ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Natural Language Processing

Argument Mining into Active Learning Systematic Reviews: unlocking the synergy between MARGOT and ASReview

CANDIDATE:

Elisa Ancarani

SUPERVISOR:

Paolo Torroni

CO-SUPERVISORS:

Rens van de Schoot

Marco Lippi

Andrea Galassi

Laura Hofstee

Jelle Teijema

Academic Year: 2022-2023

Session: 2nd

"Sei nel difficile?"

Abstract

Active learning enhances the systematic review process by effectively screening a large amount of titles and abstracts, using machine learning in combination with human expertise. However, the intricacy of full-text (traditional) abstracts can lead to issues, such as token restrictions and longer processing time. In light of these challenges, this thesis harnesses the capabilities of argument mining to distill salient information from abstracts in order to refine the screening process. Therefore, I propose the integration between ASReview LAB, an active learning tool for systematic reviews, and MARGOT, an argumentation mining software. This suggested approach leverages the power of computational argumentation, illustrating its significant value in literature processing. On this basis, I conducted an experiment based on various benchmark data, employing machine learning techniques to extract features from both traditional and Argument Mined abstracts. These features informed subsequent classification models. Next, I test the consistency of the experiment and conduct a quantitative and qualitative analysis spotlighting the benefits of Argument Mined abstracts. Results indicate marginal differences between traditional and Argument Mined abstracts. Yet, in different scenarios, Argument Mined abstracts elevate the overall quality of the systematic reviews. Furthermore, the efficiency of machine learning models depends heavily on the intrinsic attributes of the data they process.

Contents

List of Figures

List of Tables

1. Introduction

A systematic review is a comprehensive synthesis of the literature that uses explicit and reproducible methods to systematically search, evaluate and synthesise a particular topic or research question [1]. Researchers rigorously write systematic reviews to detect and fill gaps in a specific scientific field and to provide an overarching, unbiased summary of all the scholarly research on a given topic [2]. Hence, the process of a systematic review involves manually screening thousands of articles to identify studies that meet specific inclusion criteria while being fully reproducible and transparent [3]. Consequently, systematic reviews occupy a prominent position in research due to their rigorous methodological approach, exhaustive data collection and meticulous evaluation of existing evidence, resulting in an unparalleled ability to offer a comprehensive overview of the available evidence on a specific topic [4]. However, with the exponential growth of scientific papers and policy reports covering diverse subjects [5], crafting a systematic review by hand can be a very labor-intensive and time-consuming task [6]. Advances in technology and the emergence of Artificial Intelligence (AI) techniques have opened up new ways to automate and speed up this process. Since 2016, AI tools like EPPI-Reviewer and Abstrackr have been employed for evidence review [7]. The introduction of ChatGPT by OpenAI in 2022 and subsequent tools like ChatGPT for Sheets and Docs, Casper AI, and ChatPDF have further elevated the role of AI in research, particularly in text-based data analysis [7]. Active learning (AL) is a subset of the Machine Learning (ML) field that demonstrated its efficiency in selecting relevant studies from a large pool of literature as it outperforms human screening using various ML models such as feature extractors and classifiers [8], saving up to 95% of screening time [9]. This approach departs from conventional methods that randomly select which documents to label [10]. Instead, it denotes a scenario in which the ML model iteratively selects which studies should be prioritized for screening [11]. These are then presented to a human oracle, typically a domain expert or annotator, who classifies them as relevant or not and the model updates its predictions based on this feedback [12]. In AL systematic reviews, the ML model is trained just on the most relevant data, becoming more adept at identifying and presenting relevant records to the oracle [13]. This continuous improvement leads to a more accurate and efficient review process, in which the model becomes better at recognising relevant studies at each iteration [14].

This approach turns out to be beneficial as it addresses imbalance issues — given that typically less than 5% of studies are deemed relevant [15] — while reducing the annotation burden [16]. In besides systematic reviews, AL finds applications in different fields and there is different software that implements this technique for different purposes. For instance, Prodigy [17] is a data annotation tool that utilizes AL techniques for tasks like NER, Image Categorization, Object Detection and Text Classification. KNIME [18], is a software in which one of the available features is the labelling of documents with the use of AL. The branch of computer vision that studies the driving of autonomous vehicles has integrated AL for annotating driving scenario trajectories [19]. This dissertation focuses on the application of AL to systematic reviews.

In order to identify relevant literature reviewers generally evaluate the titles and the abstracts [20] of the retrieved studies. This initial assessment helps to determine whether the studies potentially meet the inclusion criteria and warrant further examination in the systematic review. Active machine-learning model (AML) can thus be incorporated into the title-abstract screening process in order to reduce the amount of time spent [21]. In this regard, ASReview LAB [14] is an open-source machine learning tool which integrates AL techniques for screening and systematically labelling a large collection of textual data [22]. The software offers a reliable and transparent title-abstract screening for deciding which studies to include in a systematic review [14]. Therefore, it does an Information Retrieval (IR) of the relevant scientific literature. It supports multiple ML models to extract features from text and classifying them as relevant/non relevant. It comes equipped with exploration and simulation modes, especially beneficial for algorithm comparison and design. ASReview LAB is highly extensible, permitting third-party developers to integrate modules that bolster the pipeline with new models, data, and other enhancements. In addition, with ASReview LAB, it is possible to simulate with the Python API, the Web Interface or the command line interface (CLI). Makita (Make it Automatic) is a workflow generator which exploits the CLI of ASReview to conduct large-scale simulation studies, making it possible to mimic the screening process for a systematic review as if a researcher were using AL [23].

The research of [16] provides a comprehensive understanding of the advantages and limitations of AL in the context of systematic reviews. A challenge that arises when incorporating AL into systematic reviews is to determine the appropriate point to stop the screening process. To tackle this issue, there are several works by [24] and [25]. Furthermore, the ASReview GitHub page hosts open discussions on the matter [26].

Another aspect to consider is the common practice in systematic reviews of relying solely on titles and abstracts due to their concise availability in scholarly databases. Yet, this approach comes with potential limitations such as token restrictions for ML models as well as textual ambiguities.

Argumentation Mining (AM) emerges as a promising remedy. AM's capability to extract argumentative structures from text [27] could enhance the process of feature extraction and help text's representativeness. MARGOT is an AM system developed by [28] which employs state-of-the-art ML and NLP techniques to extract relevant claims and evidence from text, without prior knowledge of the topic. The application of MARGOT spans different domains, from text summarization, to detect trends in newspapers, recommendation and user customisation [28]. The tool is able to support the end-user in highlighting part of text containing the most argumentative content and aims to generalize across multiple domains and genres [28], which is still an open challenge for many current AM methods. In its implementation it is closely followed the definition of [29] where a claim is *"a general concise statement that directly supports or contests the topic"*, whereas a context-dependent evidence is *"a portion of text supporting a claim in the context of a given topic."*. To detect argumentative-rich information the tool was trained on the largest available corpus dataset of argumentative text from IBM Research. The corpus contains 547 Wikipedia articles organized into 58 topics, and it has been annotated with 2294 claims and 4690 evidence facts, both context-dependent, meaning that they are labelled only if they are relevant to a given topic [28].

The system builds upon and extends the previous work on detecting context-independent claims [30] using Tree Kernels [31].

MARGOT implements an AM-pipeline that can be divided into two steps:

1. **build two distinct classifiers** - given an input text MARGOT labels the sentences of the textual corpora as claims and/or evidence.

2. **detect the boundaries of the argument components.**

The ideal structure to encode the grammatical structure of a sentence was thought to be a constituency parse tree, driven by the observation that argumentative sentences often have common syntactic structure [28]. Then, MARGOT employs tree kernels to capture similarities between pairs of trees by considering the number of common substructures or fragments they share. The higher the similarity of the common substructure between pairs of trees, the stronger the indication of similarity between the arguments represented by those trees. The second step is formulated as a sequence labeling problem. In MARGOT is considered a technique which combines Structured Support Vector Machines with Hidden Markov Models (SVM-HMM), where SVM-HMM is trained to label a sentence by identifying both claim and non-claim tokens. At the end of this two subsequent stages, the system returns a claim and evidence score which represent the confidence that a sentence is a claim and/or evidence.

AM and IR can complement each other in various domains. For instance, AMICA [32] is an AM-based search engine designed for the analysis of literature related to Covid-19, it retrieves scientific papers from matching keywords and ranks the results according to their argumentative content. This dissertation focuses on building a bridge between AM and IR in the context of systematic reviews. To this end, my goal is to seamlessly integrate the Argument Mining (AM) capabilities of MARGOT with the Information Retrieval (IR) functions of ASReview LAB. In particular, I want to analyse the impact of the use of Argument Mined abstracts (AM abstracts) in the systematic review process when using Active Learning (AL) techniques to retrieve relevant documents during scientific literature screening.

The key contribution of this dissertation lies in the in-depth analysis of the various aspects and potential benefits of the use of AM abstracts in systematic reviews with AL techniques and the possibility that they could replace full-text abstracts. Specifically, I want to identify both the benefits and challenges associated with the harmonisation of these tools in order to explore potential solutions to the problems or limitations encountered. Finally, I validate the effectiveness of the integrated approach through experimental validation.

The thesis opens with the ”Methods” section, which explains the data, the experiment setup and the analytical strategy in which are detailed all stages of the analysis. The subsequent ”Results” section shows the outcomes of the simulations of the various datasets, considering also the effect of the absence of both the abstracts and titles. This aspect is further expanded upon with a result overview, a test on the stability of the results performed by carrying out several simulations, and a comparative analysis in the ”Results Across the Simulations” section. The study wraps up with a discussion and a conclusion section.

Thesis Objective and Contributions

The objective of the present thesis is to find a synergistic integration between MARGOT and ASReview LAB. The main contribution of this dissertation is an analysis of the nuances and potential benefits of AM abstracts in systematic reviews that use active learning techniques. Thus, starting with some labelled benchmark data containing screened studies of different topics, the *research question* this work tries to answer is how AM abstracts impact in the process of systematic review and if they could possibly replace full-text abstracts.

Addressing this challenge presents both opportunities and potential pitfalls. Specific concerns include handling of textual data which pertains to different topic, the variability in terminology of scientific literature as well as noise and ambiguities. Additionally, determining appropriate evaluation metrics, designing experiments, and establishing benchmarks to answer the research question can be challenging, particularly when evaluating the combined performance of the two tools. The thesis tries to achieve the following objectives:

- Examine and understand the differences in text representation in terms of argumentative structure, coherence, quality as well as noise and ambiguities by carrying out a comparative analysis between traditional (full-text) abstracts and AM abstracts.
- Analyse different aspects of the data such as unbalance issues, completeness (presence of both abstract and title) and investigate also on the missingness of the AM abstracts and full-text abstracts, exploring the impact of their removal on the performance of the systematic reviews with ASReview LAB.
- Design an experiment by considering the appropriate benchmark data, machine learning models and evaluation metrics in order to accurately assess the performance and the outcomes resulting from integration of the tools.
- The stability, scalability and efficiency of the experiment are also considered.

Figure 1.1 shows a very general picture of the steps carried out. More in-depth details are provided in the following chapters.

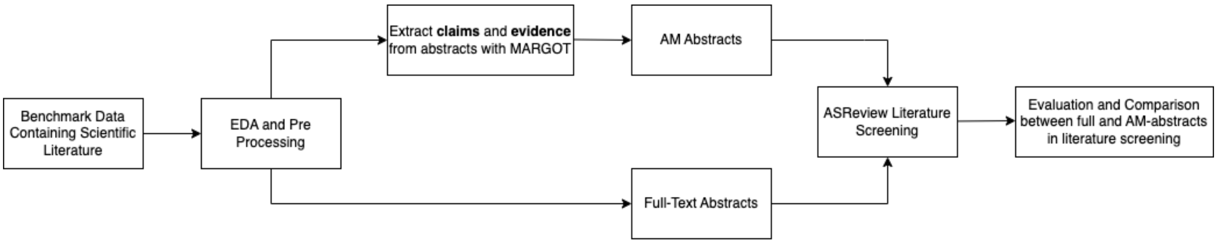

Benchmark Data Containing Scientific Literature
EDA and Pre Processing
Extract **claims** and **evidence** from abstracts with MARGOT
AM Abstracts
Full-Text Abstracts
ASReview Literature Screening
Evaluation and Comparison between full and AM-abstracts in literature screening

Figure 1.1: General pipeline of the phases of the study

The study seeks to explore the benefits of integrating MARGOT with ASReview LAB by providing the output characteristics of MARGOT (the claims and evidence extracted in the abstracts) as input to ASReview LAB, which conducts a systematic review by retrieving the most relevant documents. Thus, this dissertation holds importance due to several reasons:

- Systematic reviews play a key role in synthesising a large amount of literature [14], by integrating Argument Mining into ASReview LAB this research aims to analyse and try to enhance the screening process into different scenarios.
- Find relevant arguments acknowledges challenges related to the information loss and content missingness. Addressing these issue will contribute in a more in-depth understanding of the text representation and data quality.
- With advancements in technology and the emergence of AI techniques, there is a growing opportunity to automate and expedite the systematic review process. Thus, the aim is to explore the harmonisation of two distinct AI domains: how text processed by an AM tool influences the screening processes of an IR tool based on active learning and if AM abstracts could potentially replace the full ones.
- This project aims to generalise across different topics and domains, including healthcare and statistics. Hence, the answer to the research question is assessed from different perspectives.

Different scenarios when evaluating the integration of AM abstracts into the process of systematic review are taken into account with the aim of trying to contribute to improve the data quality and decision-making in academic research. Moreover, AM information may be valuable for researchers that manage large volumes of data in order to reduce the reliance on full-text sources and to aid in effective information management.

2. Methods

2.1 Experimental Set-Up

In order to try to further speed up the process of systematic review when using AL, this experiment employed a comparative approach to determine the efficacy of full-text (traditional) abstracts and MARGOT AM abstracts. I relied on various benchmark datasets, each containing studies labeled as either relevant or non-relevant. The choice of the data was mainly influenced by the performance described in [33]. This, allowed to assess the impact of MARGOT on data with varying quality standards and to identify potential room for improvement for systematic reviews when using the tool. Although these datasets might differ in some attributes, each includes a *title* and an *abstract* field suitable for input to ASReview LAB. The core objective was to evaluate how systematic reviews perform when utilizing full-text abstracts versus AM abstracts, both within individual datasets and across all datasets under consideration.

2.2 Data

The choice of data was made by considering both their difference in performance in terms of the WSS@95 score obtained in a previous simulation carried out by [33] as well as their difference in topics. The simulation was conducted with an ensemble classifier using NB and LR and proposed two ensembling strategies: Multiply Ensemble and Random Ensemble. The former multiplies the probabilities of each model, giving more weight to records that both models agreed upon, while the Random Ensemble strategy randomly chooses a classifier at each iteration and the probabilities are calculated using the selected classifier. The data were chosen based on the results of the Multiply Ensemble strategy.

The data in the present study belong to both the statistical and the medical field. They are all available at SYNERGY, a free and open dataset on study selection in systematic reviews [34]. The data exhibited both high and low performance in the above mentioned simulation.

