National Center for Scientific Research Demokritos

A MASTER THESIS ON

Weakly-Supervised Fine-Grained Semantic Indexing of Biomedical Literature using Citations

SUBMITTED TO THE NATIONAL CENTER
FOR SCIENTIFIC RESEARCH DEMOKRITOS
AND THE UNIVERSITY OF PELOPONNESE
IN FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

Master of Science

in Data Science

BY

Voulgari I. Eleni **Registration No: 2022201704005**

Under the Guidance of
Dr Krithara Anastasia

University of Peloponnese

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS
School of Economy, Management and Informatics
Erythroy Stavroy 28, 22131, Tripolis, Greece
2019 - 2020

National Center for Scientific Research Demokritos
Institute of Informatics & Telecommunications
Neapoleos 27, Ag. Paraskeui 15341

C E R T I F I C A T E

This is to certify that the Master's Thesis entitled

Weakly-Supervised Fine-Grained Semantic Indexing of Biomedical Literature using Citations

Submitted by

Voulgari I. Eleni **Reg. Number:17005**

is a bonafide work carried out under the supervision of Dr. Krithara Anastasia and it is submitted towards the fulfillment of the requirement of the National Center for Scientific Research Demokritos and the University of Peloponnese, for the award of the degree of Master of Science (Data Science).

Dr. Krithara Anastasia (Supervisor)	Prof. Skiadopoulos Spiros (Committee)	Dr. Giannakopoulos George (Committee)

Place: Athens Date: 14th December 2020

Acknowledgements

A fruitful work of Master Thesis is the outcome of support, encouragement, inspiration, supervision and collaboration of all the parties involved during study. I am genuinely thankful to present my master thesis titled: "Weakly-Supervised Fine-Grained Semantic Indexing of Biomedical Literature using Citations". I would like to give my candid thanks to our Director of the Institute of Informatics & Telecommunications (II&T) at NCSR "Demokritos", Dr. Karkaletsis Vangelis, for giving me the opportunity to present this thesis and to my respected guide, Dr. Krithara Anastasia, for their great support and motivation. I would also like to express my heartfelt appreciation to Nentidis Anastasios, Research Associate in NCSR Demokritos, for their assistance, inspiration and encouragement, without which my successful master thesis would not have been possible. Last but not least, I have to assert my warmth towards the involved staff members of National Center for Scientific Research Demokritos and my major gratitude to my family and friends, for their proper reinforcement and assistance.

Abstract

Semantic indexing of biomedical literature is essential for plenty of the research areas in the field of bioinformatics, such as data mining and knowledge retrieval. Annotations of biomedical research publications with Medical Subject Headings (MeSH) result in coarse grained indexing, due to the fact that the terms assigned are the MeSH descriptors, which may correspond to various related but disparate biomedical concepts. These semantic annotations may not provide adequate information to professionals in need of extracting more specific domain knowledge. In this Master's thesis, we suggest a methodology, in which a training dataset is enriched with citations' or/and references' semantic features and then used to train an available concept-level automatic annotator, so as to investigate possible changes in its performance. This approach is evaluated on Alzheimer's Disease MeSH related narrower concepts. The results indicate that, under the proper choice of classifiers and the appropriate definition of the input parameters, the performance of the classifiers, trained on the enriched dataset can surpass that of the base classifiers. The best classifier's performance is obtained, when the training dataset contains the semantic features from both citations and references.

Technical Key Words: MeSH, semantic indexing, biomedical literature, weak supervision, citation, reference

Contents

List of Figures

1 Introduction

1.1 Motivation

A considerable part of the immense amount of biomedical data, available nowadays, consists of plain text, such as explanations of clinical trials, electronic health records, information on adverse events and research publications. Such texts are written in scientific language, including definitions, expressions and terms, necessitating the creation of a terminology to formalize and catalog these scientific terms and concepts. The required terminology is also essential for the process of Information Retrieval (IR), for example when indexing publications or making queries on the text. Therefore, since the assignment of excerpting terms of a domain, by hand, is time and cost consuming, researchers focus in designing automated methods to aid know-how professionals to catalog the terms and concepts of a domain in the form of a vocabulary and automatically annotate biomedical publications. This allows users to retrieve information from a vast amount of biomedical text data efficiently and effortlessly using semantic search.

Most often the assigned tags are broad concepts, which describe the general notion of the biomedical terms, leaving out the more fine-grained concepts. This leads to the nonexistence of an indexing with such a granularity to facilitate scientists to explore more specific regions of biomedical literature. This tremendously affects the branch of diseases, in which narrower concepts very often depicts different types of a disease. The scientific community would benefit from an automated fine-grained partitioning of the related literature as it can efficiently reveal variations between the types of patients and provide accurate information for medicine applications.

This thesis belongs in the aforementioned field of biomedical semantic indexing, and specifically the annotation of biomedical text data with fine-grained scientific concepts. We are specifically interested in scientific literature annotation, thus indexing of biomedical research publications. Publication indexing is a significant part of knowledge extraction: researchers publish their work and results, which can be studied by others, to gain knowledge of work on a specific field. Fine-grained

indexing will provide more specialised search results, saving researchers' time and effort.

1.2 Purpose and Research Questions

The goal of this thesis is to attempt to exploit the semantic information provided by all articles that cite or are cited by a piece of biomedical literature to accomplish its fine-grained semantic indexing with the relevant narrower concepts of a MeSH descriptor. To this end, the preliminary work on fine-grained semantic indexing, based on weak supervision [1] is further extended, to make use of such information and the system is being tested on the case of Alzheimer's Disease.

The research questions, this thesis attempts to address, are the following:

- Are the semantic features of citations and references being selected in the top features for training of the models? Are they considered significant?
- Do the above semantic features help in the classification?
- Can the combination of both citations' and references' features make a difference to the performance of the classifiers?

1.3 Scope and Limitations

This study focuses on the effect of the semantic information, provided by cited and referenced papers, on the performance of machine learning classification algorithms. This piece of information is used to produce semantic features for the enrichment of the features of the training dataset and not as labels.

1.4 Target group

The work of this thesis can be beneficial to several types of members of the scientific community, including those involved in automated biomedical semantic

indexing, those in need to extract literature under terms of fine granularity level and those in research of the effect of semantic features of citations and references on a fine-grained biomedical indexing method.

1.5 Outline

This thesis incorporates ideas from information retrieval from document repositories, thus textual information, semantic indexing and distant learning. Following the introductory chapter is Chapter 2, which sets the background and summarizes the related work. Chapter 3 presents the experiment design and implementation, along with the dataset development and the sub-activities of the experiment. Chapter 4 describes all the experiments performed and Chapter 5 demonstrates and evaluates the experimental results. The findings achieved by the proposed method will be compared against the ones achieved by the baseline method. Finally, in Chapter 6, potential directions for future work are being discussed, a conclusion of this thesis is provided and its contribution is outlined.

2 Background & Related work

2.1 Background

The biomedical concept catalog, developed and maintained by the National Library of Medicine (NLM) of the United States, called the Medical Subject Headings (MeSH) thesaurus, is a controlled and hierarchically-organized vocabulary, used for indexing, cataloging, and searching of biomedical and health-related information. It includes the subject headings appearing in MEDLINE/PubMed, the NLM Catalog, and other NLM databases[1]. So, each piece of biomedical literature is annotated with a set of MeSH terms that best depicts its content.

MeSH headings, or descriptors, are grouped into 16 categories each divided further into subcategories. Descriptors are hierarchically arrayed within each subcategory in up to thirteen hierarchical levels, from the most generic to the most specific. These lists are called trees because of the branching structure of the hierarchies. Each MeSH descriptor appears in at least one position in the trees and can appear in as many additional trees as it may fit.

MeSH descriptors contain one or more concepts, and each concept contains one or more terms, which are synonymous, thus a concept is the shared meaning of synonymous terms. A unique identifier (ConceptUI) is assigned to each concept, so terms sharing the same ConceptUI are synonymous. Additionally, each descriptor has a preferred concept - the name most often used to refer to the descriptor - and each concept has a preferred term - being the name of the concept.

One of the most significant relationships in the produced Semantic Network is the broader-narrower relationship. This relationship encapsulates the hierarchies between the biomedical concepts. Narrower concepts may be viewed as sub-types. For example (Fig.1) one concept in a descriptor may be narrower than the preferred concept. It is essential to remember that granularity levels differ across the Network. There are descriptors with many semantic types and other with little or none.

[1] `https://www.nlm.nih.gov/mesh/meshhome.html`

Figure 1: Preferred and narrower concepts of Alzheimer Disease descriptor

The manual annotation of publications has become very challenging for indexers due to the increasingly vast amount being published and indexed each year (Fig.2 [2]). To overcome the difficulties, NLM created the Medical Text Indexer (MTI), an automated tool, which identifies the relevant MeSH terms in the text of publications and returns them to the indexers, thus being extremely helpful and overall very accurate [3]. The above information extraction task of recognizing biomedical terms in natural language text, includes identification and mapping of each extracted entity to a concept through biomedical named entity recognition.

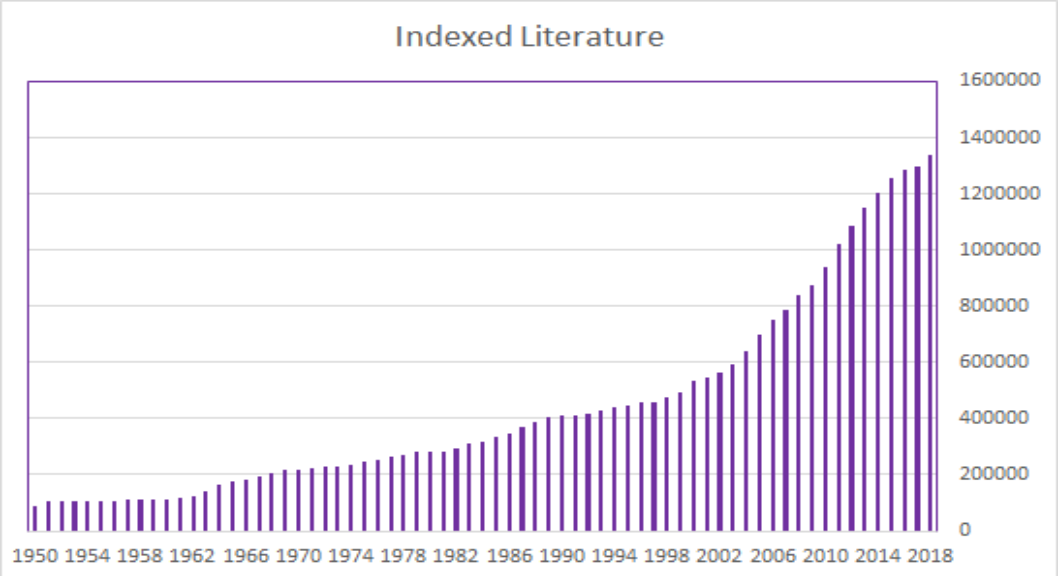


Figure 2: Number of indexed papers in Medline.

2.1.1 Semantic Scholar

In 2015, Semantic Scholar, founded by the nonprofit Allen Institute for Artificial Intelligence (AI2), began as a search engine for computer science, geoscience and neuroscience. It is an example of artificial intelligence-enabled search engine to respond to the inability of researchers to keep pace with all the publications in their disciplines. The project's goal is to automate text-based learning to cope with the increasing amount of scientific documents published each year [4]. By October 2019, its number of included papers had grown to more than 175 million, expanding to all research areas. But its success is that it eliminates the long tail of search results, allowing scientists to get up to speed on their disciplines easily, by showing only directly relevant publication.

Semantic Scholar's document representation contains all the necessary information about the citation and the references that this work want to exploit. There are two choices for accessing Semantic Scholar database. Either through its downloadable files or its RESTful API one can link to Semantic Scholar items and pull information about individual records on demand [2]. The "Paper Lookup" API method returns returns a JSON structure, that describes the specific paper, including the ids of the papers that have cited it and the ones it referenced (Fig.3).

Figure 3: Snapshot of JSON structure for the paper with PMID:24802362

[2] `http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/`

2.2 Related Work

The areas relating to this thesis can be split up in three main areas; biomedical semantic indexing, distant supervision learning and citation-based classification of literature.

2.2.1 Biomedical Semantic Indexing

Domain-based information retrieval uses predefined concepts that provide a valid source of information for indexing documents by attenuating the problem of term mismatch that IR systems encounter [5]. Document indexing can be performed manually or automatically. Human annotators with expert knowledge of terminologies, which are highly experienced on the domain, conduct manual indexing. A method less costly and time-consuming is automatic indexing and could be an assistance to or a full substitute of the manual process. Automatic indexing is basically assigning a number of terms to a document, which terms denote some concepts. A unique preferred term, used for indexing and one or more non-preferred terms, used for retrieval, represents each concept. As mentioned in the introduction, the Medical Text Indexer (MTI) is the main component of the NLM Indexing Initiative and has been used in both semi-automated and fully automated indexing based on MeSH concepts [6].

