder
245	Hypoactive Sexual Desire
246	Hypoactive Sexual Desire Disorder
247	Hypochondriasis
248	Identity Problem
249	Impulse-Control Disorder
250	Inhalant Abuse
251	Inhalant Dependence
252	Inhalant Intoxication
253	Inhalant Intoxication Delirium
254	Inhalant-Induced Anxiety Disorder
255	Inhalant-Induced Dementia
256	Inhalant-Induced Mood Disorder
257	Inhalant-Induced Persisting Dementia
258	Inhalant-Induced Psychotic Disorder
259	Inhalant-Induced Psychotic Disorder With Delusions
260	Inhalant-Induced Psychotic Disorder With Hallucinations
261	Inhalant-Related Disorder
262	Insomnia
263	Intermittent Explosive Disorder
264	Intoxication Delirium
265	Kleptomania
266	Learning Disorder
267	Major Depressive Disorder
268	Major Depressive Disorder Recurrent
269	Major Depressive Disorder Recurrent In Full Remission
270	Major Depressive Disorder Recurrent In Partial Remission
271	Major Depressive Disorder Recurrent Mild
272	Major Depressive Disorder Recurrent Moderate
273	Major Depressive Disorder Recurrent Severe
274	Major Depressive Disorder Recurrent Severe With Psychotic Features
275	Major Depressive Disorder Recurrent Severe Without Psychotic Features
276	Major Depressive Disorder Single Episode
277	Major Depressive Disorder Single Episode In Full Remission

278	Major Depressive Disorder Single Episode In Partial Remission
279	Major Depressive Disorder Single Episode Mild
280	Major Depressive Disorder Single Episode Moderate
281	Major Depressive Disorder Single Episode Severe
282	Major Depressive Disorder Single Episode Severe With Psychotic Features
283	Major Depressive Disorder Single Episode Severe Without Psychotic Features
284	Male Dyspareunia
285	Male Erectile
286	Male Erectile Disorder
287	Male Hypoactive Sexual Desire Disorder
288	Male Orgasmic Disorder
289	Male Sexual Dysfunction
290	Malingering
291	Mathematics Disorder
292	Medication-Induced Movement Disorder NOS
293	Medication-Induced Postural Tremor
294	Mental Disorder
295	Mental Retardation
296	Mild Mental Retardation
297	Mixed Receptive-Expressive Language Disorder
298	Moderate Mental Retardation
299	Mood Disorder
300	Narcissistic Personality
301	Narcissistic Personality Disorder
302	Narcolepsy
303	Neglect of Child
304	Neuroleptic Malignant Syndrome
305	Neuroleptic-Induced Acute Akathisia
306	Neuroleptic-Induced Acute Dystonia
307	Neuroleptic-Induced Parkinsonism
308	Neuroleptic-Induced Tardive Dyskinesia
309	Nicotine Dependence
310	Nicotine Withdrawal
311	Nicotine-Related Disorder
312	Nightmare Disorder

313	Noncompliance With Treatment
314	Obsessive-Compulsive Disorder
315	Obsessive-Compulsive Personality
316	Obsessive-Compulsive Personality Disorder
317	Occupational Problem
318	Opioid Abuse
319	Opioid Dependence
320	Opioid Intoxication
321	Opioid Intoxication Delirium
322	Opioid Withdrawal
323	Opioid-Induced Mood Disorder
324	Opioid-Induced Psychotic Disorder
325	Opioid-Induced Psychotic Disorder With Delusions
326	Opioid-Induced Psychotic Disorder With Hallucinations
327	Opioid-Induced Sexual Dysfunction
328	Opioid-Induced Sleep Disorder
329	Opioid-Related Disorder
330	Oppositional Defiant Disorder
331	Pain Disorder Associated With Both Psychological Factors
332	Pain Disorder Associated With Both Psychological Factors and a General Medical Condition
333	Pain Disorder Associated With General Medical Condition
334	Pain Disorder Associated With Psychological Factors
335	Panic Disorder
336	Panic Disorder With Agoraphobia
337	Panic Disorder Without Agoraphobia
338	Paranoid Personality Disorder
339	Paraphilia
340	Parasomnia
341	Parent-Child Relational Problem
342	Partner Relational Problem
343	Pathological Gambling
344	Pedophilia
345	Personality Change
346	Personality Disorder