The threshold value of the WSS@95 is 0.5. This choice was due to the fact that a WSS@95 score below 0.5 means that less than 50% of the records can be safely skipped. This indicates that a significant amount of manual screening effort is still required to achieve a high recall level. The first two datasets under analysis are *PTSD-Trajectories* from [35] and *ACEInhibitors* from [36]. The former contains screened studies on Bayesian PTSD-Trajectory Analysis with Informed Priors whereas the latter is a gold standard dataset utilised for systematic review of drug classes. I selected these datasets based on their good performance in terms of WSS@95, quantified as `0.889` and `0.789` for the first and second datasets, respectively.

Conversely, the remaining datasets were selected for their low performance. These are: *Antihistamines*, *Estrogens*, *ADHD*, *Urinary Incontinence*, *Opioids*. All are from [36], like *ACE*, and are used in research on automated classification of document citations for systematic review of drug classes. The former exhibited an almost negligible WSS@95 score of `0.055` but the others demonstrated a score ranging from `0.20` to `0.49`. Table 2.1 shows the data with their respective WSS@95 scores.

Table 2.1: WSS@95 scores of datasets from the simulation of [33]

Dataset	**WSS@95**
PTSD	0.889
ACE	0.789
Antihistamines	0.055
Estrogens	0.257
ADHD	0.411
Urinary Incontinence	0.455
Opioids	0.482

The characteristics of the data can be summarised in Tables 2.2 and 2.3. The tables include an overview of the size, domain and percentage of relevant studies, as well as the percentage of missing abstracts, titles and abstracts in the records. None of the datasets contain duplicate values.

In this work, PTSD is the first dataset under study. It comprises 5782 records and 39 attributes detailing information of a scientific article such as title, abstract, authors, publisher, type of reference, year, and so forth. It contains 38 relevant studies and 747 empty abstract fields.

The second dataset is ACE. It consists of 2544 records and 8 attributes, such as publication ID,

Data	Field	Size	Relevant Records %
PTSD	Statistics	5782	0.66%
ACE	Medicine	2544	1.61%
Antihistamines	Medicine	310	5.16%
Estrogens	Medicine	368	21.73%
ADHD	Medicine	851	2.35%
Urinary Incontinence	Medicine	327	12.23%
Opioids	Medicine	1915	0.78%

Table 2.2: Characteristics of datasets: domain, size, and relevance percentage

Data	Missing Abs.%	Missing Titles%	Missing Abs. & Titles %
PTSD	12.92%	1.11%	1.07%
ACE	12.15%	0%	0%
Antihistamines	10.65%	3.23%	3.23%
Estrogens	8.42%	3.80%	3.80%
ADHD	8.23%	2.82%	2.82%
Urinary Incontinence	17.13%	5.50%	5.50%
Opioids	10.34%	3.13%	3.13%

Table 2.3: Characteristics of datasets: percentage of missing abstracts, percentage of missing titles, percentage of missing abstracts and titles

title, keywords, authors, abstract, and label type descriptions. Within this dataset, 41 records represent relevant studies, approximately 1.61% of the total. Additionally, there are 309 empty abstract fields.

The other datasets are those with a comparatively lower level of performance. For all these data, the number of attributes is the same as that of ACE.

Antihistamines is the starting point due to its notably inferior WSS@95 score in comparison to the other data. This choice was made to facilitate an immediate comparative evaluation of the different performance of the models using data with different performance outcomes. Out of the 310 rows of Antihistamines, 16 are relevant studies. Furthermore, 33 records have an empty abstract field, which represents 10.65% of the entire dataset. The next dataset is Estrogens and comprises 368 rows, of which 80 relevant studies and 31 records without an abstract field. Urinary Incontinence presents a significantly higher proportion of missing abstracts. In fact, 56 out of 327 abstracts are empty in the original version, or 17.13%. The last dataset is Opioids which has 198 missing abstracts out of a total of 1915 records.

2.3 MARGOT Set Up

In this phase, MARGOT was configured to identify argumentatively rich information from the abstracts. Given that MARGOT requires text-based inputs, I employed a Python script to extract and export each abstract into individual `.txt` files, with each dataset having its dedicated folder. To keep the indexing of the records, missing abstracts were also included. Thus, a text file might be empty. Next, I fed these files into MARGOT using a Bash script provided by [28]. The tool then broke down each abstract into sentences, assigning them claim and/or evidence scores. A sentence is deemed relevant if its corresponding score is greater than zero. Afterward, I could export the results in various formats such as HTML, JSON, TXT, and XML. For convenience, I opted for the JSON format. After processing, I created for each dataset two separate folders: one with the full-text abstracts and another, named AM abstracts, containing the abstracts processed by MARGOT. The total amount of files produced by MARGOT is 12097, matching the combined number of records from all the datasets. Through another Python script, I reconstructed the text from the JSON files, retaining only the parts with a significant score. As the final step, I created a dataset which reflects the original one by mapping each AM abstract in place of the full one. In this way, I set the stage for a comparative analysis between full-text and AM abstracts.

The abstract of this dissertation provides an example on how MARGOT finds relevant claims and evidences in a text corpus:

Active learning enhances the systematic review process by effectively screening a large amount of titles and abstracts, using machine learning in combination with human expertise. *This suggested approach leverages the power of computational argumentation, illustrating its significant value in literature processing. Next, I test the consistency of the experiment and conduct a quantitative and qualitative analysis spotlighting the benefits of Argument Mined abstracts.*

Bold text highlights claims, while italics text represent the evidence. I reported only the claims and evidence with a score greater than zero.

2.3.1 Handling Abstracts Fields in Data

In the datasets under analysis it may be the case some records have an empty abstract field. Therefore, an abstract field left unpopulated in the original dataset, is kept as such also in the dataset processed by MARGOT. In the circumstance where MARGOT does not find any claims or evidence in a given abstract, it is labelled as empty in the MARGOT dataset and regarded as a ”missing abstract” despite its presence in the original dataset. This choice was made with the aim of enabling a comprehensive comparative analysis between the two tools. Thus, it is assumed that the pertinent information is encapsulated within the title alone, given the fact that ASReview requires a title and abstract as input for carrying out a systematic review. If neither is present, the software is still able to learn about the irrelevance of a record.

2.4 Experiment Design and Evaluation Metrics

Before moving on to the second part of the experiment, this section provides a theoretical understanding of the machine learning algorithms implemented in ASReview LAB and employed in this project. It also gives an overview of the measures used to evaluate their performance.

2.4.1 Machine Learning Models

The experiment involves multiple machine learning models to perform classification and feature extraction on text data. The feature extractor identifies relevant features from the text, which are subsequently given as input to the classifier which binary labels the study as relevant/non relevant. Here’s is the list of the models that I will use in this work. Further details about their combinations are given in the Makita’s setup section. The classifiers considered in this study are the following:

- **Logistic Regression (LR)** - is a model which estimates the probability of an event occurring [37], in our case determining the likelihood of a record being relevant. The main idea behind it is to pass a weighted sum of inputs through a sigmoid function (activation function) which maps the output to a range of probabilities between 0 and 1. Although logistic regression is commonly used for binary classification problems, there are also variants for multi-class classification.

- **Naive Bayes (NB)** - is a probabilistic algorithm rooted in the Bayes Theorem [38]. The theorem relates the conditional and marginal probabilities of two random events, and can be viewed as a way to update probabilities with new evidence. The term ”naive” in Naive Bayes arises from the assumption that all predictor variables are independent of each other given the class variable [39]. While this independence assumption often doesn’t hold in real-world scenarios, the algorithm simplifies computational complexity by treating each predictor variable independently, making it especially efficient with large datasets [40].
- **Random Forest (RF)** - is an ensamble of decision trees [41] where a decision tree is a tree structure where each internal node represents a feature, each branch represents a decision taken based on that feature, and each leaf node represents the predicted outcome [42]. In ASReview, the default number of trees is set to 100.
- **Support Vector Machine (SVM)** - The main idea is to transform the input data into a higher-dimensional feature space and find a hyperplane that maximally separates the data points of different classes while maximizing the margin, which is the distance between the hyperplane and the nearest data points of each class [43].

All this algorithms are supervised meaning that algorithms are trained on labelled dataset to make predictions on unseen data [43].
The feature extractors are:

- **Doc2Vec** - also known as paragraph vectors - enables to learn the representation of documents as vectors (document embeddings) by mapping each document as a fixed-length vector in a high dimensional space [44]. In ASReview the vector length is set to 40.
- **TFIDF** - Term Frequency - Inverse Document Frequency is a statistical measure that evaluates the relevance of the representation of a string in a document within a set of documents. It is obtained by multiplying the number of times a word appears in a specific document (Term Frequency) by its rarity in a set of documents (Inverse Document Frequency) [45]. The method penalises words that appear frequently in documents and assigns greater importance to words that appear less commonly.
- **SBERT** - Is a variation of the classical BERT model and it relies on a Siamese architecture, which consists of two identical neural network branches with shared weights to

better capture sentence-level semantics [46]. This architecture allows to process pairs of sentences and learn to encode their semantic similarity [47]. BERT architectures have a token processing limit of 512 [48]. Unlike traditional BERT, which primarily focuses on word-level representations, SBERT aims to generate meaningful sentence embeddings by considering the contextual information of each word within the broader context of the sentence.

More details on the hyperparameters that all these models (classifiers and feature extractors) have by default are available in the official ASReview documentation [22].

2.4.2 Evaluation Metrics

The metrics used to evaluate the performance of the ML models used to carry out a systematic review on both the titles and the AM abstracts as well as the titles with the full abstracts are Recall, WSS and ATD. The Recall is the main metric considered. The Recall measures the proportion of relevant records (true positives) that have been found after screening $X\%$ of the total records [49]. It can be measured as:

$$Recall = \frac{relevant\_instances\_retrieved}{all\_relevant\_instances} \qquad \text{from [50]}$$

In this work, the Recall is measured at 10% of screened records (Recall@10).
According to [36], the WSS is a measure of *”the work saved over and above the work saved by simple sampling for a given level of recall”*. Specifically, it calculates the percentage of records that a reviewer can skip after receiving prior knowledge used to train the initial model iteration compared to random reading. The WSS is typically measured at a Recall of 0.95 (WSS@95) and it a commonly used benchmark reflecting the proportion of records saved through active learning while still missing 5% of relevant publications [49].

$$WSS@95 = \frac{TP+FN}{N} - 0.05 \qquad \text{from [36]}$$

Figure 2.1 illustrates the difference between the metrics Recall (y-axis) and WSS@95 (green line). The diagonal on the stepped line represents the naive labeling method, where records are screened in random order. Each step represent a relevant record, excluding the ones taken as prior knowledge.

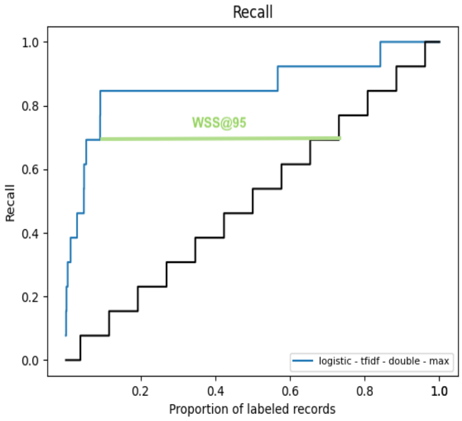


Figure 2.1: Explanation of Recall and WSS@95

ATD (Average Time to Discovery) [51] is chosen in order to investigate the difference in speed in discovering a relevant record between AM and the full-abstracts, with the measurement unit being seconds. ATD can be formulated as follows:

$$TD_i = rank_i$$

$$ATD = \sum_{i=1}^{N} \frac{TD_i}{N}$$

Where N is the total number of relevant paper, TD_i is the time to discover a relevant record i which is equal to the $rank_i$. As all record are discovered in a certain order, $rank_i$ stands for the ranking of a relevant record i among all record discovered. The ranking excludes the prior knowledge that is used to initially train the algorithms.

2.5 Makita Set Up

With Makita it is possible to mimick the screening process for a systematic review as if a researcher were using active learning [23]. Makita offers various templates tailored to different simulation needs. In this work, I carried out 14 simulations. For each of the seven datasets, I considered both the original and AM versions. I employed the 'multiple models' template, which prepares a script for running simulations with varied algorithms and a fixed set of priors. This includes one relevant and one irrelevant record that remain unchanged across all models. The default models are:

- Classifiers: Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machines (SVM)
- Feature extractors: Doc2Vec, SBERT, TF-IDF
- Impossible models: [Naive Bayes, Doc2Vec], [Naive Bayes, SBERT]

A simulation incorporates combinations of the above mentioned classifiers and feature extractors. However, certain pairings, such as the Naive Bayes classifier with Doc2Vec or Sentence BERT (SBERT) feature extractors, are incompatible. This is primarily because Naive Bayes takes as input only positive numbers, which TF-IDF provides, whereas both SBERT and Doc2Vec may produce negative values.

2.6 Experiment Reproducibility

The data can be found at SYNERGY [34]. The specific versions employed in this project include ASReview LAB version 1.1.1 and Makita version 0.6.3. To perform batch processing in a large number of documents, e.g all the abstracts, MARGOT is available as a stand-alone package, provided by [28]. The experiment is initially conducted once (number of runs = 1) to compare the outcomes of each dataset with its respective version processed by MARGOT (AM version). It is executed on a macOS equipped with a quad-core 2.3 GHz Intel Core i7 processor and 16 GB RAM. The command to set up the Makita's multiple model template that has to be launched from the terminal inside the Makita folder is:

```
asreview makita template multiple_models
```

More information on how to set up a simulation study using Makita can be found at [23].
Then, the stability of the results obtained is verified using a CPU optimized Computer cluster by running 15 simulations for 32 GB of memory. The platform that hosts these simulations is called Exoscale, an European cloud service provider. To speed up the process, the simulation processes are run in parallel, divided over the 16 available computing cores. For these simulations, it is used a custom Makita's multiple model template in which the parameter 'n_runs 15' is added to the command for the parallel processing line. Importantly, in the multiple model template all the ML models have a default random seed, instead the seed of the prior knowledge (one relevant and one irrelevant record) increases with each run, starting at 535.

2.7 Data Pre Processing

This section delves into the characteristics of the data after being processed by MARGOT. In particular, it focuses on the percentage of missing abstracts, abstracts length as well as MARGOT's ability to find relevant claims and evidence from those abstracts and its capacity to process relevant studies. Overall, this analysis sheds light on the implications of MARGOT for data completeness.

2.7.1 Margot Data

Table 2.4 provides the percentage of missing abstracts in the datasets before and after being processed by the AM tool.

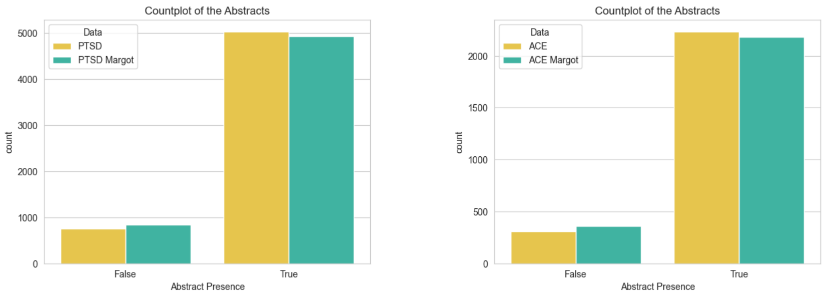
Data	Missing Abs.%	Missing AM Abs.
PTSD	12.92%	14.60%
ACE	12.15%	14.19%
Antihistamines	10.65%	12.26%
Estrogens	8.42%	8.97%
ADHD	8.23%	9.99%
Urinary Incontinence	17.13%	18.35%
Opioids	10.34%	12.17%

Table 2.4: Missing abstracts percentage of the data before and after being processed by MARGOT

In PTSD, MARGOT doesn't detect 97 abstracts resulting in a total of 844 empty abstracts fields. The ACE dataset contains 309 empty abstract fields. MARGOT doesn't detect 52 abstracts, resulting in a total of 361 missing abstracts in its AM version. Figure 2.2 shows the frequency of missing and non-missing abstracts for both PTSD and ACE.