Concept extraction is the pillar for automatic document indexing and retrieval and constitutes an essential technique for identifying concepts of specified terminologies in NLP [7]. For many reasons, automatic concept extraction from medical text is a challenging task. First, terms which represent biomedical concepts usually consist of multiple words, that as a whole, lead to a more specific concept. For example, the word "Disease" alone can mean any type of disease, whereas the term "Alzheimer's Disease" may mean any type of this specific disease and the term "Familial Alzheimer's Disease" is the degenerative disease of the brain, caused by a single genetic mutation that is transmitted through families. Second, the same concept applies to many interchangeable terms. Phrases such as "Senile Dementia", "Dementia of Alzheimer Type" and "Alzheimer Syndrome", for example, can be

used to refer to the definition of "Alzheimer's Disease", as defined in MeSH.

Studies on the extraction of biomedical concepts have been conducted extensively in the literature [8, 9]. Biomedical concept extraction approaches can be classified into four categories: dictionary-based, statistical, rule-based and machine learning methods.

Dictionary-based concept extraction methods use existing terminologies to map free text to dictionary entries. A system, using an estimated string matching to identify protein and gene names and their variations, have been proposed [10], where protein terminology and text are both encoded using the four letter (A,C,G,T) nucleotide code.

Several statistical approaches have been suggested to identify general terms. For instance, a method called "C/NC value for recognizing technical terms" is used to extract technical terms from literature, in digital libraries [11], but also has been used to identify concepts in biomedical literature [12]. Frequency is their most common measure, with term frequency (tf) that counts the frequency of a term in a document and inverse document frequency (idf), that decreases the weight of terms that occur very often in the corpus and increases the weight of terms that rarely occur. Multiplying those two, gives the tf-idf weight that shows how relevant a word is to a document in a coprus.

Rule-based methods typically contain the creation of patterns, that use lexical and morphological properties, to assign structures to specific concepts [13]. Such methods are considered to be very time-consuming, and are typically difficult to be applied to other more general tasks. The Catalog and Index of Online Health Resources in French (CISMeF) system, developed at Rouen University Hospital, categorizes text according to Metaterm (MT), using rules. [14].

Machine learning (ML) supervised methods use an annotated set of documents to train classifiers, that attempt to learn how to match free text with the predefined concepts (classes). Hidden Markov models (HMM) and specific orthographic features have been used to identify terms of a set of ten classes [15]. An essential method for supervised ML is the support vector machines (SVM). Multi-class

SVMs have been trained on manually annotated GENIA corpus for the purpose of named entity recognition (NER) [9]. More specifically, the method aims to predict tags, identifying named entities based on "position" features (e.g., POS, prefix, suffix), as well as pattern similarities and HMM state features to fix data sparseness. An ML method, submitted to the BioASQ challenge [16], uses dense word vectors, which results in significant dimensionality reduction (compared to the BOW representation), reducing the training time, without affecting the performance of the classifiers [17]. Another work, also submitted to the BioASQ challenge, that uses a MUlti-Label Ensemble method (MULE) giving very positive results [18]. More recent work includes the use of deep learning methods, such as convolutional neural networks and deep multi-tasking models, to achieve even better performance in the task of biomedical literature indexing [19, 20]. Supervised learning algorithms, however, demand for a significant amount of annotated training examples, thus facing difficulties when such training set is not available.

2.2.2 Weakly Supervised Learning

Supervised ML methods, such as classification, require a significant amount of training examples with ground-truth labels. However, in many tasks such information can be difficult to obtain, due to the high cost of data labeling process. So, it is required for the aforementioned ML methods to be able to perform under weak supervision, thus using instances annotated with weak labels. These labels can be incomplete, inexact or inaccurate [21]. Weak labelling technique often uses a heuristic or a set of heuristics to assign labels to the unlabeled instances of the dataset or to abstain.

Several studies have been conducted on text classification using weak labelling. One of these, uses a rule-based NLP algorithm to produce weak labels for the training data and then uses these pre-trained word embeddings as representation features for classification algorithms in clinical text [22]. Another work, in order to automatically distinguish multiple meanings of the same word and construct a labelled contextualized corpus, uses representations of word occurrences and user-provided seed words [23].

Last but not least, the work extended by this thesis focuses on training a fine-grained concept annotator on literature published about a specific disease. More specifically, fine-grained concept tags are assigned heuristically to abstracts, based upon those concept occurrence in the text [1]. This work will be further described on the next chapter as it is the pillar of this thesis.

2.2.3 Citation-based Methods

The most common citation-based method is citation context analysis, which refers to the analysis of the text surrounding the citation mark. This analysis can provide a useful source of concepts, that are not found in the text itself, and a richer feature representation, thus improving information retrieval [24]. Likewise, cited references are used as a query expansion method, extending the information of a piece of literature, for the retrieval of related biomedical documents [25]. Furthermore, citation context analysis gives the possibility to identify significantly related articles in a context-sensitive way [26] and can provide different kind of information, giving access to alternative biomedical literature search results than traditional search engines [27].

An alternative method to use is frequency citation analysis, which ranks documents based on their co-citation frequency with an article, and the frequency of all citations that cite or are cited by this article [28]. Another approach is based on identifying all the references cited in the document and, using the classification metadata of extracted references from existing libraries and a weighting algorithm, infers the most likely class for the document [29].

3 Method Design

The method design section aims to provide an appropriate framework for the study, determining the choices made, regarding the study approach.

3.1 Baseline Method

The pillar work, which this thesis extends [1], is based on weak supervision labelling to achieve fine-grained semantic indexing of biomedical literature, thus assign narrower MeSH concepts to documents. As mentioned in the previous chapter, classification models need labelled data for training. To overcome the difficulty of the absence of fine-grained indexed datasets, this method makes use of weak labels, generated automatically, as well as the broader-narrower relationship of concepts in MeSH descriptors. For each of the MeSH descriptors, a predefined collection of fine-grained labels is specified, based on its broader-narrower relationships.

The first stage of this method is to generate and assign weak labels to a collection of selected documents (Fig.4). The collection is obtained through Medline/Pubmed [3], using the Entrez Programming Utilities, an interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI) [4]. The abstracts and the titles of all the documents, annotated with the desired MeSH descriptor, are included in the collection. For the purpose of the weak labelling, MetaMap [30], an inclusive biomedical NLP tool for effective mapping of biomedical text to the UMLS Metathesaurus [5], is used. So, an article is given a corresponding weak label for any narrower of the preferred concept, that occurs in its text, thus the instances are multi-labelled with the probability of these labels to be noisy.

[3] https://pubmed.ncbi.nlm.nih.gov/
[4] https://www.ncbi.nlm.nih.gov/books/NBK25497/
[5] https://uts.nlm.nih.gov/home.html/

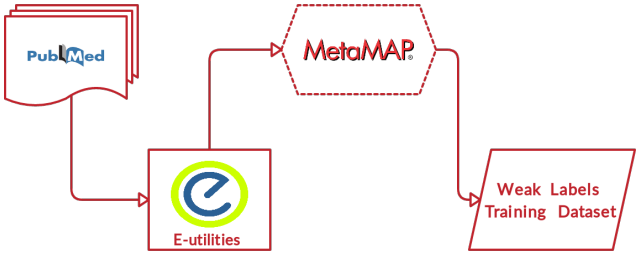


Figure 4: Weak Labelling Process

The second stage comprises of all the steps needed for the model development (Fig.5). During the data pre-processing step several useful structures are created using the documents' text along with the weak labels. The dataset is defined to be multi-labelled and it is also split into train and test. Also several transformations take place, like binarizing the multi-labels, tokenizing the text, adding the CUI information to the matrix and producing the tf-idf representation [31]. Both lexical and semantic features are produced, using the documents' titles and abstracts, as well as the concept information from MetaMap, creating the feature-label matrix. The instances of this matrix represent the documents, thus the title combined with the abstract for each document. The features consist of the lexical tokens of the documents and the added CUI features from MetaMap. The values of the matrix are the tf-idf representation for the token features and binary (0/1) or absolute frequency values for the added MetaMap features. For each instance the corresponding weak labels are assigned (Fig. 6).

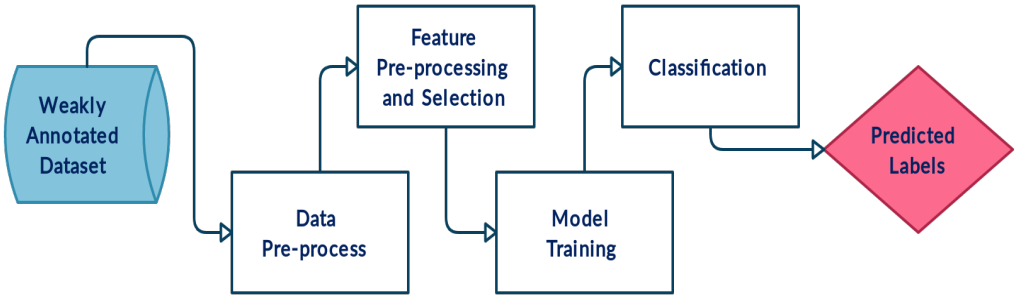


Figure 5: Suggested Model Development Process

In the feature pre-processing and selection step, an analysis defines the weights of each feature and finds the features that are most useful for predicting the fine-grained labels. Furthermore, the top (k) features are chosen, according to the number of (k) the system receives.

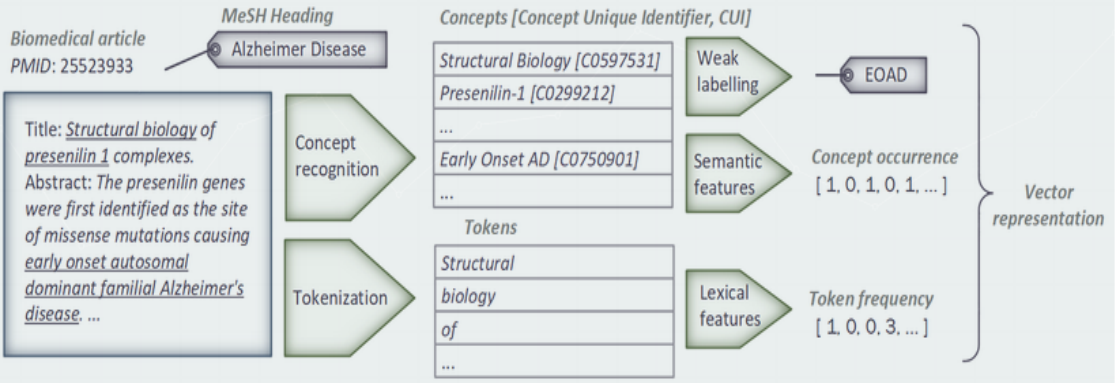


Figure 6: Baseline Method Features

Four models are trained for this multi-label classification task: a Decision Tree Classifier (DTC), a Random Forest Classifier (RFC), a Linear Support Vector Classifier (LSVC) and a Logistic Regression Classifier (LRC). For testing purposes two small sets of documents, one with random samples and one with balanced samples, were manually annotated with fine-grained concepts by experts.

Several experiments with different configurations showed that models trained with weak labels created by fine-grained concept occurrence can give better results than the heuristic of the concept occurrence alone, under certain conditions. The iterative model training on the predicted weak labels did not showed any signs of improvement. For a more detailed explanation of this method and the way to make use of it, please refer to: `https://github.com/tasosnent/BeyondMeSH`.

3.2 Proposed Method

The work of this thesis relies upon the baseline method and extends it in a way, that it exploits citations and references information. For each document of

the dataset the relevant citations and references information is downloaded as ids, through Semantic Scholar. Following the baseline method, in the proposed method, the pre-processing step is enriched by adding the concepts that are present in the citations as features (Fig.7). These features can take either binary values (existence/non existence), or the absolute frequency of the concept occurrence. A piece of publication might not have citations or references that correspond to those of the pool of publications of the specific dataset. In this case, the corresponding features are left empty and do not take part in the models' training.

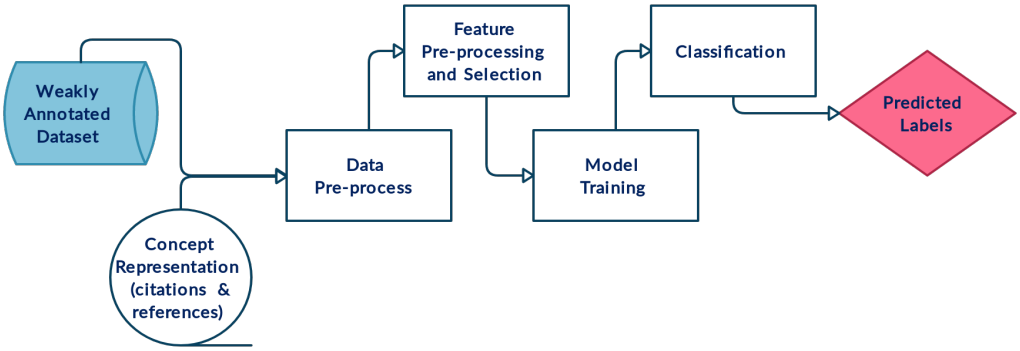


Figure 7: New Model Development Process - Added citations / references

But what do we consider as references and what as citations? These two concepts are substantially the two sides of the same coin. References are those articles which the specific article cited and citations are those articles that cited the specific paper (Fig 8).