347	Pervasive Developmental Disorder
348	Phase of Life Problem
349	Phencyclidine Abuse
350	Phencyclidine Dependence
351	Phencyclidine Intoxication
352	Phencyclidine Intoxication Delirium
353	Phencyclidine-Induced Anxiety Disorder
354	Phencyclidine-Induced Mood Disorder
355	Phencyclidine-Induced Psychotic Disorder
356	Phencyclidine-Induced Psychotic Disorder With Delusions
357	Phencyclidine-Induced Psychotic Disorder With Hallucinations
358	Phencyclidine-Related Disorder
359	Phonological Disorder
360	Physical Abuse of Adult
361	Physical Abuse of Child
362	Pica
363	Polysubstance Dependence
364	Posttraumatic Stress Disorder
365	Premature Ejaculation
366	Primary Hypersomnia
367	Primary Insomnia
368	Profound Mental Retardation
369	Psychotic Disorder
370	Psychotic Disorder With Delusions
371	Psychotic Disorder With Hallucinations
372	Pyromania
373	Reactive Attachment Disorder
374	Reactive Attachment Disorder of Infancy or Early Childhood
375	Reading Disorder
376	Relational Problem NOS
377	Religious or Spiritual Problem
378	Rett's Disorder
379	Rumination Disorder
380	Schizoaffective Disorder
381	Schizoid Personality

382	Schizoid Personality Disorder
383	Schizophrenia
384	Schizophrenia Catatonic Type
385	Schizophrenia Disorganized Type
386	Schizophrenia Paranoid Type
387	Schizophrenia Residual Type
388	Schizophreniform Disorder
389	Schizotypal Personality
390	Schizotypal Personality Disorder
391	Sedative Anxiolytic Intoxication
392	Sedative Anxiolytic Intoxication Delirium
393	Sedative Anxiolytic Withdrawal
394	Sedative Anxiolytic Withdrawal Delirium
395	Sedative Anxiolytic-Induced Amnestic Disorder
396	Sedative Anxiolytic-Induced Anxiety Disorder
397	Sedative Anxiolytic-Induced Dementia
398	Sedative Anxiolytic-Induced Mood Disorder
399	Sedative Anxiolytic-Induced Persisting Amnestic Disorder
400	Sedative Anxiolytic-Induced Persisting Dementia
401	Sedative Anxiolytic-Induced Psychotic Disorder
402	Sedative Anxiolytic-Induced Psychotic Disorder With Delusions
403	Sedative Anxiolytic-Induced Psychotic Disorder With Hallucinations
404	Sedative Anxiolytic-Induced Sexual Dysfunction
405	Sedative Anxiolytic-Induced Sleep Disorder
406	Sedative Anxiolytic-Related Disorder
407	Sedative Hypnotic
408	Selective Mutism
409	Separation Anxiety Disorder
410	Severe Mental Retardation
411	Sexual Abuse of Adult
412	Sexual Abuse of Child
413	Sexual Aversion Disorder
414	Sexual Disorder
415	Sexual Dysfunction
416	Sexual Masochism

417	Sexual Sadism
418	Shared Psychotic
419	Shared Psychotic Disorder
420	Sibling Relational Problem
421	Sleep Disorder Hypersomnia
422	Sleep Disorder Hypersomnia Type
423	Sleep Disorder Insomnia
424	Sleep Disorder Insomnia Type
425	Sleep Disorder Mixed
426	Sleep Disorder Mixed Type
427	Sleep Disorder Parasomnia
428	Sleep Disorder Parasomnia Type
429	Sleep Terror Disorder
430	Sleepwalking Disorder
431	Social Phobia
432	Somatization Disorder
433	Somatoform Disorder
434	Specific Phobia
435	Stereotypic Movement Disorder
436	Stuttering
437	Substance Abuse
438	Substance Dependence
439	Substance Induced Psychotic Disorder
440	Substance Induced Psychotic Disorder With Delusions
441	Substance Intoxication
442	Substance Withdrawal
443	Substance-Induced Amnestic Disorder
444	Substance-Induced Anxiety Disorder
445	Substance-Induced Dementia
446	Substance-Induced Mood Disorder
447	Substance-Induced Persisting Amnestic Disorder
448	Substance-Induced Persisting Dementia
449	Substance-Induced Psychotic Disorder
450	Substance-Induced Psychotic Disorder With Hallucinations
451	Substance-Induced Sexual Dysfunction

452	Substance-Induced Sleep Disorder
453	Substance-Related Disorder
454	Tic Disorder
455	Tourette's Disorder
456	Transient Tic Disorder
457	Transvestic Fetishism
458	Trichotillomania
459	Vaginismus
460	Vascular Dementia
461	Vascular Dementia With Delirium
462	Vascular Dementia With Delusions
463	Vascular Dementia With Depressed Mood
464	Vocal Tic Disorder
465	Voyeurism

Appendix C: Association Words List

No.	Verbs
1	Affect
2	Associate
3	Candidate
4	Cause
5	Connect
6	Contribute
7	Correlate
8	Demonstrate
9	Downregulate
10	Dysregulate
11	Implicate
12	Indicate
13	Interact
14	Involve
15	Lead
16	Link
17	Mediate
18	Modulate
19	Observe
20	Plays
21	Predict
22	Predispose
23	Reflect
24	Reinforce
25	Relate
26	Susceptibility
27	Susceptible
28	Target
29	Underlie
30	upregulate