Shifting the attention to the data characterized by a comparatively lower score, Antihistamines comprises 33 missing abstracts and MARGOT doesn't identify any claim and/or evidence in 5 abstracts, bringing the total number of missing abstracts fields to 38 in its AM version. In Estrogens the AM version of MARGOT results in a total number of 33 missing abstracts.

The AM version of ADHD results in a total number of 85 missing abstracts, so about 10% of the entire set of data. Urinary Incontinence is the fourth dataset under analysis and its AM version has total of 60 missing abstracts. The last dataset considered to assess the influence of AM abstracts in the process of systematic review is Opioids. It has 198 missing abstracts in the

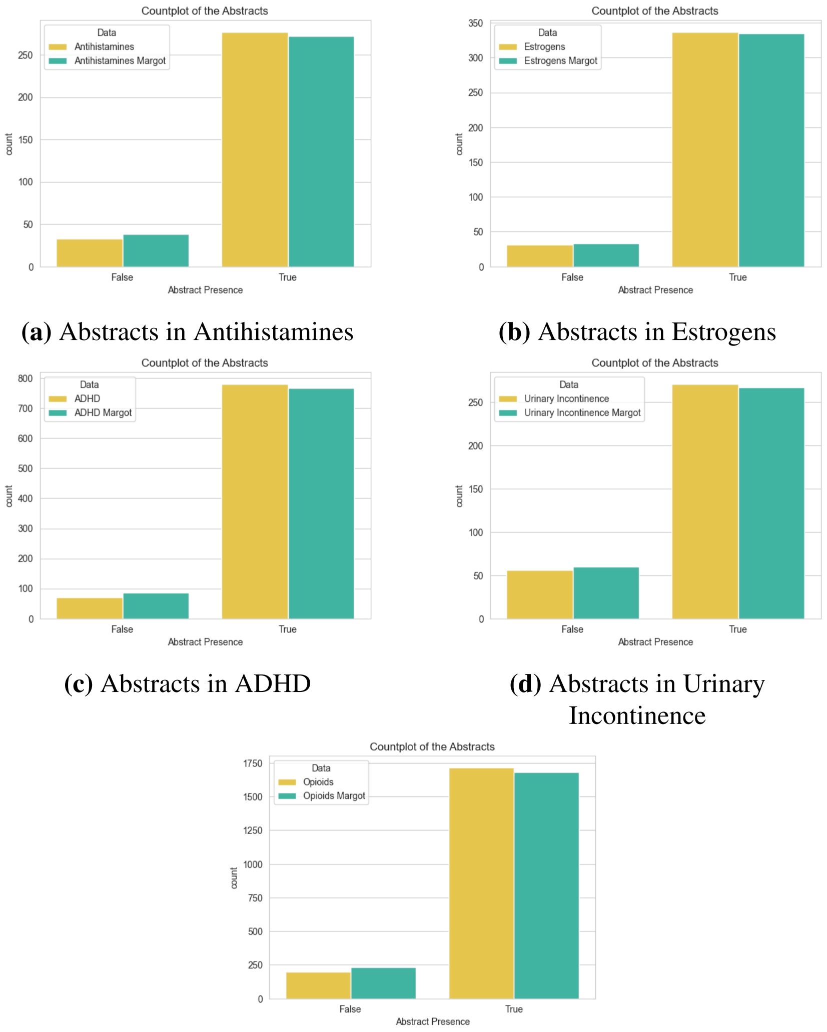


(a) Abstracts in PTSD

(b) Abstracts in ACE

Figure 2.2: Missing Abstracts in high WSS@95 data
Histograms of missing/non-missing abstracts in the data with a WSS@95 greater than 0.5

full version and 233 missing abstracts in the AM version. Figure 2.3 illustrates the frequency of missing and non-missing abstracts for the data with a low WSS@95.



(a) Abstracts in Antihistamines

(b) Abstracts in Estrogens

(c) Abstracts in ADHD

(d) Abstracts in Urinary Incontinence

(e) Abstracts in Opioids

Figure 2.3: Missing Abstracts in low WSS@95 data
Histograms of missing/non-missing abstracts in the data with a WSS@95 lower than 0.5

2.7.2 Analysis of missing abstract fields and data completeness in relevant records

For each dataset, I conduct a comprehensive examination on the record labelled as relevant to identify missing abstract fields and instances where MARGOT failed to detect argumentative rich information within the abstracts.

PTSD includes 38 relevant records, all of which have complete abstract fields. Similarly, ACE has 41 relevant records without missing abstracts. However, in Urinary Incontinence 2 of the 40 relevant records have an empty abstract field. In Antihistamines, all 16 relevant records contain the abstract field. Concerning the Estrogens data, a total of 80 records were deemed relevant, but 3 of them have missing abstracts. In ADHD, out of the 20 relevant records, 19 have complete abstracts. Finally, within the Opioids dataset, there are 15 relevant records of which 1 contains an empty abstract.

In relation to all the relevant records in the PTSD, ACE, Opioids, Estrogens, and ADHD datasets, MARGOT effectively identifies claims and evidence, demonstrating its ability in extracting argumentative rich information from the relevant abstracts. In the case of Antihistamines, MARGOT successfully finds arguments in all the relevant records. The only exception is the abstract field with index 91. This is may be due to the fact that the sentences in this abstract mainly provide information and results, rather than explicitly stating claims and supporting evidence.

2.7.3 Examining Abstract Length and Investigating MARGOT's Potential in Systematic Reviews

To shed light on potential insights gained from the use of MARGOT within the process of systematic review conducted with ASReview LAB, I carry out statistics on the corpora. In order to deepen the cases where MARGOT finds no argumentative rich information, I calculate the average and the median of the lengths of the abstracts fields in both the original and in the AM datasets. I also consider the minimum and the maximum abstract lengths where MARGOT doesn't find any claims and supporting evidence.

Here, 'length' denotes the aggregate count of characters in a string, including both spaces and punctuation. For each set of data, the average abstract length is computed using the arithmetic mean, without taking into account the empty abstracts fields. Let m_i be the number of non-empty abstracts in a dataset i and l_{ij} the length of the jth abstract in the dataset i. The average abstract length for a dataset i can be calculated as:

$$\text{Average Abstract Length}_i = \frac{1}{m_i} \sum_{j=0}^{m_i - 1} l_{ij}$$

Analogously, the average number of claims and evidence detected by MARGOT is calculated through the arithmetic mean which considers the total number of detected claims and evidence divided by the total number of files, excluding empty ones. Let me denote m as the number of non-empty abstracts that a dataset i has. Within each dataset i, let CE_{ij} represent the number of claims and evidence detected in the j-th abstract field. The average number of claims and evidence (CE) found by MARGOT in a dataset i can be calculated as:

$$\text{Average CE}_i = \frac{1}{m_i} \sum_{j=0}^{m_i-1} CE_{ij}$$

The median represents the middle value obtained by sorting the abstract lengths in ascending order. This measure is also considered as it is robust to outliers and it can help to understand the distribution of the length of the abstracts since it provides insights into the central tendency around which the lengths are distributed. Table 2.5 shows the outcomes resulting from these calculations. It can be noted that, indicatively, the average length of the AM abstracts is halved compared to the full-abstracts. Additionally, the values are normally distributed as the median and the average are usually close in values for both the original and the MARGOT data. Moreover, the average number of claims and evidence is the nearly similar for each set of data. This may be because the average abstracts length doesn't vary a lot within the dataset as academic abstracts often follow a standardised structure that usually includes specific sections such as backgrounds, methods, results, and conclusions [52]. Hence, it helps MARGOT to consistently identify claims and evidence in different abstracts.

Figure 2.4 shows respectively the average length of both the original and the MARGOT processed abstracts, as in Table 2.5. Moreover, it shows also the average, the minimum and the maximum abstracts length of the original abstracts where MARGOT failed to identify any relevant claims and evidence. It can be observed that the average length of the abstracts is longer than the average length of the abstracts that MARGOT fails to extract.

It is remarkable to say that, in the PTSD data, there is an outlier, namely the abstract in record with index 133, which has 14687 characters but MARGOT still cannot find any relevant

Dataset	Original		Margot		
	Avg	**Median**	**Avg**	**Median**	**Avg Claims-Ev.**
PTSD	1537	1459	892	834	5.00
ACE	1623	1637	846	788	4.00
Antihistamines	1511	1466	746	721	4.34
Estrogens	1789	1726	863	790	4.99
ADHD	1531	1506	781	734	4.62
Urinary Incontinence	1537	1535	766	751	4.77
Opioids	1463	1461	719	663	4.47

Table 2.5: Average and Median of Abstracts Lengths and of Claims and Evidences detected by MARGOT

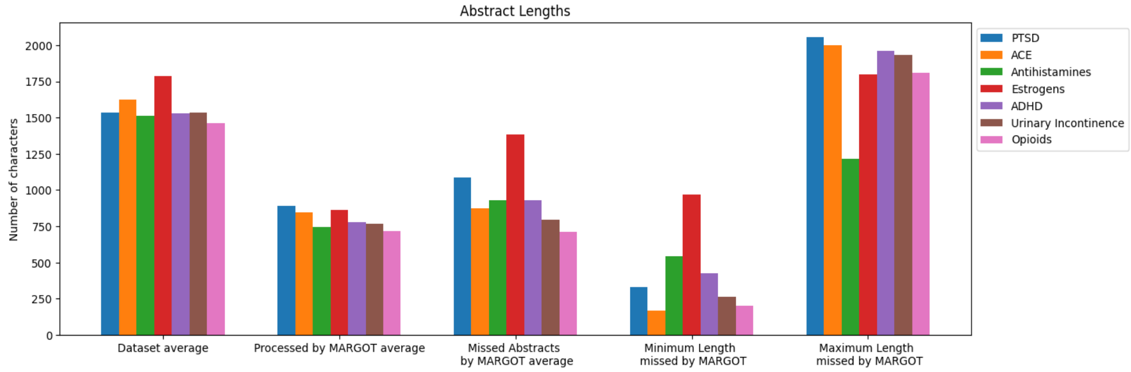


Figure 2.4: Statistics on the abstract fields

claims and evidence. The reason may be attributed to the fact that the text stored in this abstract field primarily consists of a descriptive list of topics and papers presented at a conference on child and adolescent psychopharmacology. Therefore, it does not provide specific claims or supporting evidence that can be detected by MARGOT. As it is labelled as non relevant, no operation is performed on this record. If this outlier is discarded in the plot and the second maximum is considered, the maximum abstract length that MARGOT fails to detect is in line with the ones of the others data.

2.8 Analytical Strategy of the simulation studies

After analysing the various aspects of the data and processing them, I carry out the simulations studies first on the dataset that scored an high WSS@95 and then on the dataset with low WSS@95 in the simulation of [33]. I then present the outcomes of the various models for each simulation, making a comparative analysis between the original data and the ones processed by MARGOT (AM data). Then, I evaluate the models' performance relying on the Recall@10, WSS@95 and ATD. The Recall is the main metric used as a yardstick since it is a measure of quantity. It is measured at 10% of screened records. This means that if a Recall of 1.0 is reached, all relevant records have been found at 10 % of the screening phase. The WSS@95 calculates the percentage of records that a screener can skip after receiving prior knowledge used to train the initial model iteration compared to random reading [12]. The WSS is typically measured at a Recall of 0.95 (WSS@95) and it a commonly used benchmark reflecting the proportion of records saved through active learning while still missing 5% of relevant publications [49]. The last measure is chosen in order to investigate the difference in speed in discovering a relevant record between the AM dataset and the original dataset with the measurement unit being seconds.

After concluding this part, I analyse the effect of missing abstracts on two optimally performing datasets belonging to two different topics.

In a subsequent step of this work, I investigate the presence of records that contain neither the abstract nor the title field. If these type of records are labelled as relevant, they are removed. Conversely, if deemed irrelevant, they are retained as such. This choice is mainly because, for ASReview LAB, is still possible to learn about the irrelevance from these records but not to capture meaningful insights about the relevance. Then, a new ASReview simulation is carried out on the cleaned datasets in order to analyse the influence of false positives in a systematic review. At the end of this phase, an analysis of the results is presented. In the final stage, the stability of the obtained results is examined by conducting 15 simulations for each dataset (full-abstracts and AM) on the Exoscale Cloud Service using a customised multiple model template. These simulations are run on an CPU optimised computer Cluster due to limited computational resources on the local machine used for the initial experiment. The total number of simulations is 2100, namely $10*15*14$. Where 10 is the number of models, 15 is the number of simulations performed and 14 are the datasets (as each dataset comes with its own AM version). Following this, I perform a comparative analysis of the average performance with respect to the performance obtained on the previous simulations as well as a quantitative analysis of the results obtained. Then, I set up a script in which all the outcomes are fetched from the folders of the simulations. Lastly, a new dataset that has as following features is created:

- dataset name
- classifier
- feature extractor
- seed used during the simulation
- Recall@10 score obtained from that combination of classifier + feature extractor with the full-abstracts data
- Recall@10 score obtained from that combination of classifier + feature extractor with the AM abstracts data
- WSS@95 score obtained from that combination of classifier + feature extractor with the full-abstracts data
- WSS@95 score obtained from that combination of classifier + feature extractor with the AM abstracts data

From the pairing of the results by the seed of the simulation, the dataset results in a total of 1050 rows. Finally, the Wilcoxon signed-rank test, a non-parametric statistical test, is used to compare the medians of paired samples [53]. This test assesses whether there's a statistically significant difference between the medians of two paired sets [54]. In this study, I compare the Recall@10 scores from the original data with those from the AM data. The same is done for the WSS@95 scores.

3. Results

After completing the preceding stages, through Makita it is possible to conduct a large-scale simulation study leveraging the CLI interface of ASReview LAB. This chapter offers a comprehensive analysis of the systematic review results. It presents the outcomes for each dataset, comparing the full-abstracts version with the corresponding AM version. In all the tables, the best result is displayed in bold while the second best is underlined. Each simulation study took a few hours and varied depending on the size of the dataset. The results reported are rounded to two decimal places.