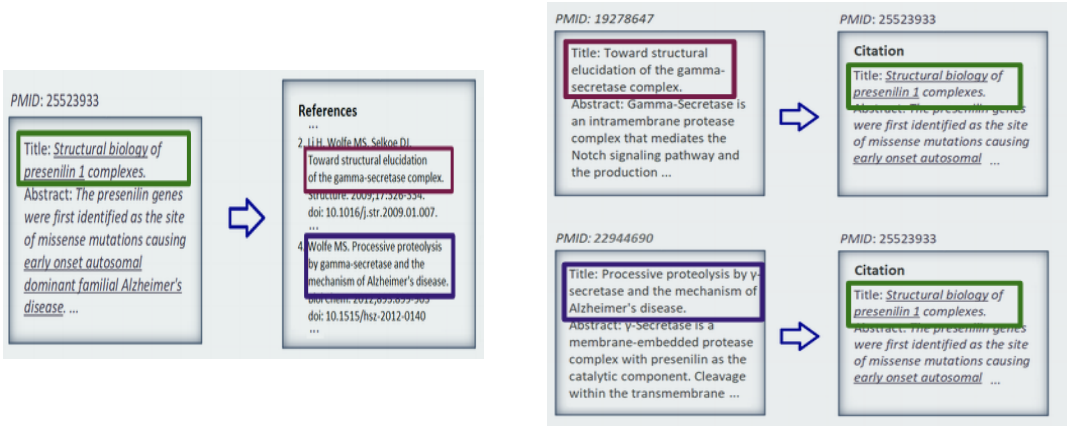


Figure 8: What are the **references** of an article (left) & what the **citations** (right)?

3.2.1 Enriched Dataset Development

As a preliminary step to further enrich the existing WS dataset and due to the fact that at the time this task was undergoing, the API did not accept the PMID information as keyword search for the database, the version of 2019-10-01 semantic scholar corpus was downloaded[6]. It consists of 178 zipped files of about a million JSON structures each, with every structure corresponding to a piece of literature. An example of the representation of a document, as well as the explanation of each of JSON's item fields can be found in `http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/`.

As the baseline dataset already contains the PMID, the title and the abstract of the documents, the extra fields of interest are the "id", "inCitations", "outCitations" and "doi" fields. Because the inCitations (list of paper IDs which cited this specific paper) and outCitations (List of paper IDs which this specific paper cited) sections of a JSON object, are lists with Semantic Scholar IDs, it is important to keep and match each document with its Semantic Scholar ID. Furthermore, there are JSON items that do not contain the PMID of the document, so a workaround to find these documents on Semantic Scholar corpus is to convert the PMID to DOI.

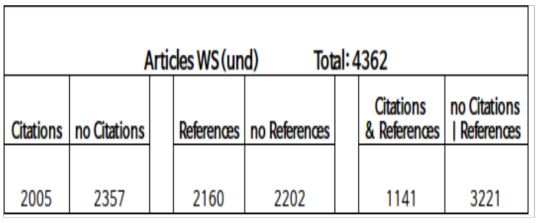
From the downloaded files, an initial selection of JSON structures, containing the word "Alzheimer" in their abstract, was performed to reduce the volume of the data in search. This led to the extraction of 133.073 JSON items. Then, a second selection according to the PMID was performed, leading to the extraction of another 40.537 items. The rest of the baseline dataset's PMIDs, not found, were transformed into DOIs, by accessing https://www.ncbi.nlm.nih.gov/pubmed/?term=%PMID, for each PMID and then using the BeautifulSoup, a Python library for pulling data out of HTML and XML files, the field of DOI is extracted from the HTML page and matched to its corresponding PMID. Then, a complementary selection on Semantic Scholar items led to the extraction of another 25.482 items. As a last effort to obtain all the documents of WS dataset, a title matching was performed against

[6] `http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/download/`

the undetected documents and the JSON items, reaching a total of 66.184 items detected.

The decision to ultimately use the under-sampled version of WS dataset (WSund), led to the selection of its 4.362 documents. By the time this decision was taken, the RESTful API of Semantic Scholar was updated to have the feature of using the PMID to access its JSON files, so the process was simplified. So, for each PMID the JSON item of Semantic Scholar database is accessed and therefore, all the necessary values from "id", "inCitations" (aka citations), and "outCitations" (aka references) fields are extracted.

Two files are created, one for the "citations" field and one for the "references" field. The aforementioned files contain the corresponding PMID, the Semantic Scholar id and the list of citation and references for each document, respectively. The need to use MetaMap extracted concepts, led to the preservation only of those citations' and references' PMIDs, that already exist in the dataset. After the exclusion of the articles not contained in the dataset, in the citation file there are 2.005 articles that have citations and in the reference file there are 2.160 articles, having references. When these files are combined the articles that have both citations and references count to 1.141 (Fig.9).



Articles WS(und) Total: 4362						
Citations	no Citations		References	no References	Citations & References	no Citations | References
2005	2357		2160	2202	1141	3221

Figure 9: Number of articles in WS(und) for each category

Then, the two aforementioned files are converted into concept representation files, utilizing the concept file, which contains the results of MetaMap weak indexing, concluding in files with the PMID and a list of strings with the frequency of concepts of each document's citations or references.

But how are the new semantic features integrated into the existing dataset? The concepts of MetaMap are obtained for each reference of an article and then all these concepts from each category, are aggregated in one representation creating the new semantic features which are then appended to the already existing features. The aggregation method is the append of each concept to a list and then the assignment of the binary (0/1) or the frequency value of a specific concept to all the citations together (Fig. 10). The same procedure is followed in the citation and combination cases.

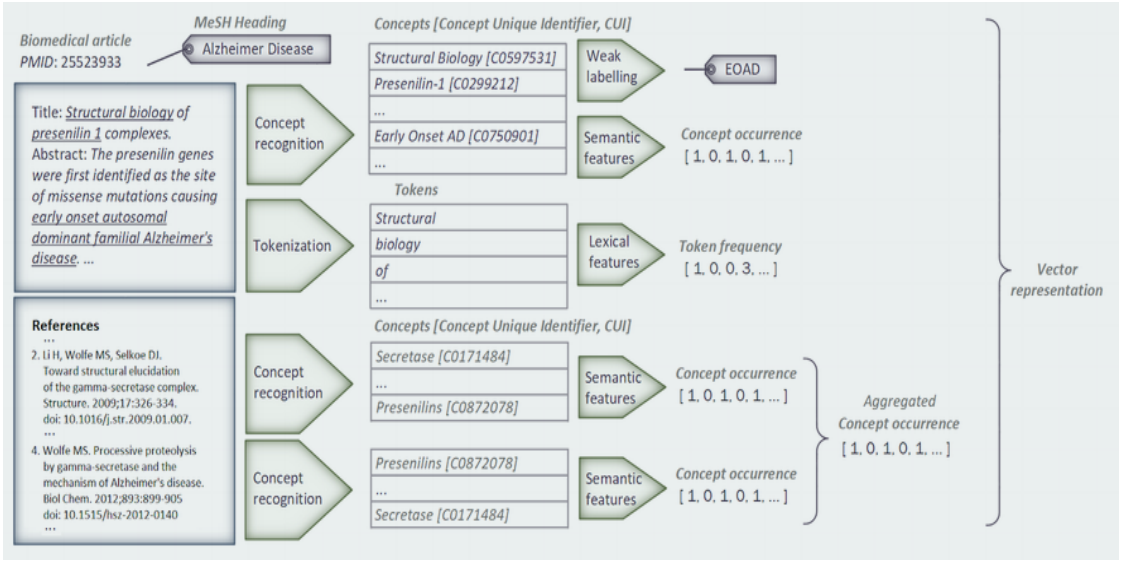


Figure 10: Integration of the new References features

4 Experiments

4.1 Experiment Settings

This section describes the settings of the experiments conducted, to assess whether the proposed method can help improving the classification accuracy. Three different settings were examined: (a) use of citations, (b) use of references, and (c) combination of both citations and references. For the purpose of evaluating the new method, the case of Alzheimer's Disease is chosen along with the under-sampled version of the WS Alzheimer's Disease training dataset (WSund), as it is reduced in size and deals with the over-representation of the preferred concept-label ("Alzheimer Disease").

Only the citations and references that already belong to the dataset are of interest, because the results of MetaMap are needed. Not all papers' citation and reference information is available through the selected source. The documents that do not have citations or references respectfully are left in the training dataset, but omitted from the test datasets, because this research aims to discover whether the classifiers are giving better predictions exactly when the aforementioned information is available.

"Alzheimer Disease" MeSH descriptor, on which the proposed new method is tested, comprises of the preferred concept "Alzheimer Disease" and six more narrower concepts: "Acute Confusional Senile Dementia (ACSD)", "Dementia, Presenile (PD)", "Alzheimer Disease, Late Onset (LOAD)", "Alzheimer's Disease, Focal Onset (FOAD)", "Familial Alzheimer Disease (FAD) and "Alzheimer Disease, Early Onset (EOAD)".

The WS(und) dataset, (under-sampled version of the initial Alzheimer Disease dataset) contains 4,362 papers annotated with the Alzheimer Disease descriptor, obtained from PubMed. For the purposes of this work, as far as textual information is concerned, only the abstract and title of papers is used. Based on the occurrence of the narrower concepts found by MetaMap, the weak labelling method

heuristically assigned fine-grained labels of narrower concepts' type to 1,923 of the articles, labels of the preferred concept's type to 3.000 articles (752 of them, also is annotated with at least one fine-grained narrower label), leaving the remaining 191 articles without any label, summing up to 5.004 MetaMap annotations. Two of the aforementioned narrower concepts, "Acute Confusional Senile Dementia (ACSD)" and "Alzheimer's Disease, Focal Onset (FOAD)" do not occur in any of the articles, according to MetaMap, thus these two labels are excluded from the experiments. The distribution of fine-grained labels in the annotations above, is shown in Fig.11.

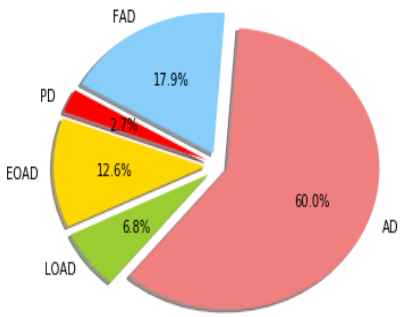


Figure 11: Distribution of fine-grained MetaMap annotations

A total of 12 experiments, with different kind of parameters, were conducted. Each of these experiments include the training of all 4 classifiers (DTC, RFC, LSVC, LRC) with 7 different set of features (5,10,20,50,100,500,1000), concluding to 336 trained models. The exact same experiments were conducted, using the baseline method, for comparison reasons. The flow of the different versions of the experiment is shown in Fig.12. We also experimented with both binary and frequency as feature values. Lastly, for the feature selection process, where all features are weighted according to their relevance to the target outcome, two approaches were used, one based on the analysis of variance (ANOVA) and one based on chi-square.

For the evaluation of the proposed method, two small sets for each category were used, one with random samples (MA1) and one with a selection of samples such that there is no over-representation of the label that represents the broader meaning, i.e. AD (MA2). Articles selected for each category contain the corresponding additional information. These sets have been golden-labeled, with their true fine-grained labels by expert scientists. The baseline method is tested again for all the experiments

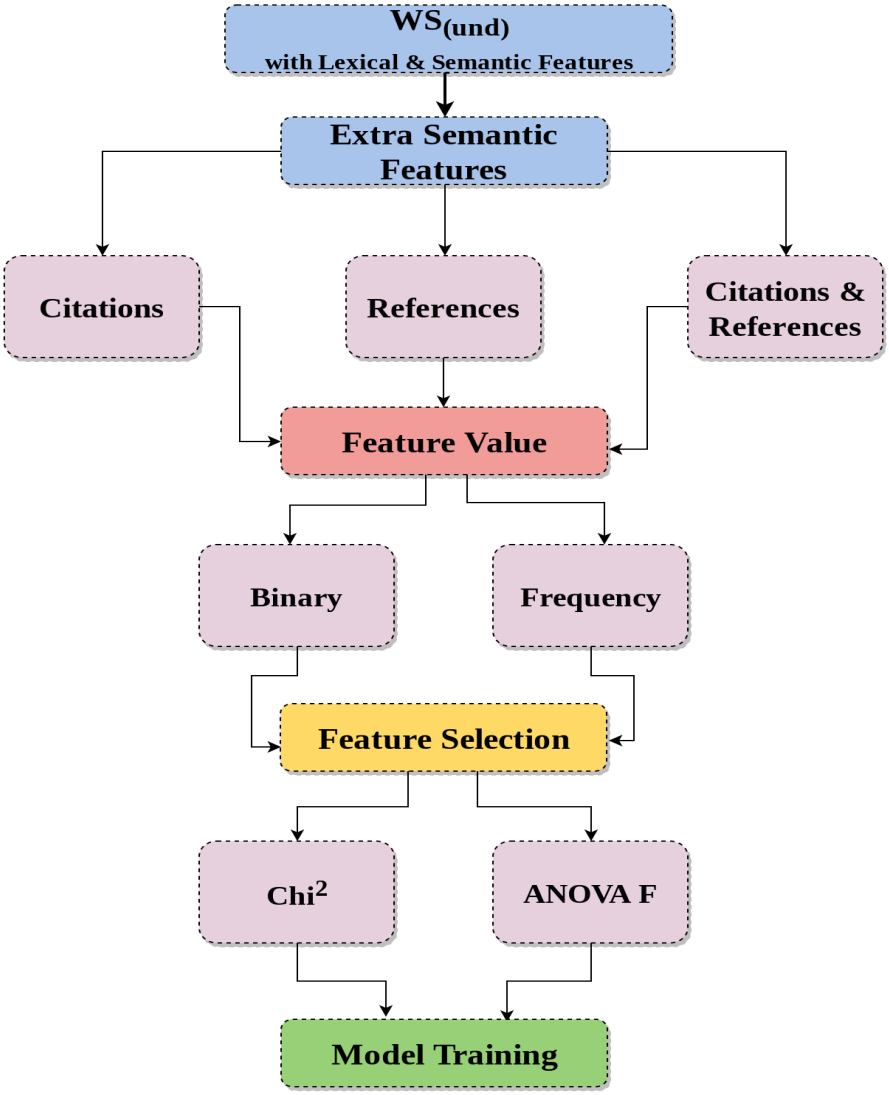


Figure 12: Flow of experiments

separately, with the corresponding test sets.

An analysis of the fine-grained weak labels in each training and each test set is shown in the following table (Fig. 13). The first column contains the fine-grained labels of the entire dataset. In the next three columns the distribution of the fine-grained labels in the training datasets are shown, but only for the articles that actually contain the corresponding additional information. The rest of the columns show the number of the fine-grained labels for the test sets of each category respectively.

Use Case	Fine-Grained Weak Labels	WS(und) Dataset (4.162)	Training Sets			Test Sets					
			With Citations	With References	With Combination	Citations Case		References Case		Combination Case	
						Random MA1 (33)	Balanced MA2 (56)	Random MA1 (46)	Balanced MA2 (58)	Random MA1 (18)	Balanced MA2 (38)
Alzheimer Disease (AD)	EOAD	629	402	345	224	1	37	2	32	1	26
	LOAD	342	180	217	119	1	22	2	18	1	14
	PD	135	52	42	24	1	9	1	8	1	5
	FAD	898	563	514	354	8	38	6	40	4	29
	labels / total labels	2004 / 5004	1197 / 5004	1118 / 5004	721 / 5004	11 / 34	106 / 107	11 / 49	98 / 105	7 / 19	74 / 75

Figure 13: Analysis of the fine-grained weak labels contained in the training and in the test set

4.2 Evaluation Method

The proposed approach is evaluated in two levels. The first one concerns the feature selection approaches, while the second one concerns the overall accuracy of the approach. For the later, the macro-F1-score is used as a comparison metric between the results derived from the proposed and the baseline methods. For each experiment setting the aforementioned two approaches are being presented in the next chapter.

5 Results & Discussion

5.1 Results

This section aims to present an analysis of the feature selection results, for each category of experiments described in the previous chapter, as well as the comparison of the models' performance with that of the corresponding models of the baseline method and is divided into two subsections. The subsections are both further split into three parts, each corresponding to the experiments using the semantic features of a specific modality (citations, references, combination) for model training.

5.1.1 Feature Selection Analysis

The enriched datasets contain more than 45.000 features and only about 4.362 instances, making it necessary to select the features that have direct effect on the target values. Thus, just like in the baseline method, two statistical techniques are used, chi-Squared test and ANOVA correlation coefficient, as mentioned in the previous chapter. The maximum number of features selected is 1.000, enabling the algorithms to train faster and reducing the complexity of the models, making them easier to interpret. The results of the feature selection step can provide valuable insight on whether the new added features are important, playing a significant role in the training of the models.

Citations' Semantic Features

The enrichment of the dataset with extra citation semantic features leads to the addition of 10.308 new features, that correspond to the CUIs assigned by Metamap to the citation articles. Therefore, the proportion of these new features in a total of 48.032 features is 21.46%. Using the feature selection method for dimensionality reduction to 1.000 features, the proportion of the newly added features slightly increases in the case of chi-square statistic, but becomes more than tripled in the case

of ANOVA-F. The choice of binary or frequency values does not seem to play any role in the aforementioned proportion during the feature selection process (Fig.14). For cases where the number of selected features is small, the proportion changes, so the performance of the models might explain which proportion is best for training.

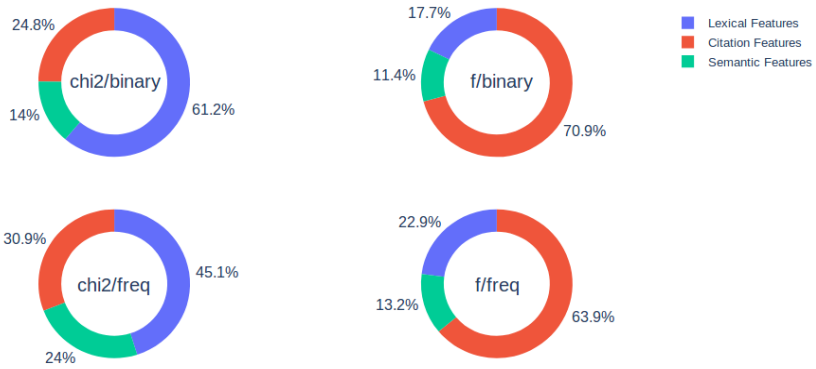


Figure 14: Proportions of selected top features using **Citations** - k=1000

Even though the choice of binary or frequency values does not seem to affect the amount of the selected new features, it surely has an influence on the position in which the first citation feature appears, combined with the statistical feature selection method used. As presented, in Fig.15, in the case of chi-square statistic there is a large gap between the position of the first citation feature. whereas in the AVONA-F case, no significant change is made in the aforementioned position.

	Binary Values	Frequency Values
Chi-square	86	5
ANOVA-F	16	14

Figure 15: Position of the first **Citation** feature in the top features

The position of the first new feature is important as it shows whether the statistical methods find that the new features are significant enough to be used in the models' training. For example, in the case of using the chi-square statistic and frequency values for the citation semantic features, the first citation feature appears in position number 5, having two semantic and two lexical features preceding. This first citation semantic feature represents the CUI of the term "psen1 gene" (C1418985). The corresponding identified lexical term "ps1" appears in position

number 17, showing that the citation semantic feature is considered more important for training than the actual lexical term. The same applies to the second citation semantic feature, representing the CUI of the term "mutations" (C0026882), for which the citation semantic feature lies in position number 9, whereas the lexical feature is found in position number 36 (Fig.16).

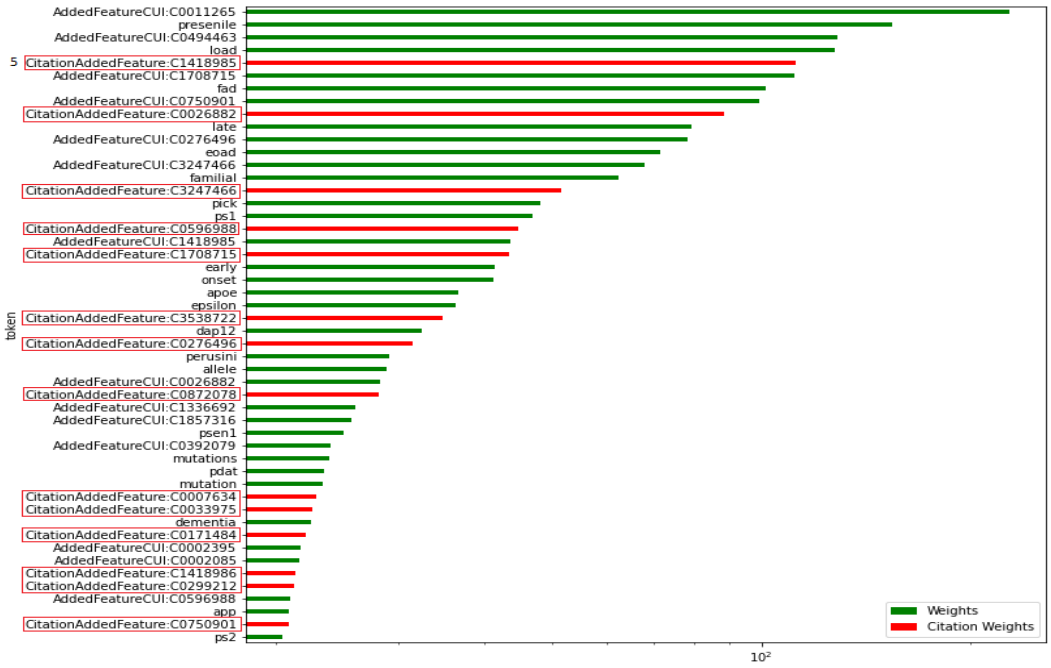


Figure 16: Features' Weights (chi2/frequency) - **Citations** Case

In some cases, where a small number of features (e.g. 5, 10 or even 20 features) are chosen by the user to be selected, the new features are not even present in the training dataset, but they seem to indirectly affect the ranking and, therefore, the selection of the top features. This happens because the occurrence of the new citation features have an impact on the weights assigned to all features, during the feature selection step.

The examination of the top 100 features selected, reveals the changes caused by the addition of the citations' semantic features. Fig.17 shows the features removed and the ones added to the 100-feature dataset of the baseline method. As observed, the dataset is enriched with 32 new citation semantic features, that represent several concepts, like "19q13" (C1520855) and others, not occurring in the features of the baseline method. Furthermore, for some CUIs, like "gamma-Secretase" (C0379528) and "Exons" (C0015295), the feature selection process prefers to choose the citation

semantic representation and not the article's itself. As expected, the features representing concepts that are actually the fine-grained concepts of AD, like "Alzheimer Disease, Early Onset" (C0750901), are being chosen in both the citation and the initial semantic form.

Removed Features	Added Features
AddedFeatureCUI:C0015295	CitationAddedFeature:C0015295
AddedFeatureCUI:C0015576	CitationAddedFeature:C0015576
AddedFeatureCUI:C0299212	CitationAddedFeature:C0299212
AddedFeatureCUI:C0379528	CitationAddedFeature:C0379528
AddedFeatureCUI:C1883559	CitationAddedFeature:C1883559
abeta42	CitationAddedFeature:C0007634
abo	CitationAddedFeature:C0010813
AddedFeatureCUI:C0005558	CitationAddedFeature:C0017337
AddedFeatureCUI:C0031734	CitationAddedFeature:C0026882
AddedFeatureCUI:C0052201	CitationAddedFeature:C0027882
AddedFeatureCUI:C1332412	CitationAddedFeature:C0030705
AddedFeatureCUI:C2347741	CitationAddedFeature:C0030956
AddedFeatureCUI:C2350277	CitationAddedFeature:C0032416
AddedFeatureCUI:C3642141	CitationAddedFeature:C0033975
adeoad	CitationAddedFeature:C0085151
atd	CitationAddedFeature:C0162638
bin1	CitationAddedFeature:C0171484
ctt	CitationAddedFeature:C0241888
eofad	CitationAddedFeature:C0276496
fibroblasts	CitationAddedFeature:C0542341
gene	CitationAddedFeature:C0596988
linked	CitationAddedFeature:C0750901
mutants	CitationAddedFeature:C0868928
photic	CitationAddedFeature:C0872078
precursor	CitationAddedFeature:C1418985
presenilins	CitationAddedFeature:C1418986
psen2	CitationAddedFeature:C1520855
risk	CitationAddedFeature:C1705543
s320f	CitationAddedFeature:C1708715
varad	CitationAddedFeature:C1833334
wild	CitationAddedFeature:C3247466
ε4	CitationAddedFeature:C3538722

Figure 17: Removed & Added Features (k=100, chi-square/frequency values) - **Citations**. The removed highlighted features on the left column represent the same concepts as the highlighted ones on the right column that were added.

References' Semantic Features

The enrichment of the dataset with extra reference semantic features leads to the addition of 8.956 new features, that correspond to the CUIs assigned by Metamap to the reference articles. Therefore, the proportion of these new features in a total of 46.680 features is 19.19%. Using the feature selection method for dimensionality reduction to 1.000 features, the proportion of the newly added features increases in both cases (chi-square & ANOVA-F) with the increase in the ANOVA-F case being more dramatic. Again, the choice of binary or frequency values seem to not play any role in the aforementioned proportion during the feature selection process (Fig.18).

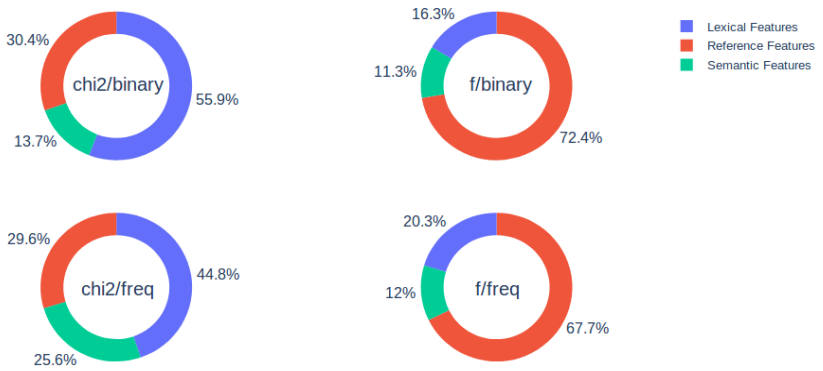


Figure 18: Proportions of selected top features using **References** - k=1000

As far as the effect of binary/frequency value on the position, in which the first reference feature appears, the results are very similar to those of the citations' features, with the difference being more profoundly in the case of chi-square statistic and not significant in the case of ANOVA-F (Fig.19). This could lead us to the conclusion that the source of the extra semantic features does not matter for the method to decide whether they are important.