3.1 Case study 1: Systematic Review on data with high WSS@95

3.1.1 The PTSD Data

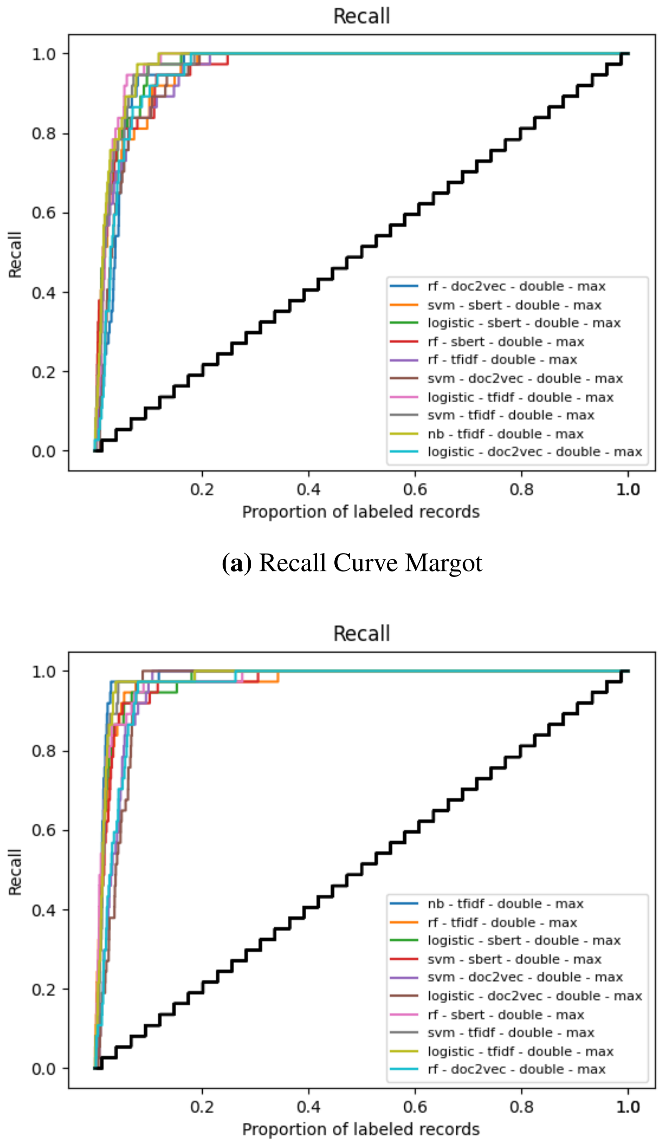
The outcomes presented in Table 3.1 indicate that, under two circumstances, PTSD and its AM version (AM PTSD) have identical Recall@10 values: when using LR and NB as classifiers and TF-IDF as feature extractor. However, the full-abstracts version of the data outperforms the AM one with respect to ATD and WSS@95 in both cases. Although when using LR-TFIDF with AM PTSD there's a small decrease in WSS@95 and ATD scores, it is achieved first the highest recall score (the proportion of labelled records is smaller). Specifically, the original dataset reaches the maximum recall score with roughly 20% of labelled records, while AM PTSD requires approximately 15%. This behaviour can be observed by comparing the recall scores of LR-TFIDF in Figure 3.1. It may be due to the fact that the representation of the text is denser with MARGOT as it captures only the most argumentative rich information.

Nevertheless, when NB is deployed as the classifier, the change in proportion is not significantly different. This leads to the assumption that the difference between NB and LR plays a key role. Based on this results, LR appears to achieve the recall faster with the AM data due to the nature of condensed text and the fact that this model is designed to learn the rela-

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	**PTSD**	**1.0**	**0.86**	**246.08**
		AM PTSD	0.89	0.82	259.27
LR	SBERT	PTSD	0.95	0.86	153.49
		AM PTSD	0.97	0.83	185.81
LR	TFIDF	PTSD	0.97	0.90	117.30
		AM PTSD	0.97	0.87	156.41
NB	TFIDF	PTSD	0.97	0.90	91.81
		AM PTSD	0.97	0.85	151.62
RF	Doc2Vec	PTSD	0.97	0.86	236.54
		AM PTSD	0.95	0.85	257.19
RF	SBERT	PTSD	0.97	0.86	149.92
		AM PTSD	0.84	0.82	227.24
RF	TFIDF	PTSD	0.97	0.88	160.27
		AM PTSD	0.84	0.77	257.70
SVM	Doc2Vec	PTSD	0.95	0.84	224.35
		AM PTSD	0.86	0.80	277.92
SVM	SBERT	PTSD	0.92	0.83	181.24
		AM PTSD	0.84	0.78	240.16
SVM	TFIDF	PTSD	0.97	0.89	124.95
		AM PTSD	0.95	0.86	177.08

Table 3.1: Performance results on the PTSD dataset

tionship between predictors, in this case the TF-IDF features, and outcomes. In general, the dataset in its original form exhibits superior performance compared to the version processed by MARGOT. However, AM PTSD surpasses the Recall@10 level of PTSD when using a LR as classifier and SBERT as feature extractor. This result may be explained by the discrepancy in word processing capabilities between ASReview and BERT-based models. The former has no limits in the number of words it can process whereas the latter is programmed to restrict input tokens to the initial 512 [48], equivalent to roughly 400 words for English text [55]. Hence, by condensing the richer parts of the text with MARGOT, only the fundamental features can be extracted and the text will probably be truncated to a lesser extent or not at all. Additionally, the shorter abstract fields could result in less noise, which may make it easier for the models to identify true positives. It is important to emphasise that this data in its original form already exhibits excellent performance and AM data did not lead to any particular improvement. However, MARGOT manages to improve the Recall@10 at the expense of a slight deterioration of WSS@95 and ATD when carrying out a systematic review using LR-SBERT.



(a) Recall Curve Margot

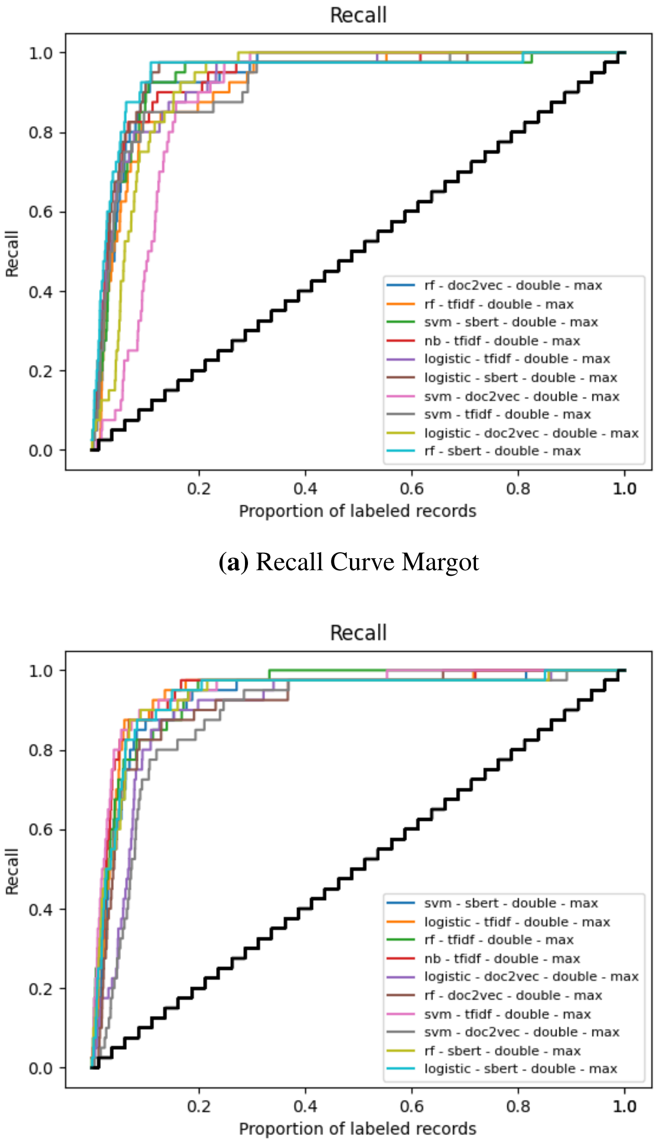
(b) Recall Curve Original

Figure 3.1: Recall Plot of PTSD Data

3.1.2 The ACE Data

In the context of the ACE dataset, changes in the outcomes of the models have been observed. Table 3.2 reveals that the application of a LR classifier with SBERT as feature extractor leads to an improvement in the performance of systematic reviews, similar to PTSD. This enhancement is not limited to a 0.02 increase in Recall@10 alone, but also includes a 0.03 increase in WSS@95 and a significant decrease in ATD. The recall curves in Figure 3.2 demonstrate that MARGOT attains the peak recall level at approximately 70% of labelled records, compared to the original ACE data, which achieves its maximum recall at around 85% of the labelled records.

The performance of the systematic review in the PTSD dataset was significantly reduced due to the use of the combination of RF classifier and SBERT with MARGOT. Here, by contrast, AM ACE exhibits a significant enhancement. This is demonstrated by an increase of 0.03 in Recall@10 level, a 0.07 increase in WSS@95, and a reduction in ATD. The recall curves in Figure 3.2 indicate that AM ACE attained a maximum recall at a faster pace, at approx. 80% of labelled records compared to the 85% of the original data for this combination of models.
Notably, the highest level of performance was achieved by the AM data when used in conjunction with the aforementioned models (RF-SBERT), outperforming all the other models in both the versions of the data. In this way, it exceeds the best performance previously obtained for the full-abstract version. MARGOT demonstrates efficacy in another combination of models, namely RF in conjunction with Doc2Vec. While its effectiveness slightly diminished in the PTSD dataset, it exhibited considerable enhancement in this dataset, resulting in a 0.05 increase in Recall@10, a 0.13 increase in WSS@95, and a significant reduction in ATD.
Furthermore, AM ACE attains the maximum recall level at a substantially faster rate than ACE. It achieves this level at approximately 30% of labelled records, as opposed to around 65% in the full-abstracts version. A small room for improvement is also noticeable when using the SVM-SBERT models: the Recall@10 remains the same for both datasets, but the ATD and WSS@95 scores improve when using MARGOT. A singular behaviour is observed with SVM-Doc2Vec in AM ACE. In fact, there is an important decrease of the Recall@10 but an increase in performance at the ATD and WSS@95. To conclude, for some model combinations, better performance is achieved with MARGOT than with the original data. In particular, the RF-SBERT models with MARGOT led to the most effective systematic reviews in the ACE dataset, achieving the highest performance among all models tested.



(a) Recall Curve Margot

(b) Recall Curve Original

Figure 3.2: Recall Plot of ACE Data

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	ACE	0.80	0.62	251.65
		AM ACE	0.75	0.74	205.18
LR	SBERT	ACE	0.88	0.79	166.4
		<u>AM ACE</u>	<u>0.90</u>	<u>0.82</u>	<u>143.88</u>
LR	TFIDF	ACE	0.88	0.80	148.33
		AM ACE	0.80	0.70	176.83
NB	TFIDF	ACE	0.88	0.78	141.38
		AM ACE	0.83	0.72	165.70
RF	Doc2Vec	ACE	0.83	0.57	212.75
		AM ACE	0.88	0.70	152.33
RF	SBERT	ACE	0.90	0.76	166.13
		AM ACE	**0.93**	**0.83**	**134.65**
RF	TFIDF	ACE	0.83	0.76	142.15
		AM ACE	0.83	0.65	211.13
SVM	Doc2Vec	ACE	0.73	0.65	295.23
		AM ACE	0.45	0.69	293.40
SVM	SBERT	ACE	0.85	0.75	185.15
		AM ACE	0.85	0.78	174.18
SVM	TFIDF	ACE	0.90	0.75	124.70
		AM ACE	0.85	0.64	211.30

Table 3.2: Performance results on the ACE dataset

3.2 Case study 2: Systematic Review on data with low WSS@95

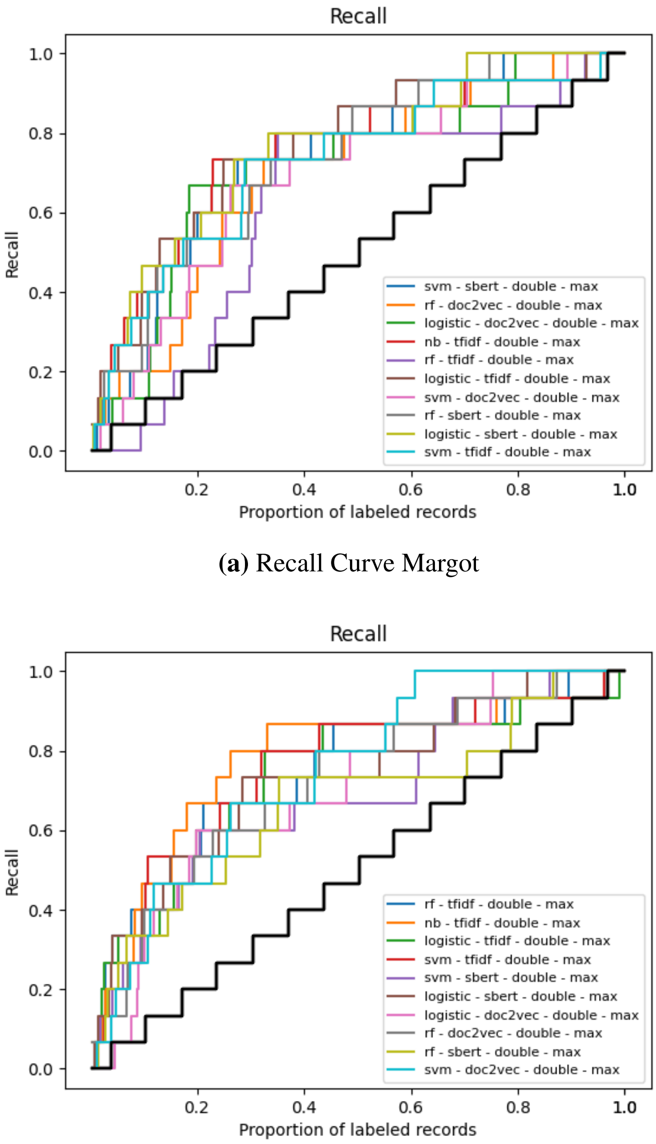
In this section, we focus on data with a low (< 0.5) WSS@95. This approach aims to investigate potential benefits that may arise from integrating MARGOT in the process of systematic review, under the hypothesis that extracting topics from the abstracts of low-performing data can potentially improve the quality of the systematic review.

3.2.1 The Antihistamines Data

Antihistamines is the dataset which obtained the lowest WSS@95 score in the simulation of [33]. The findings presented in Table 3.3 indicate that MARGOT, in combination with LR and SBERT, enhances the Recall@10 score, as in the previous datasets. Additionally, as depicted in Figure 3.3, the proportion of labelled data corresponding to the recall level attained through MARGOT is approximately 0.70%, whereas it is marginally higher when using the original data, i.e 0.80%. When combined with LR and TFIDF, with MARGOT there's an increase in both Recall@10 and WSS@95 and a decrease in ATD. Although the Recall@10 is 0.07 lower than the highest score, the WSS@95 and the ATD register the best scores demonstrating remarkable validity when using the AM data with these models. It is of note that when using RF in conjunction with SBERT, the AM data achieves the maximum recall level earlier than the original dataset, despite the slight worsening in terms of Recall@10 score. For the same models, the AM data has a difference of 0.18 higher than the original data and a decrease in ATD of approximately 20 seconds. The exact same behaviour can be observed when using SVM together with SBERT, but this time the Recall@10 shows a significant decrease. The use of MARGOT improves the WSS@95 scores for several models, such as RF-SBERT, NB-TFIDF, SVM-TFIDF, SVM-SBERT and LR-TFIDF. Looking at all three metrics, with a preference for both the Recall@10 and WSS@95, LR-SBERT models when using MARGOT leads to the highest performance.