	Binary Values	Frequency Values
Chi-square	123	4
ANOVA-F	17	14

Figure 19: Position of the first **Reference** feature in the top features

In the references case, the same first two concept CUIs, representing the terms "psen1 gene" (C1418985) and "mutations" (C0026882), are in positions 4 and 10, with their lexical corresponding features again being further down the list. In the chi-square statistic and frequency values case, more CUIs similar to the first are recognized as important, like "presenilin-1" (C0299212) and "mutant" (C0596988). The corresponding initial semantic features of these CUIs are found later on the list, showing that the references features might be more significant (Fig.20).

What is notable here is that the new semantic features (CUIs), corresponding to the fine-grained annotations of Alzheimer Disease, like "Presenile dementia" (C0011265), "Alzheimer Disease, Late Onset" (C0494463) and "Alzheimer Disease, Early Onset" (C0750901) are not found in the first 50 positions as one would expect, but they are found in the dataset, when 1.000 features are selected.

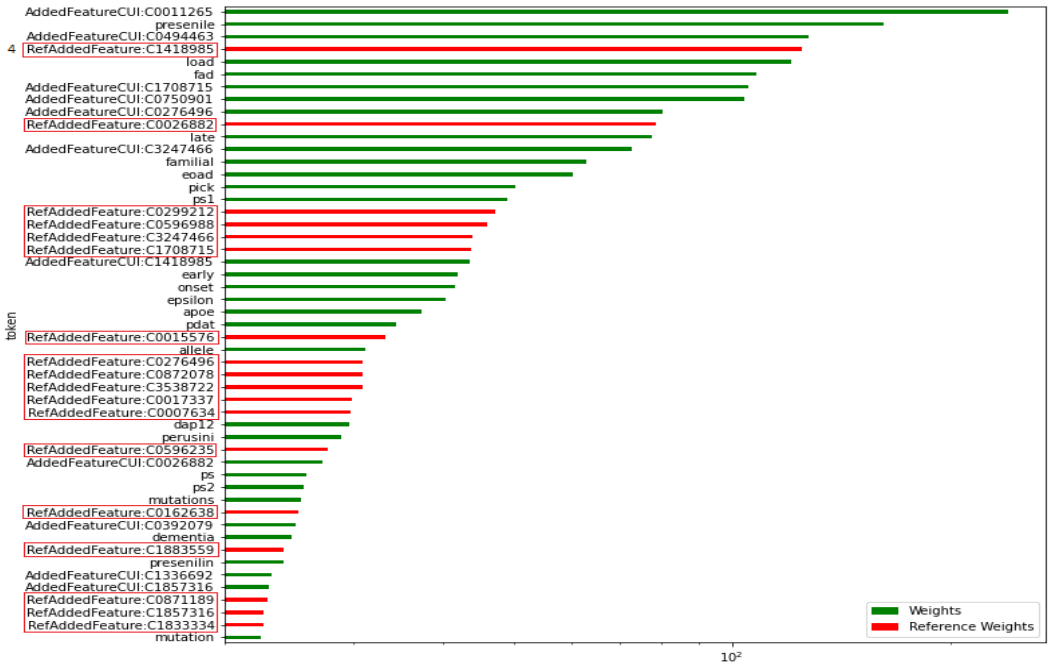


Figure 20: Features' Weights (chi2/frequency) - **References** Case

Observing the first 50 features, a difference in the proportion of the new features chosen arises. References features represent the 36% of the features, whereas citations features represent a lower portion of them, just 28%. This comes in contradiction to the percentages of the chi-square, frequency value case, concerning 1.000 features. References semantic features seem to affect more the chi-square statistic, biasing it towards the new references features.

Examining the top 20 features selected, the changes caused by the addition of the references' semantic features is being obvious. Fig.21 shows the features removed and the ones added to the 20-feature dataset of the baseline method. As observed, the dataset is enriched with 6 new references semantic features, that represent several concepts, like "Presenilin-1" (C0299212) and others, not occurring in the features of the baseline method. Furthermore, for some CUIs, like "Mutation" (C0026882), the feature selection process prefers to choose the reference semantic representation and not the article's itself. Even though important lexical features, like "early" and "onset" have been removed, the model improves its performance by using the new semantic features.

Removed Features	Added Features
AddedFeatureCUI:C0026882	RefAddedFeature:C0026882
AddedFeatureCUI:C1418985	RefAddedFeature:C1418985
apoe	RefAddedFeature:C0299212
early	RefAddedFeature:C0596988
mutations	RefAddedFeature:C1708715
onset	RefAddedFeature:C3247466

Figure 21: Removed & Added Features (k=20, chi-square/frequency values - **References**. The removed highlighted features on the left column represent the same concepts as the highlighted ones on the right column that were added.

Citations & References Semantic Features

Accordingly to the previous modalities, enriching the dataset with both semantic citation and reference features, leads to the addition of 19.264 new features, that correspond to the CUIs assigned by Metamap to both the citation and reference articles. Therefore, the proportion of these new features in a total of 56.988 features is 33.80%. The feature selection method for dimensionality reduction to 1.000 features, rises the sum of proportions of the newly added features in both statistical cases (chi-square & ANOVA-F) with again the ANOVA-F case having a major increase. Again, the choice of binary or frequency values does not seem to play any major role in the aforementioned proportions (Fig.22).

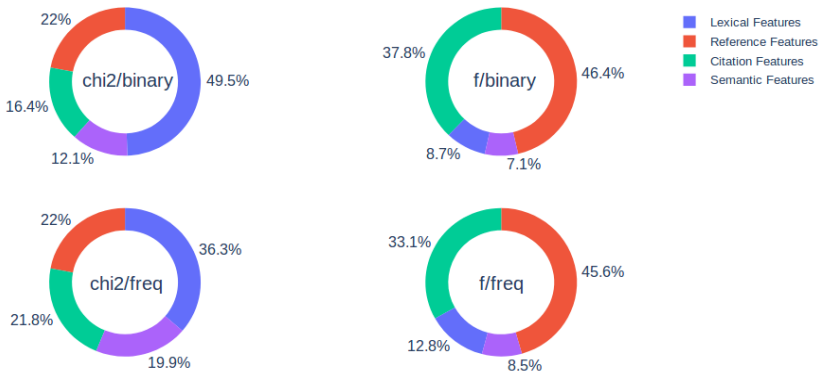


Figure 22: Proportions of selected top features using **Citations & References** - k=1000

Equivalently to the behavior of the feature selection process in the two separate

modalities (citations or references), the same principles seem to apply when using the two modalities combined. So, the positions, that the first new features appear in the dataset follow the same rules as before. In chi-square with binary values case, the new either citations or references features do not seem to be very significant for training, something that does not occur in any other combination of parameters (Fig.23).

	Binary Values		Frequency Values	
	citations	references	citations	references
Chi-square	93	135	7	5
ANOVA-F	17	19	15	14

Figure 23: Position of the first **Citation & Reference** features in the top features

The same semantic features are recognized as important in the case of combined modalities. Thus, the CUIs of the terms "psen1 gene" (C1418985) and "mutations" (C0026882) are high in the weight list, with the corresponding lexical terms being further down or not even in the first 50 features (Fig.24). Again, CUIs corresponding to similar terms are being recognized as significant, like "presenilin-1" (C0299212) and "mutant" (C0596988). Most of the CUIs of citations or references, representing the fine-grained annotations are not in the first 50 features, with the exception of "Familial Alzheimer Disease (FAD)" (C0276496), being in the top 50 features for all three categories of semantic features (initial, citation, reference).

The chi-square statistic with frequency values case shows that out of the first 50 features, 22 are citations or references features, thus 46% of them, divided in 16% for the citations features and 30% for the references features. This is another indication that references features are considered more significant for the predicted outcome than the citations features, by the feature selection method. This is further examined in the next subsection, where the performance of the models are being compared.

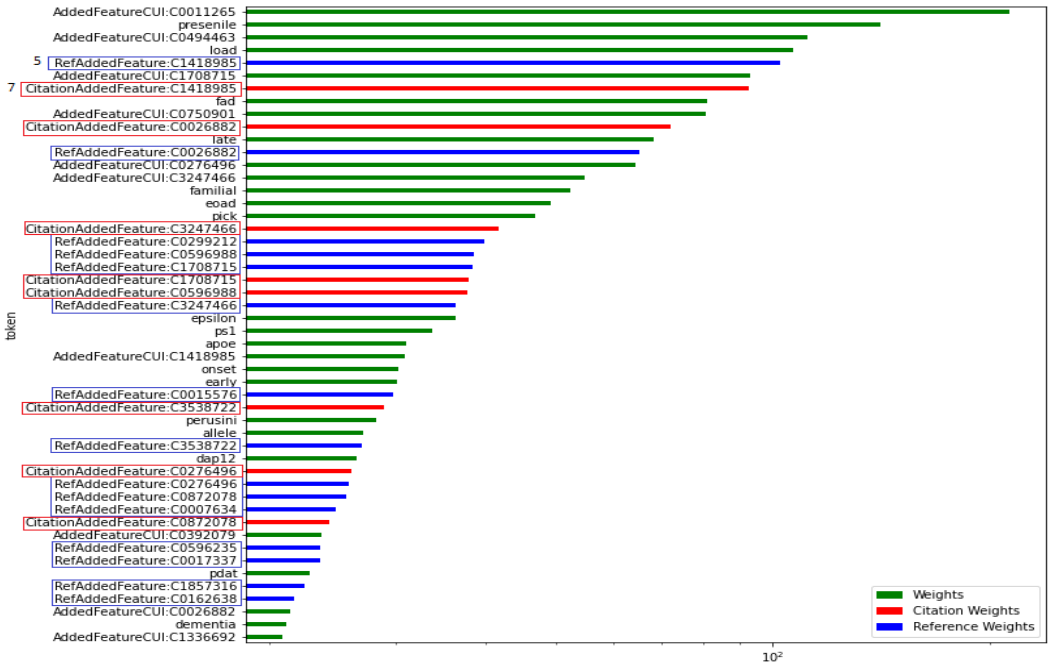


Figure 24: Features' Weights (chi2/frequency) - **Citations & References** Case

In the top 50 features selected, the changes caused by the addition of the new semantic features is being observed. Fig.25 shows the features removed and the ones added to the 50-feature dataset of the baseline method. The dataset in this case is not enriched with any semantic features of the new kind (citations or references) but 5 of them have changed, because the new features affect the assignment of the weights and thus the ranking of the features. Here, the concept "Mutation" (C0026882) is removed along with some lexical features, that do not seem very important, and other more significant lexical features are added, like "apolipoprotein" and "hakola" (Nasu-Hakola disease - early onset dementia disease).

Removed Features	Added Features
AddedFeatureCUI:C0026882	apolipoprotein
gene	dntc
linked	hakola
psen2	nasu
wild	som

Figure 25: Removed & Added Features (k=50, chi-square/binary values - **Citations & References**. The dataset is not enriched with any citations or references semantic features.

5.1.2 Models' Performance

This subsection aims is to compare the models' performance with the corresponding models' performance of the baseline method. The number of trained models, obtained from the conducted experiments, sum up to 336 as described in the *Experiments* section. This total is divided into 112 models for each modality and further into 56 models for each category of values (binary/frequency). The aforementioned models are evaluated on the two tests sets (MA1, MA2) and a total of 672 scores is produced. Another 672 scores are produced by the baseline method models, used for the comparison of the two methods. As a comparison metric, the macro-averaged F1-measure is used.

Citations Feature Models

As a first step to evaluate the results of the models using citations' semantic features, compared to the baseline ones, is to detect the high level changes in the macro-averaged F1-measure scores. If there is not even one model that improves the aforementioned score in the proposed method, further analysis and research on its results would not be useful.

MA1	Category of values	
	binary	frequency
better	10	13
same	38	31
worst	8	12

MA2	Category of values	
	binary	frequency
better	19	17
same	21	16
worst	16	23

Figure 26: Number of models changing the F1-score - **Citations**

Fig.26 shows that in both the results on MA1 & MA2 test datasets, there are more than ten models that improve the macro-averaged F1-scores. Thus, is it worth further researching the results.

Due to the fact that the total number of the scores is large, a mean-based approach is followed to discover if there is a classifier that improves the scores in all or at least in most of the cases. For each test dataset, the scores of the classifiers that have difference between the proposed and the baseline method are kept and

the mean value of the scores of each classifier has been calculated.

For MA1 test dataset and the same experiment setting, only LinearSVC and Logistic Regression classifiers, for both features selection methods, produced scores that differ between the proposed and the baseline method. LinearSVC classifier, trained on the enriched with binary citations semantic features dataset, generates higher score when the feature selection method is performed with chi-square statistic, but lower score when performed with ANOVA-F. The exact opposite occurs, when the values of the new features are the corresponding frequencies. Logistic Regression classifier exhibits exactly the same behavior as LinearSVC (Fig.27).