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	ANT	0.27	0.15	91.67
		AM ANT	0.13	0.12	87.80
LR	SBERT	ANT	0.40	0.22	82.47
		AM ANT	**0.47**	**0.21**	**73.73**
LR	TFIDF	ANT	0.33	0.10	81.0
		AM ANT	0.40	0.33	73.33
NB	TFIDF	ANT	0.47	0.14	69.80
		AM ANT	0.40	0.20	75.67
RF	Doc2Vec	ANT	0.33	0.21	87.67
		AM ANT	0.20	0.19	93.73
RF	SBERT	ANT	0.33	0.11	101.40
		AM ANT	0.27	0.29	80.73
RF	TFIDF	ANT	0.40	0.13	79.07
		AM ANT	0.07	0.02	115.13
SVM	Doc2Vec	ANT	0.27	0.33	78.80
		AM ANT	0.20	0.20	95.47
SVM	SBERT	ANT	0.40	0.22	96.20
		AM ANT	0.20	0.33	78.67
SVM	TFIDF	ANT	0.40	0.18	76.60
		AM ANT	0.33	0.26	84.47

Table 3.3: Performance results on the Antihistamines dataset



(a) Recall Curve Margot

(b) Recall Curve Original

Figure 3.3: Recall Plot of Antihistamines Data

3.2.2 The Urinary Incontinence Data

This subsection reports the analysis of the Urinary Incontinence dataset. Table 3.4 shows the outcomes of the simulations. In the AM version of the data, a remakable improvement in terms of Recall@10 and WSS@95 is obtained when using LR as a classifier Doc2Vec as a feature extractor. In addition to the positive results, the proportion of labelled records is even lower, as Figure 3.4 shows. Classifiers like RF and SVM with TF-IDF as feature extractors, exhibit good Recall@10 performance with the AM data, even outperforming the original version with a dif-

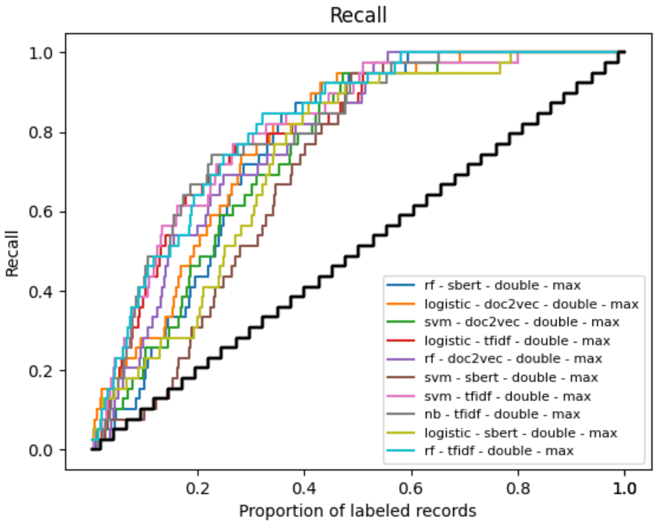
Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	UR	0.23	0.32	72.10
		AM UR	0.28	0.48	71.18
LR	SBERT	UR	0.13	0.36	85.23
		AM UR	0.21	0.40	86.51
LR	TFIDF	UR	0.44	0.46	57.90
		AM UR	0.38	0.43	64.79
NB	TFIDF	**UR**	**0.46**	**0.42**	**58.08**
		AM UR	0.41	0.38	64.79
RF	Doc2Vec	UR	0.28	0.44	68.67
		AM UR	0.28	0.42	72.46
RF	SBERT	UR	0.28	0.31	79.62
		AM UR	0.15	0.46	77.18
RF	TFIDF	UR	0.38	0.46	56.49
		AM UR	0.41	0.42	60.59
SVM	Doc2Vec	UR	0.23	0.16	93.26
		AM UR	0.21	0.46	82.26
SVM	SBERT	UR	0.21	0.36	76.18
		AM UR	0.08	0.45	94.82
SVM	TFIDF	UR	0.10	0.48	69.10
		AM UR	0.38	0.43	62.74

Table 3.4: Performance results on the Urinary Incontinence dataset

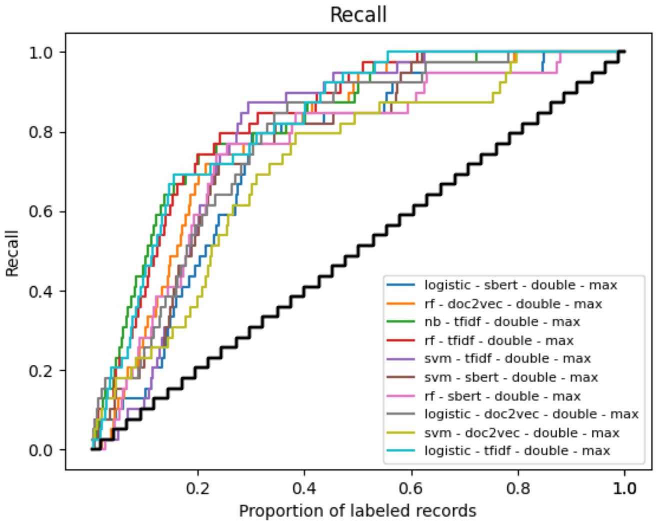
ference of 0.28 using SVM together with TFIDF. This time, this latest combination of models with AM Urinary Incontinence improves the Recall@10, whereas in previous cases it did not. While the Recall@10 value is lower, an increase in the WSS@95 is observed using RF and SVM in conjunction with SBERT and SVM with Doc2Vec. These findings supports the assumptions that the performance of machine learning models is highly dependent on the structure, organisation, and size of the dataset at hand and that the models are highly data-dependent.

A recurring pattern in this study is the use of LR and SBERT models, which leads improvements when using MARGOT. Analogously to the previous cases, the results can be explained by the fact that the combination of SBERT embeddings and LR may complement each other: SBERT generates semantically rich embeddings and LR effectively models the relationships between those embeddings and labels. If the embeddings capture significant semantic information about the abstracts, LR can effectively discern meaningful connections with the labels.

In this case, however, it appears that the AM data do not outperform the systematic review with the original data, where the best overall performance is achieved when using NB-TFIDF.



(a) Recall Curve Margot



(b) Recall Curve Original

Figure 3.4: Recall Plot of Urinary Incontinence Data

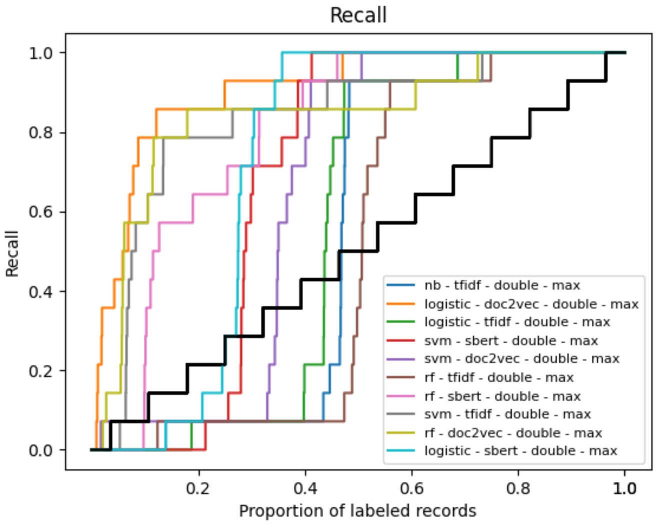
3.2.3 The Opioids Data

The results of the systematic review carried out for Opioids are presented in Table 3.5. The outcomes are highly unsatisfactory, thereby suggesting an error in the data. Moreover, for both AM and full-abstracts Opioids, there are models that have the recall curve that grows slower than the random curve, suggesting that the models are learning incorrectly. This behaviour is evident in Figure 3.5. Looking at the data metric file generated during the simulation, there are some records that present an negative Extra Relevant records Found (ERF) score. This score represents *”the proportion of relevant records found after correcting for the number of relevant records found via random screening (assuming a uniform distribution of relevant records)”* [49]. The records which have a negative ERF score are 1174, 518 and 850. The hypothesis is that as the multiple model template selects one irrelevant and one relevant record that are the same across all models. During the systematic review, the default seed initially selects one of these erroneously labelled record and begins to carry out a systematic review where the models are learning in a wrong way. In order to correct this error, another dataset was created without these rows.

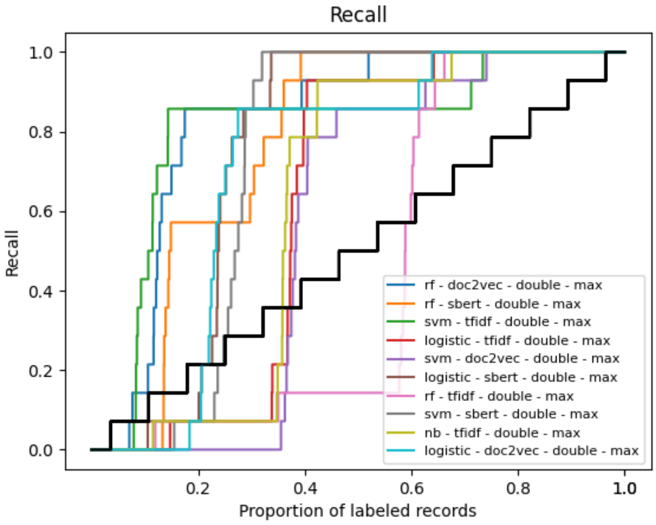
The outcomes from the simulations are shown in Table 3.6, a significant increase in performance can be seen, which means that the problem has been solved. The best and the second best model have identical performance, the only difference is that the best model is minimally faster. Figure 3.6 shows the recall plot of these outcomes.

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	OP	0.0	0.28	547.86
		AM OP	0.79	0.64	180.64
LR	SBERT	OP	0.0	0.56	464.79
		AM OP	0.0	0.55	519.86
LR	TFIDF	OP	0.0	0.49	721.86
		AM OP	0.0	0.42	840.07
NB	TFIDF	OP	0.0	0.47	715.43
		AM OP	0.1	0.41	879.07
RF	Doc2Vec	OP	0.14	0.50	327.93
		AM OP	0.57	0.29	307.21
RF	SBERT	OP	0.0	0.53	431.00
		AM OP	0.21	0.50	380.71
RF	TFIDF	OP	0.0	0.25	1052.21
		AM OP	0.0	0.33	959.21
SVM	Doc2Vec	OP	0.0	0.27	819.07
		AM OP	0.07	0.48	668.21
SVM	SBERT	OP	0.1	0.59	505.86
		AM OP	0.0	0.51	590.14
SVM	TFIDF	OP	0.43	0.18	369.36
		AM OP	0.57	0.45	322.21

Table 3.5: Performance results on the Opioids dataset

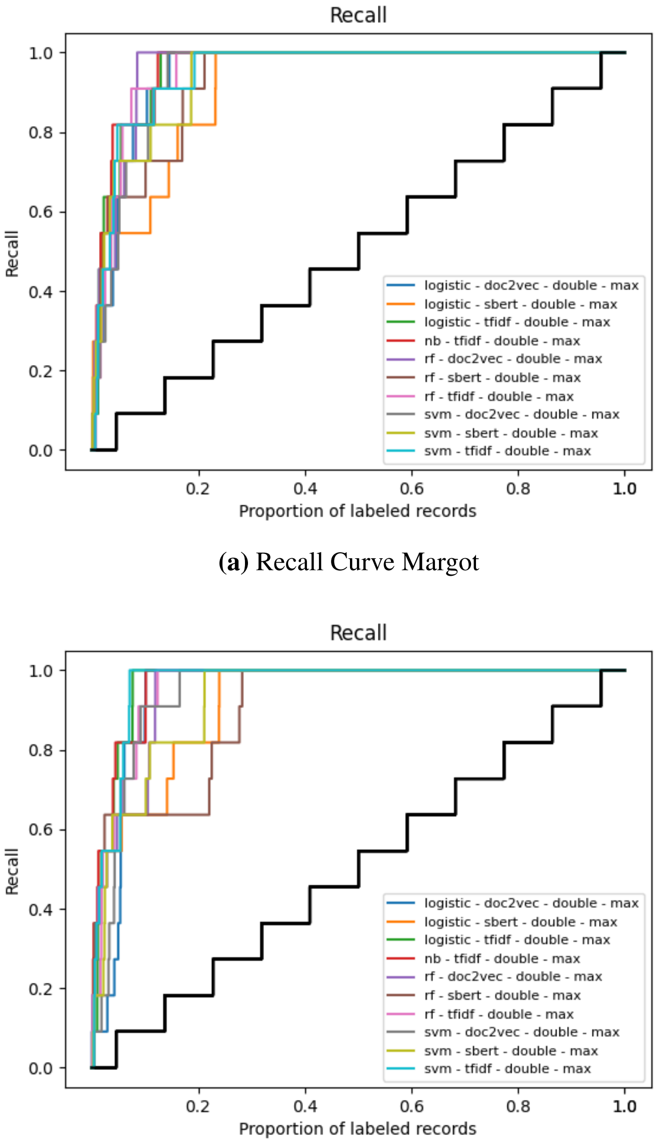


(a) Recall Curve Margot



(b) Recall Curve Original

Figure 3.5: Recall Plot of Opioids Data



(a) Recall Curve Margot

(b) Recall Curve Original

Figure 3.6: Recall Plot of Opioids Data

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	OP	0.91	0.76	105.73
		AM OP	0.82	0.76	99.64
LR	SBERT	OP	0.64	0.62	155.09
		AM OP	0.55	0.63	163.45
LR	TFIDF	OP	**1.0**	**0.79**	**62.55**
		AM OP	0.82	0.75	78.45
NB	TFIDF	OP	0.82	0.76	63.36
		AM OP	0.82	0.75	71.91
RF	Doc2Vec	OP	0.64	0.74	97.36
		AM OP	1.0	0.78	76.82
RF	SBERT	OP	0.64	0.59	188.45
		AM OP	0.64	0.69	132.64
RF	TFIDF	OP	0.91	0.78	83.18
		AM OP	0.91	0.79	87.18
SVM	Doc2Vec	OP	0.91	0.77	109.09
		AM OP	0.73	0.75	107.82
SVM	SBERT	OP	0.64	0.65	136.09
		AM OP	0.73	0.68	111.00
SVM	TFIDF	OP	1.0	0.79	64.27
		AM OP	0.82	0.75	94.27

Table 3.6: Performance results on the Opioids dataset without the rows with the negative ERF score

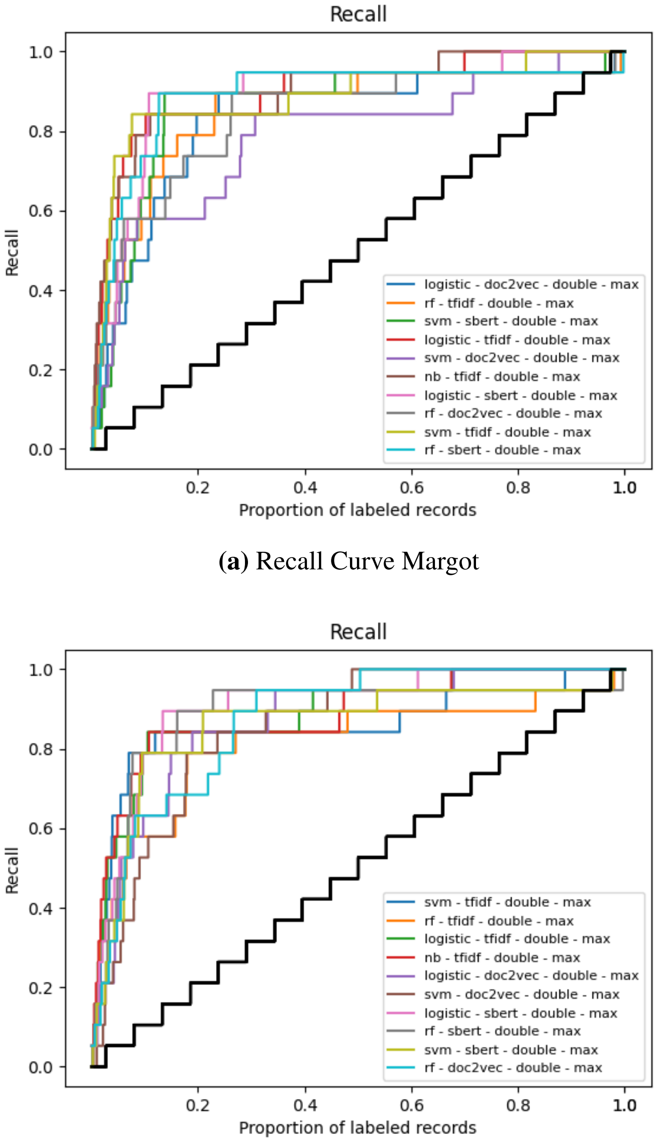
3.2.4 The ADHD Data

Table 3.7 shows the results of the ADHD dataset. Interestingly, the performance of all models that included SBERT as a feature extractor was found to be lower when using MARGOT, in contrast to the previous datasets. It is plausible that, this time, MARGOT has summarised the abstracts in such a way that SBERT encountered difficulty in extracting relevant features.