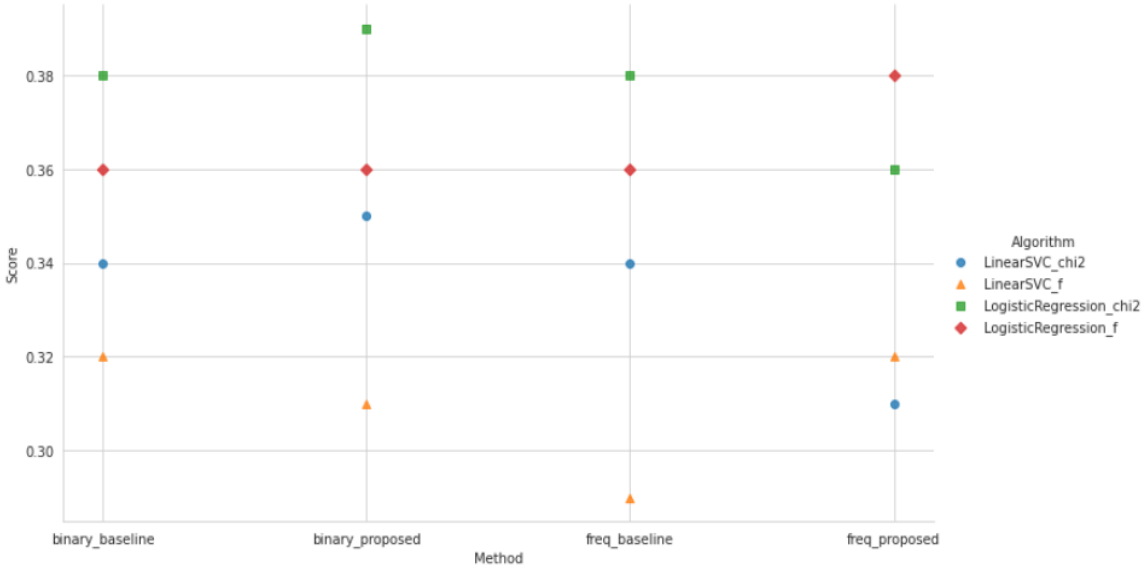


Figure 27: Mean values of classifiers for MA1 - **Citations**

For MA2 test dataset, in addition to LinearSVC and Logistic Regression, also Random Forest classifier, for both features selection methods, produced scores that differ between the proposed and the baseline method. LinearSVC classifier, trained on the enriched with binary citations semantic features dataset, generates higher score no matter what feature selection method is used. When the values of the new features are the corresponding frequencies, lower scores are generated in almost all models. Logistic Regression generates lower score in all cases and Random Forest generates higher score only in the case of binary values and chi-square feature selection method (Fig.28).

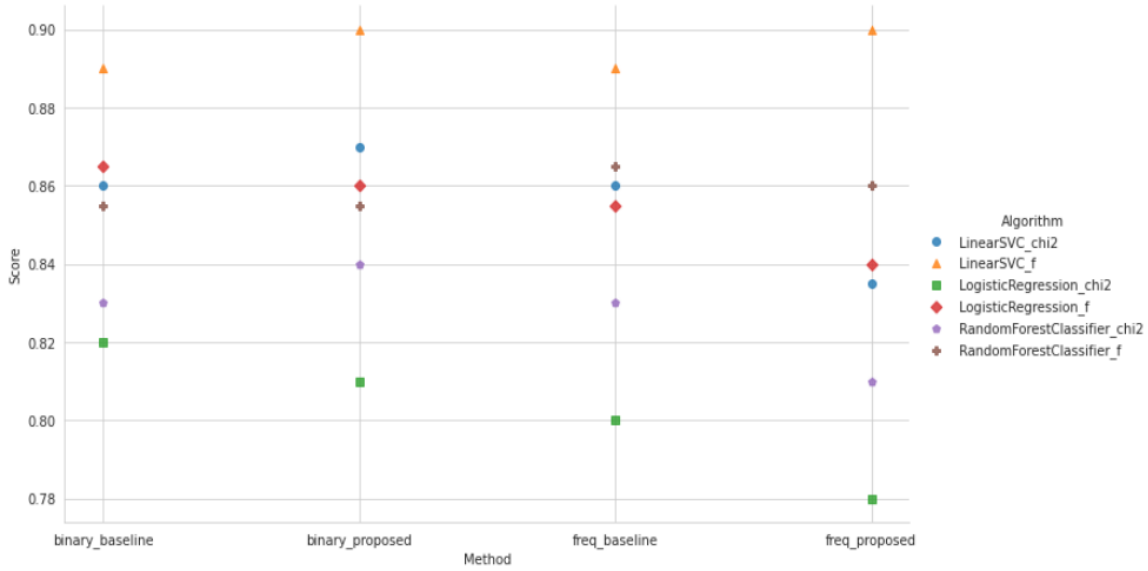


Figure 28: Mean values of classifiers for MA2 - **Citations**

As observed and according to the mean value, the classifier in the proposed method that generate the higher scores, is LinearSVC.

Irrespective of the general conclusion, to which the mean values of classifiers led, a closer look to the scores of the models trained under specific parameters gives a better understanding of whether the new enriched feature set can be useful for the improvement of classification and thus of the fine-grained semantic indexing. In Fig.29 some of the best F1-scores generated from models under specific parameters, both for the citations and the baseline method, are presented. Each sub-figure represents the best resulting scores for each of the test datasets (MA1 & MA2).

For MA1 test dataset, LinearSVC classifier improves the F1-score when the chi-square feature selection method along with binary values, as well as the ANOVA-F feature selection method along with frequency values are used. This observation is consistent with the observation made on the mean values of classifiers. Logistic Regression classifier's behavior, on the other hand, is not consistent with the previous results, as it improves the scores in the combination chi-square/frequency values and ANOVA-F/binary values cases.

The most important observation is that Logistic Regression classifier, with a combination of chi-square/frequency values as parameters and the use of 100 top

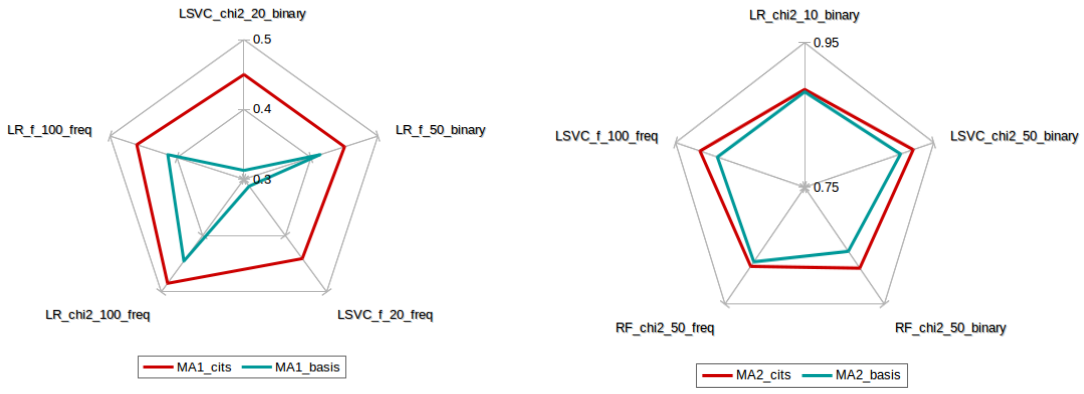


Figure 29: Best F1-scores for **Citations** & baseline methods - MA1 & MA2

features generates the highest F1-score of all the models of the baseline method, as well as all the other models of the enriched method (F1-score of this model = 0.484, next best F1-score = 0.460 from proposed method - LogisticRegression_f_100-frequency values).

For MA2 test dataset, LinearSVC classifier improves the F1-score when the chi-square feature selection method along with binary values, as well as the ANOVA-F feature selection method along with frequency values are used, exactly like the case of MA1 dataset. Logistic Regression classifier, with chi-square/binary values combination, has almost the same performance as its corresponding baseline model. Lastly, Random Forest Classifier improves its performance under the combination of chi-square and both the cases of binary or frequency values.

The most important observation is that LinearSVC classifier, with a combination of chi-square/binary values as parameters and the use of 50 top features generates the highest F1-score of all the models of the baseline method, as well as all the other models of the enriched method (F1-score of this model = 0.918, next best F1-score = 0.915 from proposed method - LinearSVC_chi2_50-frequency values).

References Features Models

The same first level check for changes in the macro-averaged F1-measure scores between the proposed and the baseline method, this time using references' semantic features, is being conducted.

MA1	Category of values	
	binary	frequency
better	7	14
same	38	32
worst	11	10

MA2	Category of values	
	binary	frequency
better	16	10
same	24	16
worst	16	30

Figure 30: Number of models changing the F1-score - **References**

Fig.30 shows that in both the results on MA1 & MA2 test datasets, there is a set of models that actually improve the macro-averaged F1-scores. The same comparison of mean values of the classifiers is performed to gain insight whether there is a pattern, like in the case of the models trained with the extra citations' semantic features.

For MA1 test dataset, again LinearSVC and Logistic Regression classifiers, for both features selection methods, produced different scores. LinearSVC classifier, trained on the enriched with binary citations semantic features dataset, generates lower score, for chi-square feature selection method, but higher score for ANOVA-F method. The same behavior arises, when the values of the new features are the corresponding frequencies. Logistic Regression classifier exhibits almost the same behavior as LinearSVC, with the difference being in the ANOVA-F / frequency values case where it generates lower score, instead of higher. (Fig.31).

For MA2 test dataset, again LinearSVC and Logistic Regression produced scores that differ between the proposed and the baseline method, for both features selection methods. LinearSVC classifier, trained on the enriched with binary citations semantic features dataset, generates higher score no matter what feature selection method is used. When the values of the new features are the corresponding frequencies, chi-square method generates lower scores and ANOVA-F method generates higher score. Logistic Regression generates lower score in all cases (Fig.32).

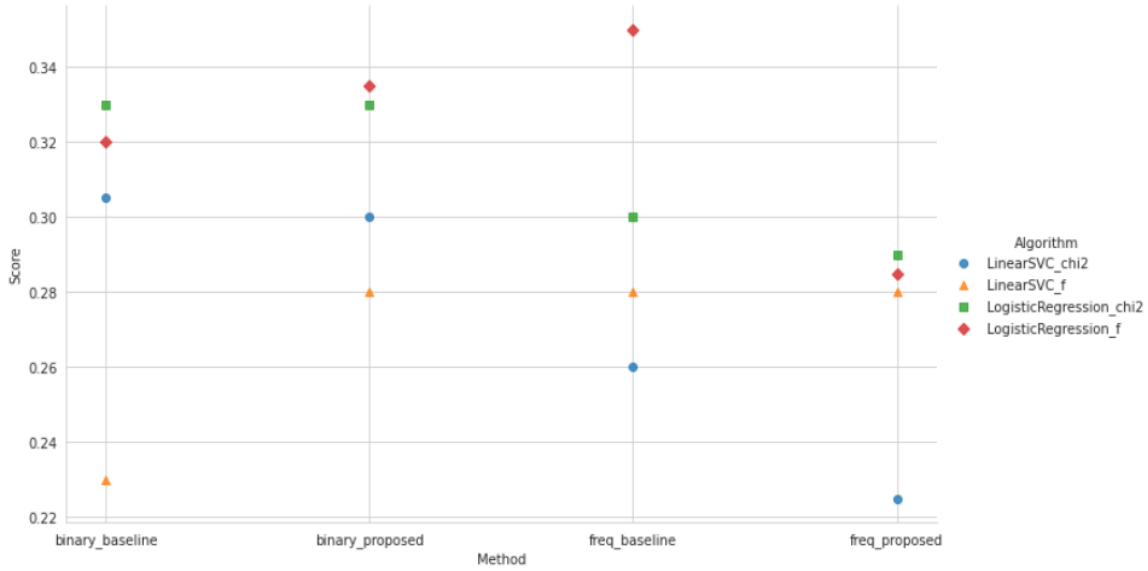


Figure 31: Mean values of classifiers for MA1 - **References**

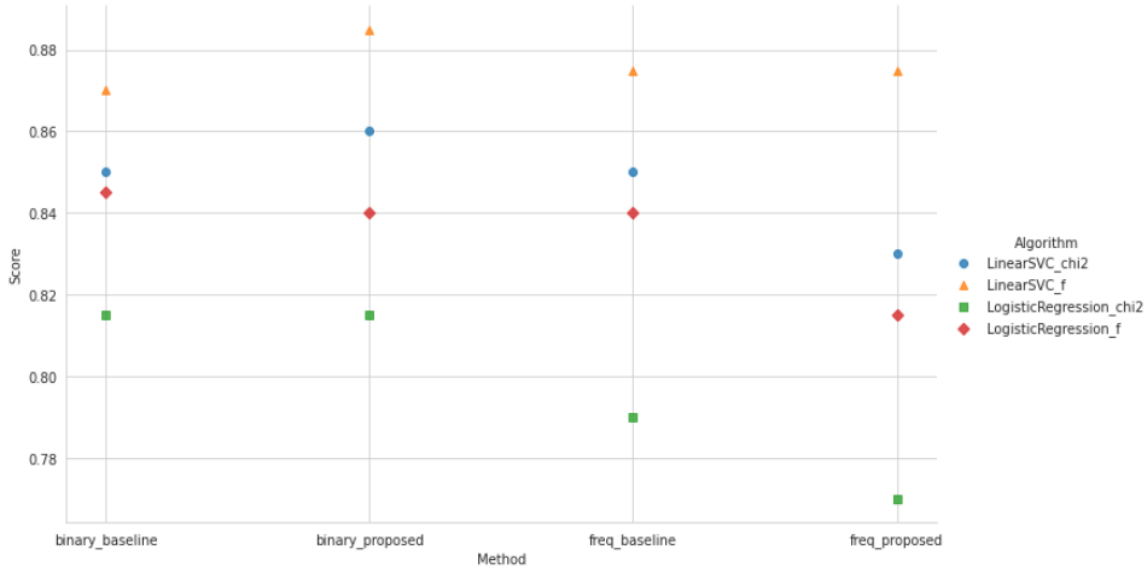


Figure 32: Mean values of classifiers for MA2 - **References**

In the case of the enriched with the references' semantic features dataset, the only pattern observed is that the score gets lower for every classifier, when the combination of chi-square feature selection method and frequencies as values for the semantic features is used.