MARGOT proved to be essential in achieving the highest level of Recall@10, reaching a value of 84% with the SVM-TFIDF models. Figure 3.7 indicates that this value was attained more quickly. Moreover, MARGOT, with classifiers like RF, NB and LR, demonstrated promising results in terms of WSS@95 when using TFIDF as feature extractor.

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	ADHD	0.63	0.58	112.16
		AM ADHD	0.47	0.31	144.32
LR	SBERT	ADHD	0.79	0.67	80.63
		AM ADHD	0.68	0.64	89.84
LR	TFIDF	ADHD	0.79	0.51	95.84
		AM ADHD	0.79	0.56	84.89
NB	TFIDF	ADHD	0.79	0.45	98.05
		AM ADHD	0.79	0.55	85.42
RF	Doc2Vec	ADHD	0.63	0.61	109.37
		AM ADHD	0.58	0.35	138.68
RF	SBERT	ADHD	0.79	0.69	98.26
		AM ADHD	0.74	0.65	96.53
RF	TFIDF	ADHD	0.58	0.09	160.0
		AM ADHD	0.58	0.42	126.52
SVM	Doc2Vec	ADHD	0.53	0.48	124.42
		AM ADHD	0.58	0.21	182.00
SVM	SBERT	ADHD	0.79	0.39	120.95
		AM ADHD	0.63	0.47	119.11
SVM	TFIDF	ADHD	0.79	0.26	121.68
		AM ADHD	**0.84**	**0.44**	**97.68**

Table 3.7: Performance results on the ADHD dataset



(a) Recall Curve Margot

(b) Recall Curve Original

Figure 3.7: Recall Plot of ADHD Data

3.2.5 The Estrogens Data

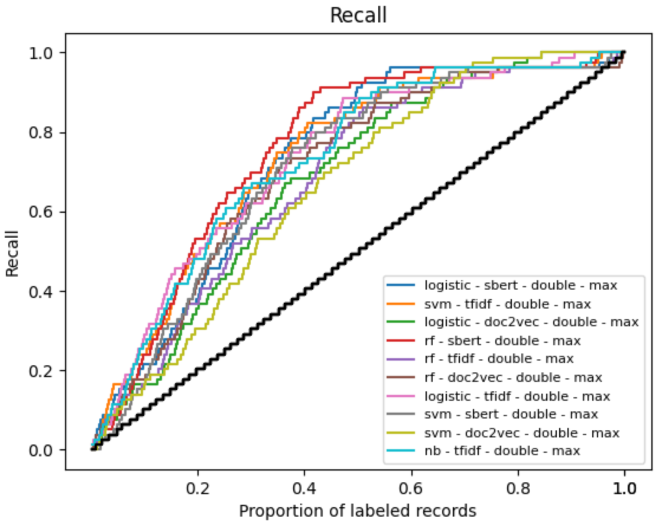
The final dataset under study is the Estrogens dataset. From Table 3.8, it is worth noting that, with the AM data, LR-SBERT models help to achieve better results. Also, when using RF together with SBERT the Recall@10 is increased by 0.05.

The best performing model is NB-TFIDF with original data, the second best model is LR-TFIDF with AM data. Although the overall performance of the systematic review is not very high, it must be taken into account that with the 21% of relevant records the maximum Re-

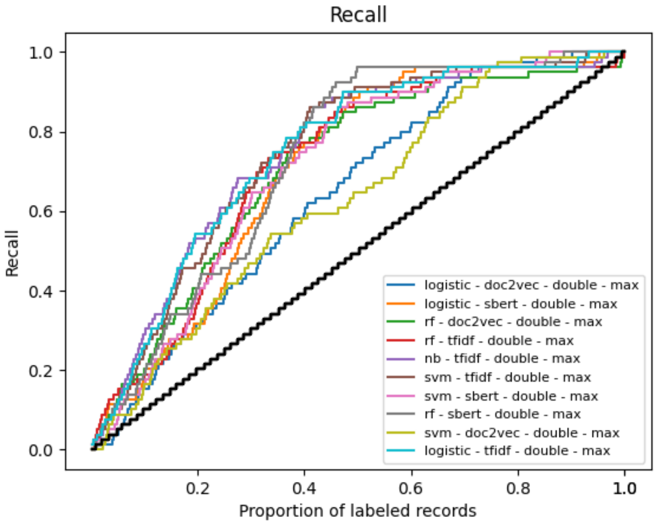
Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	ES	0.15	0.23	130.81
		AM ES	0.16	0.23	116.68
LR	SBERT	ES	0.18	0.36	107.76
		AM ES	0.22	0.39	99.29
LR	TFIDF	ES	0.25	0.28	92.97
		AM ES	0.28	0.20	97.70
NB	TFIDF	**ES**	**0.29**	**0.26**	**92.14**
		AM ES	0.27	0.30	98.75
RF	Doc2Vec	ES	0.19	0.13	106.53
		AM ES	0.18	0.26	180.09
RF	SBERT	ES	0.19	0.45	99.62
		AM ES	0.24	0.36	89.25
RF	TFIDF	ES	0.18	0.28	104.24
		AM ES	0.16	0.22	116.11
SVM	Doc2Vec	ES	0.15	0.20	135.97
		AM ES	0.15	0.25	123.70
SVM	SBERT	ES	0.15	0.28	106.10
		AM ES	0.15	0.27	105.67
SVM	TFIDF	ES	0.24	0.31	93.38
		AM ES	0.24	0.19	97.19

Table 3.8: Performance results on the Estrogens dataset

call@10 level achievable is approximately 45%.



(a) Recall Curve Margot



(b) Recall Curve Original

Figure 3.8: Recall Plot of Estrogens Data

3.3 Effect of Missing Abstracts on systematic reviews

3.3.1 The PTSD Data

Due to the significant percentage of missing abstract in PTSD data, i.e 12.9%, a complementary dataset was generated to examine the potential impact of absent abstracts on the systematic review. This new dataset comprises 4938 entries. It excludes all the missing abstracts, including the ones not detected by MARGOT.

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	PTSD	0.95	0.87	226.86
		AM PTSD	0.89	0.81	254.86
LR	SBERT	PTSD	0.95	0.87	144.73
		AM PTSD	0.95	0.84	168.27
LR	TFIDF	PTSD	0.97	0.89	98.41
		AM PTSD	0.95	0.87	138.68
NB	TFIDF	**PTSD**	**0.97**	**0.90**	**88.35**
		AM PTSD	0.97	0.85	138.30
RF	Doc2Vec	PTSD	0.97	0.86	144.38
		AM PTSD	0.89	0.74	254.22
RF	SBERT	PTSD	0.92	0.81	168.03
		AM PTSD	0.84	0.80	198.22
RF	TFIDF	PTSD	0.97	0.86	132.05
		AM PTSD	0.78	0.73	277.30
SVM	Doc2Vec	PTSD	0.97	0.86	210.41
		AM PTSD	0.86	0.78	264.68
SVM	SBERT	PTSD	0.95	0.85	167.41
		AM PTSD	0.89	0.79	201.76
SVM	TFIDF	PTSD	0.97	0.89	112.43
		AM PTSD	0.97	0.86	154.83

Table 3.9: Performance results on the PTSD dataset without missing abstract

From Table 3.9, the overall systematic review scores are slightly lower than in the dataset with the missing abstracts. For ASReview LAB, titles also play a key role. Thus, if records with missing abstracts are removed, there is a risk of discarding titles that hold importance in discerning the relevance of a given record. Moreover, ASReview LAB is also able to learn patterns from the empty fields. Nevertheless, a decrease in ATD is noticed in almost all the models. This may support the hypothesis that the presence of abstracts is decisive for the improvement of ATD. It is interesting to note that despite the worsening of some models, there is a slight improvement of others both with both MARGOT and the original version of the data. Furthermore, models using SVMs as classifiers have a slight improvement in Recall@10 and WSS@95. This may be due to the fact that the SVM model can make more accurate predictions when it comes to smaller set of data [56]. In conclusion, although the differences are not remarkable, the version of the datasets with the missing abstracts is slightly better.

3.3.2 Opioids Data

As the percentage of missing abstracts is very high, i.e. 10.34% and the topic is different from PTSD, I want to investigate whether there is room for improvement by removing the missing abstracts. In the AM Opioids there are 233 empty abstracts field, this number includes both abstracts in which MARGOT found no relevant information (35) and the number of empty abstracts in the dataset. In order to make a fairer comparison, I remove these entries from both the versions of the data without the rows with negative ERF. Table 3.10 displays the outcomes of the simulation. This clean version of the data, gives commendable results for all the models. It is worth to notice that SBERT-based models improve the WSS@95 and ATD when using the AM data. The best overall performance, is achieved with RF-Doc2Vec with the AM data. Since the differences are minimal compared to the previous version and this dataset shows better performance, I keep this new version of Opioids for the next steps.

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	OP	0.82	0.76	110.64
		AM OP	0.91	0.78	78.36
LR	SBERT	OP	0.64	0.65	128.27
		AM OP	0.64	0.72	100.55
LR	TFIDF	OP	1.0	0.78	60.91
		AM OP	0.82	0.73	78.27
NB	TFIDF	OP	0.82	0.75	59.00
		AM OP	0.82	0.74	66.00
RF	Doc2Vec	OP	0.73	0.73	93.55
		AM OP	**1.0**	**0.82**	**46.64**
RF	SBERT	OP	0.73	0.71	93.27
		AM OP	0.73	0.75	94.45
RF	TFIDF	OP	0.82	0.71	95.73
		AM OP	0.73	0.74	91.09
SVM	Doc2Vec	OP	0.82	0.74	104.73
		AM OP	0.82	0.74	105.18
SVM	SBERT	OP	0.82	0.64	110.82
		AM OP	0.73	0.76	72.91
SVM	TFIDF	OP	1.0	0.77	63.36
		AM OP	0.82	0.74	93.45

Table 3.10: Performance results on the Opioids Cleaned dataset

3.4 Effect of Missing Abstracts and Titles on ASReview Simulations

In all the data, with the exception of ACE there are some records which contain neither the abstract field nor the title field. These records are left as such if labelled as irrelevant since ASReview LAB can still learn about the irrelevance of a record. However, information about the relevance cannot be obtained if these features are missing. In the Opioids, Estrogens and ADHD datasets there are some relevant records where these fields are missing. To address this lack, it is worthwhile to try running another simulation by removing the mislabelled records in both the AM and full-abstracts versions of these data and compare the results with the previous simulations. The outcomes are presented in the tables within this section. The data with labelled '-C' excludes the records with missing titles and abstracts.

For Estrogens, the relevant records without abstract and title are: 29, 70, 111. For ADHD it is the one with index 213 and for Opioids it is the one with index 1174. For Opioids, the comparison of the simulations is done with the version of the data in which all records with a negative ERF score and all missing abstracts have been removed. The results for the new versions of Estrogens are shown in Table 3.11 which reports also the ones of the previous simulations.

With Estrogens the best performance was obtained when using NB with TFIDF. This model combination is still the best but, with the clean version of the data, there is an increase in WSS@95 and a decrease in ATD.
The outcomes from the simulations conducted using the ADHD data are presented in Table 3.12. The best performance is achieved with the clean version of the data, emphasising the importance of data quality.

The new version of the Opioids (OP-C) containing all records present in the full version except record 1174 was compared with the version of it where both records 1174, 518 and 850 and all records with missing abstracts were removed. Since the comparison is made with an even cleaner version the outcomes of this simulation are worse. However, the results prove that the removal of the record 1174 is sufficient to correct the error in the data.

In conclusion, the removal of the relevant records with both missing abstracts and titles help to achieve better results. The findings suggest that addressing the issue of missing abstracts and titles in the datasets can have a significant impact on the performance of the systematic review, emphasizing the importance of data quality and preprocessing before starting the process.