The same closer look to the models that improve the F1-scores is conducted for the method using the enriched with references' semantic features dataset. In Fig.33 some of the best F1-scores generated from models under specific parameters, both for the references and the baseline method, are presented. Each sub-figure represents the best resulting scores for each of the test datasets (MA1 & MA2).

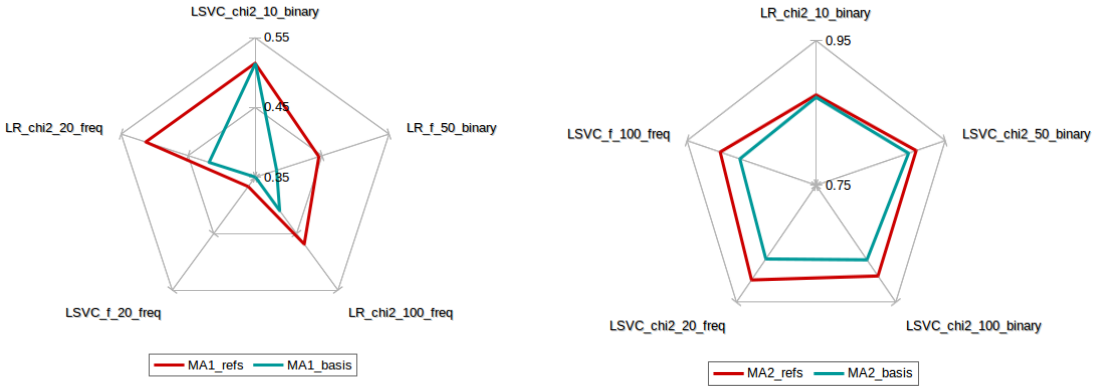


Figure 33: Best F1-scores for **References** & baseline methods - MA1 & MA2

For MA1 test dataset, LinearSVC classifier has the same performance when the chi-square feature selection method along with binary values is used, but improved performance, when the ANOVA-F feature selection method along with frequency values is used. Logistic Regression classifier improves the scores in the combination chi-square/frequency values and ANOVA-F/binary values cases, just like in the citation's method.

The most important observation is that, even though some of the models' performance is improved individually, there is no overall improvement, as the best F1-score of the enriched method is the same as the best one of the baseline method.

For MA2 test dataset, the LinearSVC classifier improves the F1-score in cases of all combinations of parameters, except when the ANOVA-F along with binary values is used. Logistic Regression classifier, with chi-square/binary values combination, has almost the same performance as its corresponding baseline model, just like in the citation's method.

The most important observation is that LinearSVC classifier, with a combination

of chi-square/frequency values as parameters and the use of 20 top features generates the highest F1-score of all the models of the baseline method, as well as all the other models of the enriched method (F1-score of this model = 0.911, next best F1-score = 0.905 from proposed method - LinearSVC_chi2_100-binary values).

Citations & References Features Models

For the third category of experiments, where both the semantic features of the citations and references are used, the first level check for changes in the macro-averaged F1-measure scores is shown in Fig.34. Both the results on MA1 & MA2 test datasets, showed a number of models that improve the macro-averaged F1-scores.

MA1	Category of values	
	binary	frequency
better	5	7
same	43	38
worst	8	11

MA2	Category of values	
	binary	frequency
better	15	14
same	25	15
worst	16	27

Figure 34: Number of models changing the F1-score - **Citations & References**

The observations of the comparison showed that for MA1 test dataset, again LinearSVC and Logistic Regression classifiers produced different scores. LinearSVC classifier generates higher score in cases with every possible combination of parameters. Logistic Regression exhibits the exact opposite behavior, thus generates lower score in all cases (Fig.35).

For MA2 test dataset, similarly to the citations case, LinearSVC, Logistic Regression and Random Forest classifiers produced scores that differ between the proposed and the baseline method. LinearSVC classifier, trained on the enriched with binary citations semantic features dataset, generates higher score no matter what feature selection method is used. When the values of the new features are the corresponding frequencies, LinearSVC produces lower score for both feature selection methods. Logistic Regression generates lower score in all cases and Random Forest generates higher score only in the case of binary values and chi-square feature selection method (Fig.36).

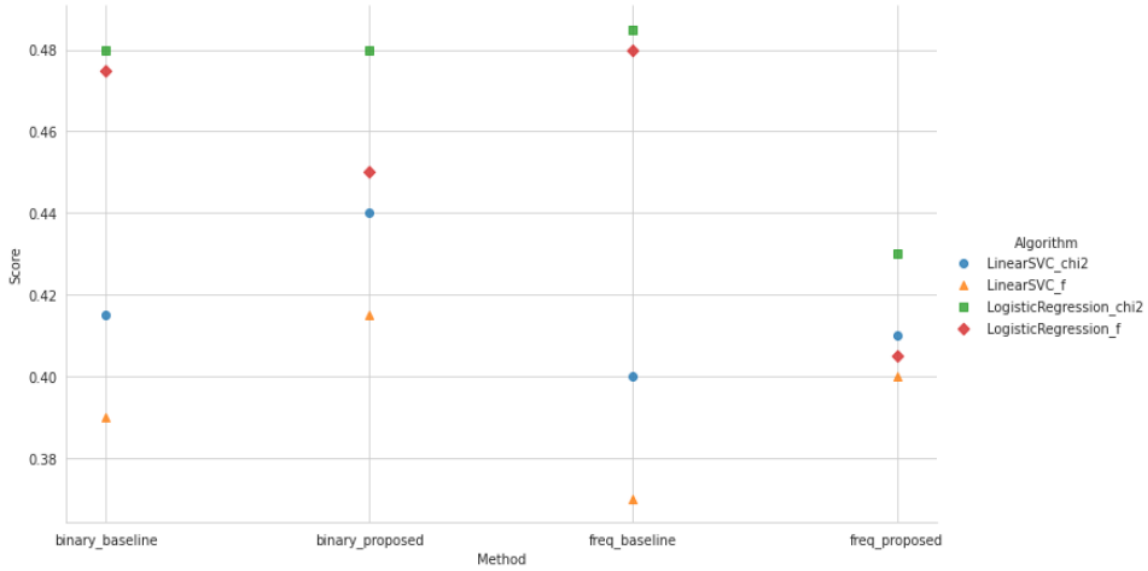


Figure 35: Mean values of classifiers for MA1 - **Citations & References**

As observed, the only classifier that shows consistency in the changes it causes, is Logistic Regression, which generates lower score in every case.

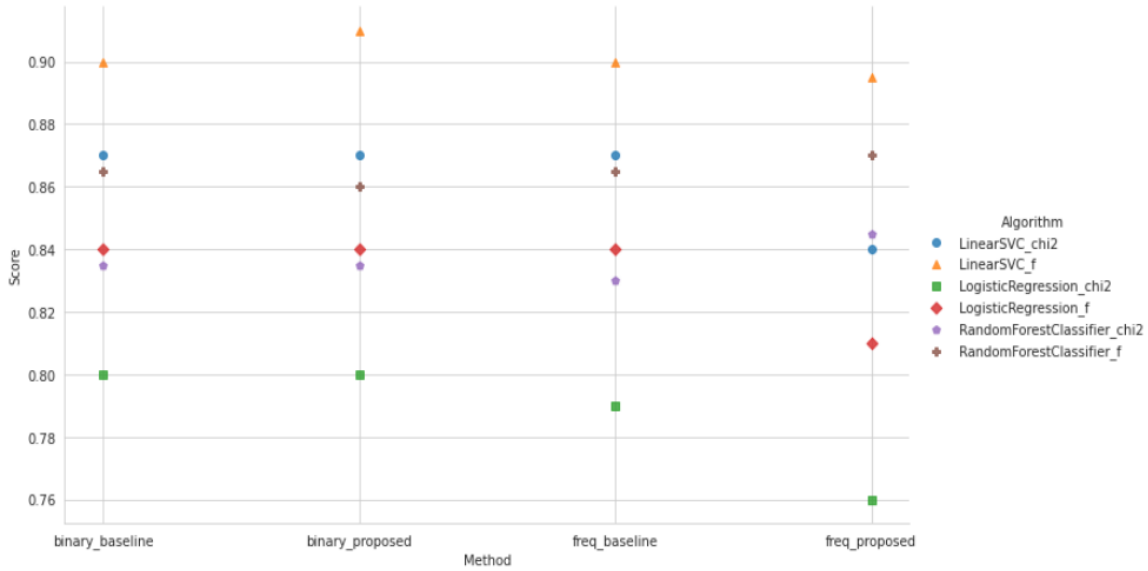


Figure 36: Mean values of classifiers for MA2 - **Citations & References**

Again, the same closer look to the models that improve the F1-scores is conducted for the method using the enriched, with both citations' and references' semantic features, dataset.

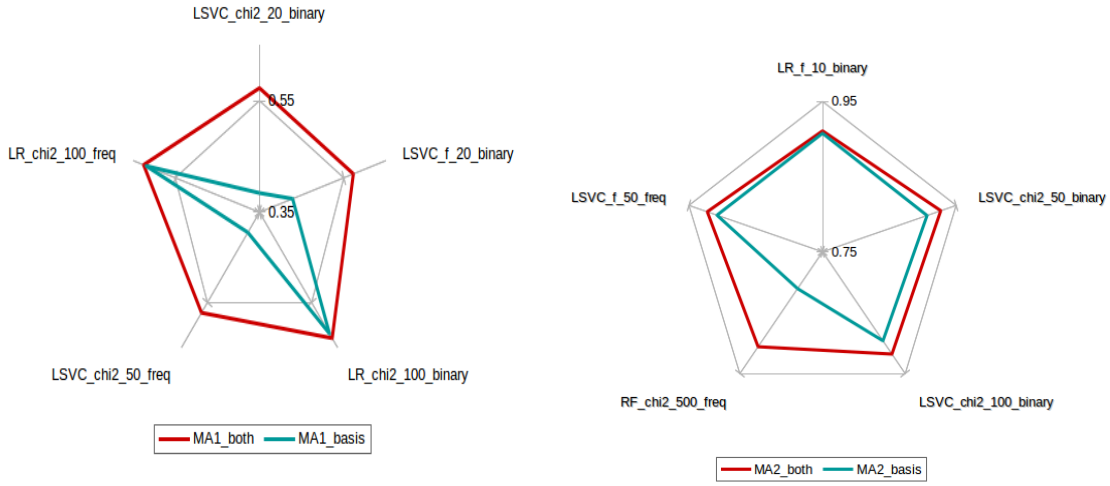


Figure 37: Best F1-scores for both **Citations & References** & baseline methods - MA1 & MA2

In Fig.37 some of the best F1-scores generated from models under specific parameters, both for the enriched and the baseline method, are presented. Each sub-figure represents the best resulting scores for each of the test datasets (MA1 & MA2).

For MA1 test dataset, LinearSVC classifier improves the F1-score for the cases with all combination of parameters, except when the ANOVA-F feature selection method along with frequency values is used. Logistic Regression classifier slightly improves the scores for the cases with the combination chi-square/binary values and chi-square/frequency values.

The most important observation is that Logistic Regression classifier, with a combination of chi-square/binary values as parameters and the use of 100 top features generates the highest F1-score of all the models of the baseline method, as well as all the other models of the enriched method (F1-score of this model = 0.629, next best F1-score = 0.625 from proposed method - LogisticRegression_chi2_100-frequency values).

For MA2 test dataset, LinearSVC classifier improves the F1-score when the chi-square feature selection method along with binary values, as well as the ANOVA-F

feature selection method along with frequency values are used. Logistic Regression classifier, with ANOVA-F/binary values combination, has almost the same performance as its corresponding baseline model. Lastly, Random Forest Classifier improves its performance under the combination of chi-square/frequency values.

The most important observation here is that LinearSVC classifier, with a combination of chi-square/binary values as parameters and the use of 50 top features generates the highest F1-score of all the models of the baseline method, as well as all the other models of the enriched method (F1-score of this model = 0.927, next best F1-score = 0.922 from proposed method - LinearSVC_f_50-frequency values).

5.2 Discussion

The combination of the results, produced by the three categories of experiments, can lead to some interesting conclusions. The proportions of the features in the enriched datasets, after the feature selection step, show that the proportions of the features are almost the same in the cases of citations and references. Combined with the fact that both modalities provide the approximately same amount of features to the dataset, the actual type of modality seem to not affect the feature selection process. In the case where both modalities are used, the amount of the new features is doubled and the proportion is a bit higher than in the other two cases. In any case, with 1.000 top feature selection, the combination of ANOVA-F statistics with binary values seem to produce the results with the largest number of new features.

Concerning the position of the first new feature selected, it is observed that in all three categories of experiments the results are very similar. ANOVA-F always selects the first new feature within the top 20 features, something that occurs only when the chi2-square statistic is combined with frequency values. The fact that both the feature selection methods, choose the new features to be among the top, as well as the proportion of them in the top 1.000 features, indicates that these new semantic features are considered significant by statistics as most contributing to the prediction variable.

According to the analysis of the performance of the models, there are noteworthy

signs that the semantic information of the articles that cites a paper, as well as of the articles that a paper references, can improve its automated fine-grained indexing. There is a significant amount of models that do better when provided this kind of information, with the improvement in the category of citations to be more noteable.

The most prominent performance improvement occurs in the experiments where both the citations' and references' semantic features are used. For both MA1 and MA2 test datasets, this category of experiments may have the lowest amount of models, which improve their scores, but it also produces the highest macro-averaged F1-score of all, as shown in Fig.38 and Fig.39.

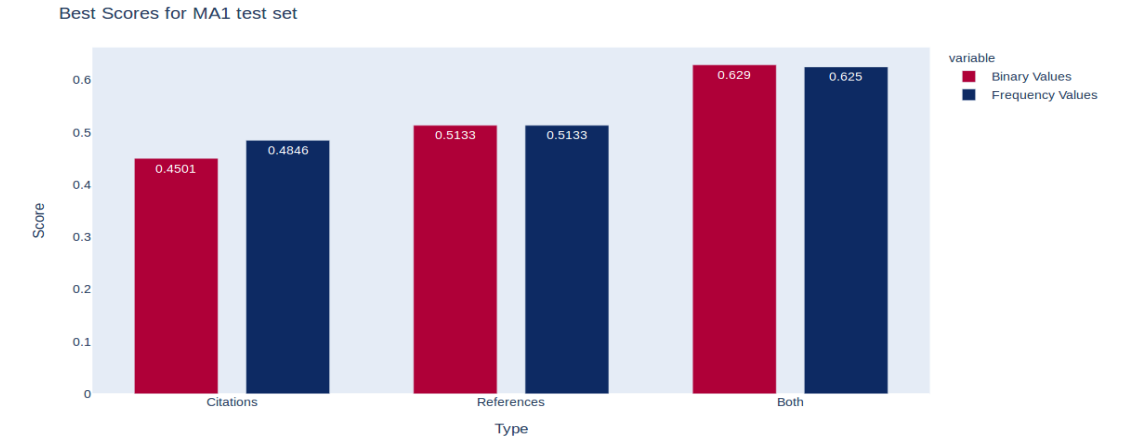


Figure 38: Best F1-scores for MA1 test set

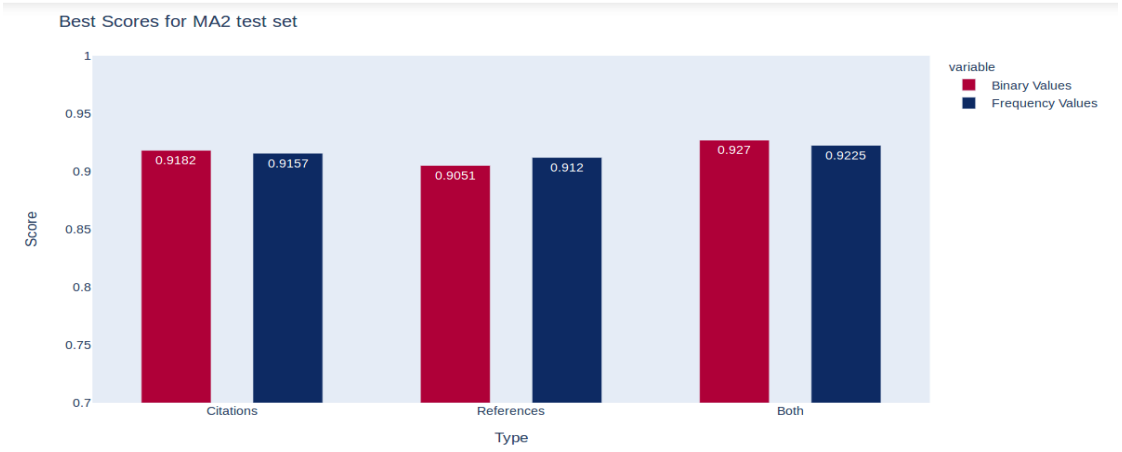


Figure 39: Best F1-scores for MA2 test set

A closer examination to the models that produced the best results for MA1 test dataset, reveals two very interesting patterns. Out of 6 best scores, each for a modality along with a category of value, 5 are produced by Logistic Regression Classifier and also 5 are produced by models that use chi-square statistic for the feature selection process. The best score, compared to all others, is produced by Logistic Regression with chi-square feature selection method, using binary values and the top 100 features (Fig. 40).

MA1					
CITATIONS_binary		*REFS_binary*		*BOTH_binary*	
LogisticRegression_f_50	0.4501633987	LinearSVC_chi2_10	0.5133333333	LogisticRegression_chi2_100	0.6290322581
CITATIONS_frequency		REFS_frequency		BOTH_frequency	
LogisticRegression_chi2_100	0.4846938776	LogisticRegression_chi2_20	0.5133333333	LogisticRegression_chi2_100	0.625

Figure 40: Models of Best F1-scores for MA1 test set

Last, but not least, for MA2 test dataset holds that out of 6 best scores, each for a modality along with a category of value, all 6 are produced by LinearSVC Classifier and also 5 are produced by models that use chi-square statistic for the feature selection process. The best score, compared to all others, is produced by LinearSVC with chi-square feature selection method, using binary values and the top 50 features (Fig. 41).

MA2					
CITATIONS_binary		*REFS_binary*		*BOTH_binary*	
LinearSVC_chi2_50	0.9182341496	LinearSVC_chi2_100	0.9051160834	LinearSVC_chi2_50	0.927007843
CITATIONS_frequency		*REFS_frequency*		*BOTH_frequency*	
LinearSVC_chi2_50	0.9157164009	LinearSVC_chi2_20	0.9119837147	LinearSVC_f_50	0.922541494

Figure 41: Models of Best F1-scores for MA2 test set

6 Future Work & Conclusion

6.1 Future Work

Future investigations are necessary to validate the conclusions that can be drawn from this work. Studies should aim to replicate the results in a larger scale and apply this method to other datasets, such as the WS dataset of Alzheimer Disease or datasets concerning other biomedical subjects, for which there are a lot of publications.

Furthermore, an extension of this method could use the concept occurrence information of citations and/or references as weak labelling instead of features, to investigate whether "extra labels" from this kind of sources can give better classification results. Another idea would be to completely exclude lexical features from the model training procedure and investigate whether training with just semantic features is enough to get the same or better performance.

Additionally, the enrichment of the dataset with extra features could include only the CUIs that represent the fine-grained annotations that we want to predict and not all the CUIs found in an article.

Last but not least, the concept occurrence information, not only from the articles that already exist in the dataset, but from all the citations/references of a publication could be used, to test if the prediction results are improved. Nevertheless, this would require the extraction of MeSH fine-grained concepts, through MetaMap, from the text of all this articles, a procedure rather time and resource consuming.

6.2 Conclusion

In this work we investigated if the inclusion of semantic features from citations and references articles in a dataset for fine-grained semantic indexing, can potentially help classification algorithms.

The new semantic features play a significant role in the classifiers' ability to predict the correct fine-grained labels, something that is backed by the fact that more than one statistical methods select them as important features. There is also strong evidence that models using these new, extra semantic features can outperform the baseline models. Classifiers with specific input parameters are found to be more accurate for balanced and unbalanced test datasets. On this basis, more classifier-targeted research can further improve the results.

Although this research is preliminary on this specific branch of the field and further studies should be conducted for the validation of the initial results, it may be considered a promising aspect of fine-grained biomedical semantic indexing, as the scientific community requires more specialized information, as well as the ability to retrieve it effortlessly.

References

[1] A. Nentidis, A. Krithara, G. Tsoumakas, and G. Paliouras, "Beyond mesh: Fine-grained semantic indexing of biomedical literature based on weak supervision," *Information Processing & Management*, vol. 57, p. 102282, Sep 2020.

[2] A. Dan Corlan, "Medline trend: automated yearly statistics of pubmed results for any query," 2004. Accessed: 2020-08-28.

[3] J. Mork, A. Aronson, and D. Demner-Fushman, "12 years on – is the nlm medical text indexer still useful and relevant?," *Journal of Biomedical Semantics*, vol. 8, 12 2017.

[4] S. Fricke, "Semantic scholar," *Journal of the Medical Library Association : JMLA*, vol. 106, pp. 145 – 147, 2018.

[5] N. Stokes, Y. Li, L. Cavedon, and J. Zobel, "Exploring criteria for successful query expansion in the genomic domain," *Information Retrieval*, vol. 12, pp. 17–50, 2008.

[6] A. Aronson, J. G. Mork, C. W. Gay, S. Humphrey, and W. Rogers, "The nlm indexing initiative's medical text indexer," *Studies in health technology and informatics*, vol. 107 Pt 1, pp. 268–72, 2004.

[7] M. Krauthammer and G. Nenadic, "Term identification in the biomedical literature," *Journal of biomedical informatics*, vol. 37 6, pp. 512–26, 2004.

[8] P. Ruch, "Automatic assignment of biomedical categories: toward a generic approach," *Bioinformatics*, vol. 22 6, pp. 658–64, 2006.

[9] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, "Tuning support vector machines for biomedical named entity recognition," in *ACL Workshop on Natural Language Processing in the Biomedical Domain*, 2002.

[10] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, "Using blast for identifying gene and protein names in journal articles.," *Gene*, vol. 259 1-2, pp. 245–52, 2000.

[11] K. T. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the c-value/nc-value method," *International Journal on Digital Libraries*, vol. 3, pp. 115–130, 2000.

[12] A. Hliaoutakis, K. Zervanou, and E. G. M. Petrakis, "The amtex approach in the medical document indexing and retrieval application," *Data Knowl. Eng.*, vol. 68, pp. 380–392, 2009.

[13] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Toward information extraction: Identifying protein names from biological papers," pp. 707–718, 1998.

[14] A. Névéol, L. Soualmia, M. Douyère, A. Rogozan, B. Thirion, and S. Darmoni, "Using cismef mesh "encapsulated" terminology and a categorization algorithm for health resources," *International journal of medical informatics*, vol. 73, pp. 57–64, 03 2004.

[15] N. Collier, C. Nobata, and J. ichi Tsujii, "Extracting the names of genes and gene products with a hidden markov model," pp. 201–207, 2000.

[16] G. Balikas, A. Krithara, I. Partalas, and G. Paliouras, "Bioasq: A challenge on large-scale biomedical semantic indexing and question answering," in *Revised Selected Papers from the First International Workshop on Multimodal Retrieval in the Medical Domain - Volume 9059*, (Berlin, Heidelberg), p. 26–39, Springer-Verlag, 2015.

[17] A. Kosmopoulos, I. Androutsopoulos, and G. Paliouras, "Biomedical semantic indexing using dense word vectors in bioasq," 2015.

[18] Y. Papanikolaou, G. Tsoumakas, M. Laliotis, N. Markantonatos, and I. Vlahavas, "Large-scale online semantic indexing of biomedical articles via an ensemble of multi-label classification models," *Journal of Biomedical Semantics*, vol. 8, 2017.

[19] A. Rios and R. Kavuluru, "Convolutional neural networks for biomedical text classification: application in indexing biomedical articles," *ACM-BCB: the ACM Conference on Bioinformatics, Computational Biology and Biomedicine.*

ACM Conference on Bioinformatics, Computational Biology and Biomedicine, vol. 2015, pp. 258–267, 2015.

[20] Y. Du, Y. Pan, C. Wang, and J. Ji, "Biomedical semantic indexing by deep neural network with multi-task learning," *BMC Bioinformatics*, vol. 19, 2018.

[21] Z. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, pp. 44–53, 2018.

[22] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. Atkinson, S. Amin, and H. Liu, "A clinical text classification paradigm using weak supervision and deep representation," *BMC Medical Informatics and Decision Making*, vol. 19, 2019.

[23] D. Mekala and J. Shang, "Contextualized weak supervision for text classification," in *ACL*, 2020.

[24] B. Aljaber, D. Martínez, N. Stokes, and J. Bailey, "Improving mesh classification of biomedical articles using citation contexts," *Journal of biomedical informatics*, vol. 44 5, pp. 881–96, 2011.

[25] F. M. O. Guzman, I. Rojas, M. A. Andrade-Navarro, and J.-F. Fontaine, "Using cited references to improve the retrieval of related biomedical documents," *BMC Bioinformatics*, vol. 14, pp. 113 – 113, 2012.

[26] M. Doslu and H. Bingol, "Context sensitive article ranking with citation context analysis," *Scientometrics*, vol. 108, pp. 653–671, 2016.

[27] A. Volanakis and K. Krawczyk, "Sciride finder: a citation-based paradigm in biomedical literature search," *Scientific Reports*, vol. 8, 2018.

[28] A. Janssens, M. Gwinn, J. Brockman, K. Powell, Michael., and Goodman, "Novel citation-based search method for scientific literature : A validation study," 2019.

[29] A. Joorabchi and A. E. Mahdi, "An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata," *Journal of Information Science*, vol. 37, pp. 499 – 514, 2011.

[30] A. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program," *Proceedings. AMIA Symposium*, pp. 17–21, 2001.

[31] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of Documentation*, vol. 60, pp. 503–520, 2004.