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	ES	0.15	0.23	130.81
		ES-C	0.11	0.17	129.38
		AM ES	0.16	0.23	116.68
		AM ES-C	0.14	0.26	121.37
LR	SBERT	ES	0.18	0.36	107.76
		ES-C	0.21	0.40	98.97
		AM ES	0.22	0.39	99.29
		AM ES-C	0.20	0.42	92.42
LR	TFIDF	ES	0.25	0.28	92.97
		ES-C	0.26	0.43	89.91
		AM ES	0.28	0.20	97.70
		AM ES-C	0.18	0.33	108.22
NB	TFIDF	ES	0.29	0.26	92.14
		ES-C	**0.29**	**0.44**	**82.07**
		AM ES	0.27	0.30	98.75
		AM ES-C	0.16	0.40	100.64
RF	Doc2Vec	ES	0.19	0.13	106.53
		ES-C	0.26	0.35	101.74
		AM ES	0.18	0.26	180.09
		AM ES-C	0.12	0.29	111.16
RF	SBERT	ES	0.19	0.45	99.62
		ES-C	0.21	0.52	86.26
		AM ES	0.24	0.36	89.25
		AM ES-C	0.20	0.50	86.50
RF	TFIDF	ES	0.18	0.28	104.24
		ES-C	0.25	0.41	86.46
		AM ES	0.16	0.22	116.11
		AM ES-C	0.22	0.35	97.42
SVM	Doc2Vec	ES	0.15	0.20	135.97
		ES-C	0.09	0.18	146.74
		AM ES	0.15	0.25	123.71
		AM ES-C	0.12	0.25	128.41
SVM	SBERT	ES	0.15	0.28	106.10
		ES-C	0.20	0.40	104.07
		AM ES	0.15	0.27	105.67
		AM ES-C	0.12	0.36	109.14
SVM	TFIDF	ES	0.24	0.31	93.38
		ES-C	0.22	0.45	87.80
		AM ES	0.24	0.19	97.19
		AM ES-C	0.13	0.33	109.18

Table 3.11: Performance results on the Estrogens dataset

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	ADHD	0.63	0.58	112.16
		ADHD-C	0.61	0.67	103.39
		AM ADHD	0.47	0.31	144.32
		AM ADHD-C	0.56	0.52	169.22
LR	SBERT	ADHD	0.79	0.67	80.63
		ADHD-C	0.72	0.77	66.56
		AM ADHD	0.68	0.64	89.84
		AM ADHD-C	0.67	0.76	70.44
LR	TFIDF	ADHD	0.79	0.51	95.84
		ADHD-C	**0.89**	**0.52**	**67.33**
		AM ADHD	0.79	0.56	84.89
		AM ADHD-C	0.78	0.60	64.17
NB	TFIDF	ADHD	0.79	0.45	98.05
		ADHD-C	0.83	0.47	74.06
		AM ADHD	0.79	0.55	85.42
		AM ADHD-C	0.83	0.58	61.28
RF	Doc2Vec	ADHD	0.63	0.61	109.37
		ADHD-C	0.67	0.76	74.61
		AM ADHD	0.58	0.35	138.68
		AM ADHD-C	0.50	0.29	123.67
RF	SBERT	ADHD	0.79	0.69	98.26
		ADHD-C	0.72	0.77	59.33
		AM ADHD	0.74	0.65	96.53
		AM ADHD-C	0.78	0.77	66.56
RF	TFIDF	ADHD	0.58	0.09	160.00
		ADHD-C	0.61	0.32	121.00
		AM ADHD	0.58	0.42	126.53
		AM ADHD-C	0.78	0.68	80.72
SVM	Doc2Vec	ADHD	0.53	0.48	124.42
		ADHD-C	0.61	0.65	105.89
		AM ADHD	0.58	0.21	182.00
		AM ADHD-C	0.50	0.50	138.83
SVM	SBERT	ADHD	0.79	0.39	120.95
		ADHD-C	0.72	0.78	79.56
		AM ADHD	0.63	0.47	119.11
		AM ADHD-C	0.56	0.70	81.11
SVM	TFIDF	ADHD	0.79	0.26	121.68
		ADHD-C	0.83	0.27	95.06
		AM ADHD	0.84	0.44	97.68
		AM ADHD-C	0.83	0.54	73.39

Table 3.12: Performance results on the ADHD dataset

Model		Dataset	Metrics		
Classifier	**Features Extractor**		**Recall@10**	**WSS@95**	**ATD**
LR	Doc2Vec	OP	0.82	0.76	110.64
		OP-C	0.69	0.24	328.08
		OP-Margot	0.91	0.78	78.36
		OP-C-Margot	0.69	0.54	225.15
LR	SBERT	OP	0.64	0.65	128.27
		OP-C	0.62	0.66	169.62
		AM OP	0.64	0.72	100.55
		OP-C-Margot	0.46	0.64	219.15
LR	TFIDF	OP	1.0	0.78	60.91
		OP-C	0.85	0.32	264.62
		AM OP	0.82	0.73	78.27
		OP-C-Margot	0.69	0.47	241.85
NB	TFIDF	OP	0.82	0.75	59.00
		OP-C	0.69	0.44	242.69
		AM OP	0.82	0.74	66.00
		OP-C-Margot	0.69	0.53	236.62
RF	Doc2Vec	OP	0.73	0.73	93.55
		OP-C	0.69	0.50	220.15
		OP-Margot	**1.0**	**0.82**	**46.64**
		OP-C-Margot	0.54	0.63	217.46
RF	SBERT	OP	0.73	0.71	93.27
		OP-C	0.54	0.60	234.38
		AM OP	0.73	0.75	94.45
		OP-C-Margot	0.54	0.60	235.92
RF	TFIDF	OP	0.82	0.71	95.73
		OP-C	0.85	0.55	201.00
		AM OP	0.73	0.74	91.09
		OP-C-AM	0.69	0.56	247.69
SVM	Doc2Vec	OP	0.82	0.74	104.73
		OP-C	0.77	0.62	174.23
		AM OP	0.82	0.74	105.18
		OP-C-Margot	0.54	0.58	245.62
SVM	SBERT	OP	0.82	0.64	110.82
		OP-C	0.54	0.62	197.08
		AM OP	0.73	0.76	72.91
		OP-C-Margot	0.62	0.65	172.08
SVM	TFIDF	OP	1.0	0.77	63.36
		OP-C	0.85	0.26	255.15
		AM OP	0.82	0.74	93.45
		OP-C-Margot	0.69	0.52	250.69

Table 3.13: Performance results on the Opioids Cleaned dataset

4. Results Overview

This section presents a comparative analysis on the performance. Figure 4.1 shows the results that led to the most efficient systematic review for each dataset. The data ending with '-C' refer to data that have been pre-processed by removing relevant records that contain neither the title nor the abstract field.

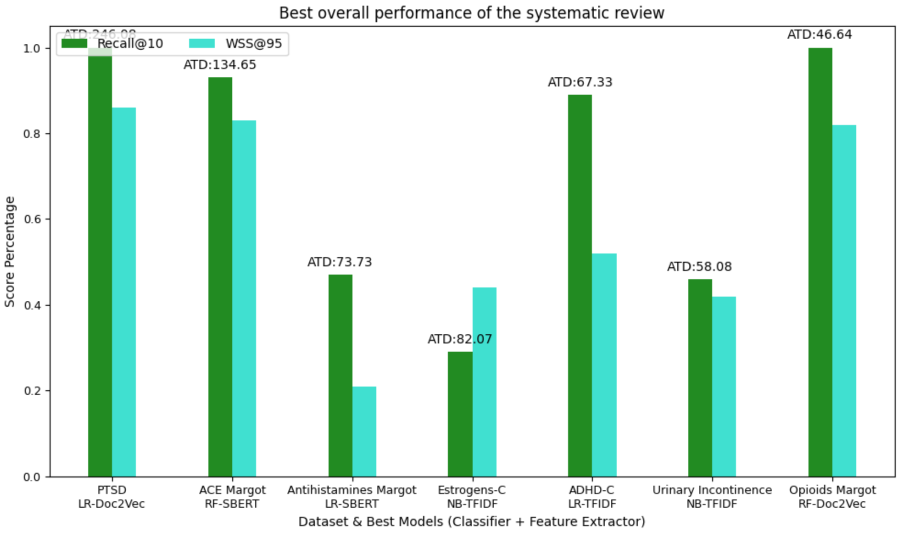


Figure 4.1: Best overall performance of the systematic review for each dataset
This table displays the best performance in terms of Recall@10 and WSS@95 for AM (MARGOT) data and full-abstracts data

Table 4.1 provides some insights about the size, the percentage of relevant records, missing abstracts and the scores obtained with the best models. The inverse correlation between the percentage of relevant records and the Recall@10 is given by the fact that the Recall is measured at 10% of screened records. Therefore, the higher the percentage of relevant records, the more difficult it is for all relevant records to be screened before 10%. For instance, let me consider Estrogens where the number of relevant record is $0.21 * 365 \approx 77$. As the recall is measured at 10%, the maximum number of screened study at 10% is $365 * 0.10 = 36.5$. Since I am looking for the highest possible Recall@10 score, let me assume that all relevant records (True Positives) are all correctly identified in the first 36 records checked. Hence, the maximum recall after screening 10% of the record is $36/77 \approx 0.47$. Instead, the WSS@95 is a measure of sampling over the normal rate (random screening) [49]. Thus, this percentage has no influence on the relevance of the records. In any case, a positive value means that a certain amount of time is saved compared to manual screening.

Table 4.1: Characteristics of the data and their best overall performance

This table shows the dataset that obtained the best overall performance with their characteristics. These are: PTSD, Argument Mined ACE, Argument Mined Antihistamines, Estrogens, ADHD, Urinary Incontinence and Argument Mined Opioids

Data	Size	Relevant Records	Missing Abs.	Recall@10	WSS@95	Models
PTSD	5782	0.66%	12.92%	1.00	0.86	LR-Doc2Vec
AM-ACE	2544	1.61%	14.19%	0.93	0.83	RF-SBERT
AM-Ant.	310	5.16%	12.26%	0.47	0.21	LR-SBERT
Estrogens	365	21.73%	7.67%	0.29	0.44	NB-TFIDF
ADHD	850	2.24%	8.12%	0.89	0.52	LR-TFIDF
Urinary Inc.	327	12.23 %	17.12%	0.46	0.42	NB-TFIDF
AM-Opioids	1680	0.71%	0%	1.00	0.82	RF-Doc2Vec

Table 4.2 compares the difference in performance metrics (Recall@10 and WSS@95) between both the original and the AM data processed by MARGOT for each combination of classifier and feature extractor. Considering that in the multiple model template there are 10 different combinations of models, let me denote a specific combination of models as M_i and $Score(M_i)$ as both the Recall@10 and the WSS@95 metrics. The following formula gives a representation of the difference in performance between the AM data and the full-abstracts ones, normalized by nd which are the number of datasets under analysis, 7 in this case.

$$Score(M_i) = \frac{\sum_{n=1}^{nd} ScoreAMData_n(M_i) - ScoreOriginalData_n(M_i)}{nd}$$

A positive Recall@10 or WSS@95 value indicates that MARGOT has achieved better performance, while a negative value indicates a decrease in performance compared to the original dataset.

MARGOT demonstrates its ability to improve the performance by increasing, on average, the Recall@10 when employing LR-SBERT. Furthermore, the outcomes reveal that MARGOT

Model		Recall@10		WSS@95	
Classifier	**Features Extractor**	**Mean**	**STD**	**Mean**	**STD**
LR	Doc2Vec	-0.026	0.085	**0.024**	0.108
LR	SBERT	**0.019**	0.045	**0.016**	0.035
LR	TFIDF	-0.063	0.080	0	0.118
NB	TFIDF	-0.043	0.048	-0.004	0.065
RF	Doc2Vec	-0.02	0.152	-0.051	0.197
RF	SBERT	-0.034	0.075	**0.054**	0.084
RF	TFIDF	-0.054	0.155	-0.006	0.169
SVM	Doc2Vec	-0.077	0.103	**0.013**	0.151
SVM	SBERT	-0.106	0.065	**0.026**	0.083
SVM	TFIDF	-0.0186	0.144	**0.001**	0.135

Table 4.2: Differences in performance for the models between AM and original data
This table highlights the performance disparities between various models using Recall@10 and WSS@95 metrics, comparing the AM (MARGOT) data with the full-abstracts data.

consistently leads to improvements across multiple models in terms of WSS@95, underlining the positive impact of using MARGOT.

It can be observed that, when the difference between the absolute mean and the standard deviation, $abs(mean(Recall@10)) - std(Recall@10)$ and $abs(mean(WSS@95)) - std(WSS@95)$, is negative for certain model combinations, there are datasets where the optimal score is achieved when using the AM data, as well as datasets where the best score is obtained with the original data. This phenomenon occurs in almost all models (except SVM-SBERT when looking at the Recall@10) and highlights the interdependence between models and data, emphasizing the significance of selecting appropriate data for optimal performance.
Similarly, with the aim of comparing the difference in performance in term of ATD, the following formula is used:

$$ATD(M_i) = \frac{\sum_{n=1}^{nd} ATDOriginalData_n(M_i) - ATDAMData_n(M_i)}{nd}$$

If the outcome is a positive value, it indicates that MARGOT is faster in the process of systematic review as fewer records need to be screened on average in order to find all the relevant records in a dataset n with a combination models (classifier and feature extractor) M_i.

Model		**ATD**	
Classifier	**Features Extractor**	**Mean**	**STD**
LR	Doc2Vec	**1.79**	35.978
LR	SBERT	**4.007**	19.764
LR	TFIDF	-14.191	16.758
NB	TFIDF	-15.643	22.703
RF	Doc2Vec	**2.621**	38.245
RF	SBERT	-4.483	34.918
RF	TFIDF	-24.659	46.683
SVM	Doc2Vec	-10.353	25.635
SVM	SBERT	**-2.539**	30.732
SVM	TFIDF	-24.279	36.575

Table 4.3: Comparison of the difference in performance in terms of ATD metric for different models between the full-abstracts and the AM data.

Table 4.3 shows that MARGOT is averagely faster in some models. However, for both the versions of the data the standard deviation deviates significantly from the mean. Therefore, it implies a greater variability and dispersion in the data. For all the combination of models, the subtraction of the absolute value of the mean from the standard deviation - $abs(mean(ATD)) - std(ATD)$ - gives always a negative outcome. This means that there are some datasets in which the version with the AM abstracts is faster compared to the original one and vice versa.

5. Check the Stability of the Results Obtained

The last part of this analysis consists of verifying the stability of the results obtained in the previous simulations. In this chapter, the results of 15 further simulations performed for each dataset on the Cloud platform are presented. The evaluation of the results is based on the calculation of the mean and standard deviation of Recall@10 and WSS@95 of these 15 simulations and their comparison with the original results obtained in the first simulation. I do not consider the ATD because its values vary significantly between the simulations. Most importantly, there is no significant difference in this metric between AM and full-abstracts data.

The values in the tables in the following sections are indicative of both performance (Recall@10 and WSS@95) and stability (Mean and Std). Theoretically, to have a perfectly stable model, its standard deviation resulting from these then simulations must be zero. So, the average of the scores must be equal to the ones obtained during the first simulation study.

5.1 Stability of PTSD

Table 5.1 shows the comparison between the results from the previous simulation conducted on the PTSD data and the mean and the standard deviation derived from the 15 simulations on the same dataset. The same comparison is also made for the AM version of the data.

- All the models show slight variations in performance for both versions of the data.
- Although in the first simulation the best performance was given by LR-Doc2Vec, now on average, the LR-TFIDF model is the best model and NB-TFIDF is the second best. The outcomes are almost equal, but LR-TFIDF is slightly more stable in terms of Recall@10. These best outcomes are all obtained with the full-abstracts data.
- LR with SBERT generally improves scores with AM data. However, in this case, the mean of Recall@10 decreases slightly, suggesting a potential instability in this pattern.

Model		Dataset	Recall@10			WSS@95		
Classifier	**Features Extractor**		**Recall**	**Mean**	**Std**	**WSS**	**Mean**	**Std**
LR	Doc2Vec	PTSD	1.0	0.98	0.032	0.86	0.85	0.012
		AM-PTSD	0.89	0.90	0.013	0.82	0.81	0.008
LR	SBERT	PTSD	0.95	0.95	0.011	0.86	0.87	0.009
		AM-PTSD	0.97	0.94	0.032	0.83	0.83	0.01
LR	TFIDF	**PTSD**	**0.97**	**0.98**	**0.009**	**0.90**	**0.90**	**0.003**
		AM-PTSD	0.97	0.96	0.013	0.87	0.88	0.004
NB	TFIDF	PTSD	0.97	0.98	0.011	0.90	0.90	0.002
		AM-PTSD	0.97	0.97	0.01	0.85	0.86	0.005
RF	Doc2Vec	PTSD	0.97	0.97	0.015	0.86	0.86	0.012
		AM-PTSD	0.95	0.91	0.032	0.85	0.82	0.014
RF	SBERT	PTSD	0.97	0.88	0.236	0.86	0.84	0.025
		AM-PTSD	0.84	0.81	0.157	0.82	0.79	0.028
RF	TFIDF	PTSD	0.97	0.97	0.014	0.88	0.88	0.008
		AM-PTSD	0.84	0.87	0.035	0.77	0.79	0.014
SVM	Doc2Vec	PTSD	0.95	0.98	0.024	0.84	0.84	0.008
		AM-PTSD	0.86	0.86	0.025	0.80	0.79	0.008
SVM	SBERT	PTSD	0.92	0.94	0.015	0.83	0.85	0.017
		AM-PTSD	0.84	0.89	0.020	0.78	0.81	0.012
SVM	TFIDF	PTSD	0.97	0.98	0.011	0.89	0.89	0.004
		AM-PTSD	0.95	0.97	0.011	0.86	0.86	0.005

Table 5.1: Analysis of model stability on PTSD
The *Recall* and *WSS* columns stand for the result obtained from the initial simulation in terms of Recall@10 and WSS@95, while the *mean* and *standard deviation* are derived to verify the stability of Recall@10 and WSS@95 based on 15 simulations.

The most unstable model is RF-SBERT as it has the highest standard deviation for the Recall@10. Yet, across all models and for both data versions, the standard deviation is generally low. This suggests minimal obscillations in Recall@10 and WSS@95 across simulations. Figure 5.1 illustrates the stability of the best model across the 15 simulations.

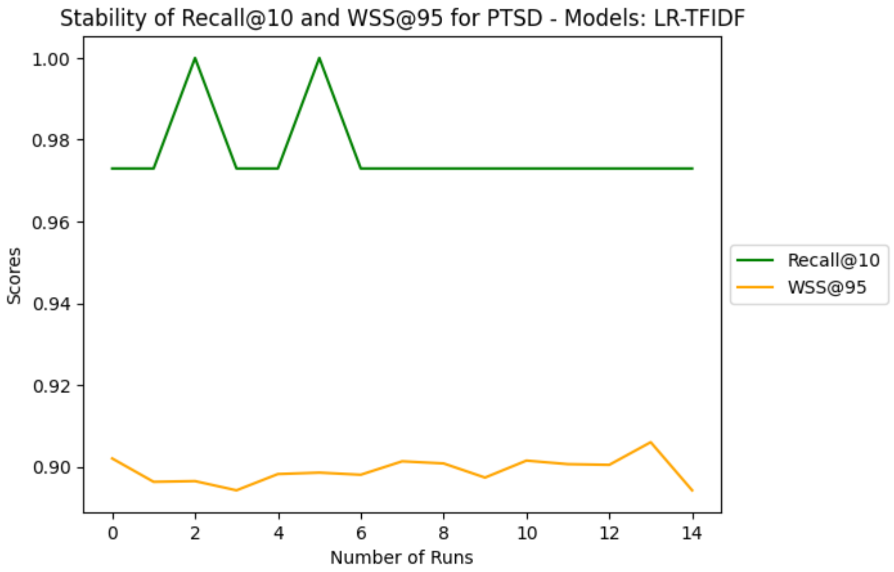


Figure 5.1: Stability of the best model of PTSD
This figure illustrates the stability of the best model, NB-TFIDF, for the full-abstracts version of PTSD across 15 simulations

5.2 Stability of ACE

Table 5.2 shows the comparison between the results from the previous simulation conducted on both the versions of the ACE data and the mean and the standard deviation derived from the 15 simulations on the same datasets. For ACE:

- No model exhibits perfect stability.
- The overall best model (RF-SBERT) and the second-best model (LR-SBERT) still have the highest average scores.
- SVM-Doc2Vec with the AM version of the data performs considerably better on average in terms of Recall@10 compared to the initial simulation. However, it has the highest standard deviation value, indicating greater variability across the simulations. This difference depends on the selection of the training data as, different seeds selects different prior knowledge.

Figure 5.2 shows the stability of the best model across simulations, revealing fluctuations between runs, especially in terms of Recall@10. Although on average this is the best model, the Recall@10 metric is not perfectly stable. In contrast, the WSS@95 metric has greater stability. The same scenario applies to the second best model.

Model		Dataset	Recall@10			WSS@95		
Classifier	**Features Extractor**		**Recall**	**Mean**	**Std**	**WSS**	**Mean**	**Std**
LR	Doc2Vec	ACE	0.80	0.76	0.028	0.62	0.65	0.037
		AM-ACE	0.75	0.70	0.042	0.74	0.74	0.006
LR	SBERT	ACE	0.88	0.87	0.019	0.79	0.78	0.009
		AM-ACE	0.90	0.89	0.046	0.82	0.83	0.013
LR	TFIDF	ACE	0.88	0.86	0.015	0.80	0.78	0.017
		AM-ACE	0.80	0.82	0.022	0.70	0.72	0.012
NB	TFIDF	ACE	0.88	0.88	0.012	0.78	0.78	0.007
		AM-ACE	0.83	0.83	0.014	0.72	0.72	0.007
RF	Doc2Vec	ACE	0.83	0.82	0.028	0.57	0.55	0.123
		AM-ACE	0.88	0.86	0.014	0.70	0.75	0.041
RF	SBERT	ACE	0.90	0.87	0.023	0.76	0.78	0.019
		AM-ACE	**0.93**	**0.90**	**0.044**	**0.83**	**0.83**	**0.012**
RF	TFIDF	ACE	0.83	0.86	0.033	0.76	0.78	0.025
		AM-ACE	0.83	0.78	0.029	0.65	0.67	0.052
SVM	Doc2Vec	ACE	0.73	0.71	0.024	0.65	0.66	0.047
		AM-ACE	0.45	0.69	0.022	0.69	0.71	0.027
SVM	SBERT	ACE	0.85	0.84	0.024	0.75	0.74	0.014
		AM-ACE	0.85	0.86	0.025	0.78	0.79	0.024
SVM	TFIDF	ACE	0.90	0.89	0.02	0.75	0.78	0.021
		AM-ACE	0.85	0.82	0.016	0.64	0.65	0.032

Table 5.2: Analysis of model stability on ACE
The Recall and WSS columns stand for the result obtained from the initial simulation in terms of Recall@10 and WSS@95, while the mean and standard deviation are derived to verify the stability of Recall@10 and WSS@95 based on 15 simulations

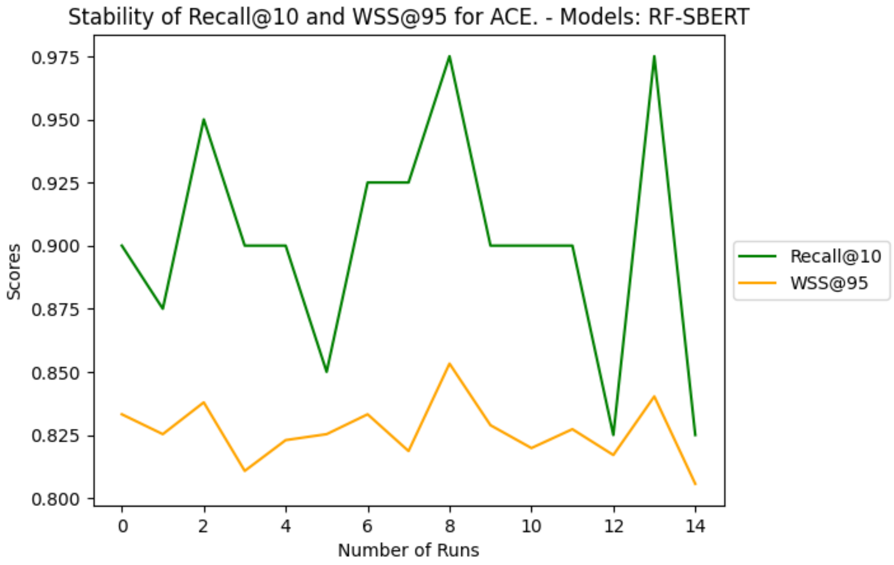


Figure 5.2: Stability of the best model of AM-ACE
This figure illustrates the stability of the best model, RF-SBERT, for the AM version of ACE across 15 simulations

5.3 Stability of Antihistamines

Table 5.3 shows the outcomes regarding the models' stability for Antihistamines.

- This dataset generally has very low scores. The standard deviation values from the 15 simulations are quite high. Moreover, the average Recall@10 scores are almost always significantly lower than the ones obtained during the first simulation. This stresses the instability of the models.
- Considering that this dataset has about 5% relevant records, it could potentially achieve a Recall@10 score of 1.0. However, it reaches only about 20% on average. In some cases, after about 10% of screened records, it fails to find any relevant records (Figure 5.3).
- The WSS@95 score shows more stability throughout the simulations.
- Due to the minimal differences between the models, it is challenging to determine if the data processed by MARGOT (AM data) performs better on this dataset or not. However, although the best model from the first simulation was LR-SBERT, better performance is now achieved on average by using RF-SBERT on the AM data. The second best combination is LR-SBERT on the full-abstracts version of the data.

Figure 5.3 illustrates the performance of the best model during the 15 simulations. Due to the different prior knowledge selected across runs, noticeable fluctuations can be seen between the recall values, with model performance reaching 50% in some simulations and dropping to 0% in others.

Model		Dataset	Recall@10			WSS@95		
Classifier	**Features Extractor**		**Recall**	**Mean**	**Std**	**WSS**	**Mean**	**Std**
LR	Doc2Vec	ANT	0.27	0.16	0.114	0.15	0.24	0.057
		AM-ANT	0.13	0.12	0.11	0.12	0.11	0.037
LR	SBERT	ANT	0.40	0.24	0.179	0.22	0.18	0.047
		AM-ANT	0.47	0.17	0.161	0.21	0.27	0.048
LR	TFIDF	ANT	0.33	0.19	0.165	0.10	0.18	0.107
		AM-ANT	0.40	0.19	0.144	0.33	0.24	0.063
NB	TFIDF	ANT	0.47	0.22	0.149	0.14	0.20	0.147
		AM-ANT	0.40	0.23	0.168	0.20	0.24	0.040
RF	Doc2Vec	ANT	0.33	0.16	0.15	0.21	0.30	0.078
		AM-ANT	0.20	0.12	0.132	0.19	0.17	0.079
RF	SBERT	ANT	0.33	0.18	0.161	0.11	0.23	0.050
		AM-ANT	**0.27**	**0.24**	**0.18**	**0.29**	**0.28**	**0.054**
RF	TFIDF	ANT	0.40	0.21	0.134	0.13	0.21	0.084
		AM-ANT	0.07	0.11	0.09	0.02	0.07	0.10
SVM	Doc2Vec	ANT	0.27	0.13	0.114	0.33	0.31	0.044
		AM-ANT	0.20	0.11	0.108	0.20	0.10	0.045
SVM	SBERT	ANT	0.40	0.20	0.153	0.22	0.23	0.072
		AM-ANT	0.20	0.20	0.155	0.33	0.29	0.040
SVM	TFIDF	ANT	0.40	0.17	0.119	0.18	0.25	0.099
		AM-ANT	0.33	0.18	0.145	0.26	0.27	0.046

Table 5.3: Analysis of model stability on Antihistamines
The *Recall* and *WSS* columns stand for the result obtained from the initial simulation in terms of Recall@10 and WSS@95, while the *mean* and *standard deviation* are derived to verify the stability of Recall@10 and WSS@95 based on 15 simulations.

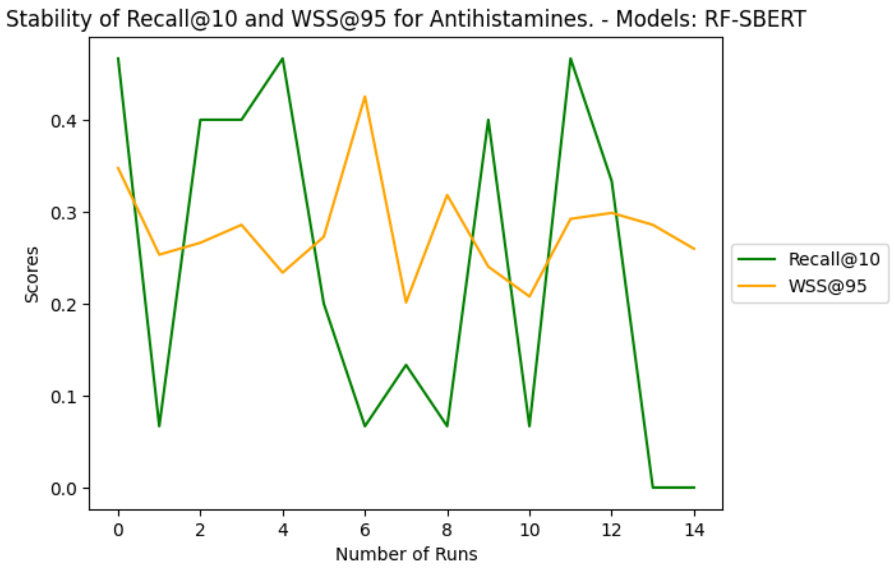


Figure 5.3: Stability of the best model of AM-Antihistamines
This figure illustrates the sability of the best model, RF-SBERT, for the AM version of Antihistamines across 15 simulations

5.4 Stability of Urinary Incontinence

Considering the fact that the maximum Recall@10 level of Urinary Incontinence could be ≈ 59% this dataset achieves fairly high performance.

- On average, NB-TFIDF is the top-perfroming model and the second best is LR-TFIDF.
- The most unstable model is RF-SBERT with the full-abstracts data but in any case the standard deviation is not too high.

Urinary Incontinence performs very well in terms of both Recall@10 and WSS@95 and the results are quite stable across runs, even if some seeds lead to a deterioration in performance. Figure 5.4 shows the stability across runs of the top-performing model, i.e NB-TFIDF with the full-abstracts version of Urinary Incontinence.

Model		Dataset	Recall@10			WSS@95		
Classifier	**Features Extractor**		**Recall**	**Mean**	**Std**	**WSS**	**Mean**	**Std**
LR	Doc2Vec	UR	0.23	0.22	0.046	0.32	0.29	0.09
		UR-Margot	0.28	0.22	0.053	0.48	0.42	0.040
LR	SBERT	UR	0.13	0.22	0.093	0.36	0.37	0.047
		UR-Margot	0.21	0.22	0.084	0.40	0.40	0.043
LR	TFIDF	UR	0.44	0.42	0.046	0.46	0.49	0.026
		UR-Margot	0.38	0.35	0.069	0.43	0.42	0.021
NB	TFIDF	**UR**	**0.46**	**0.46**	**0.041**	**0.42**	**0.41**	**0.035**
		UR-Margot	0.41	0.40	0.062	0.38	0.41	0.013
RF	Doc2Vec	UR	0.28	0.30	0.101	0.44	0.42	0.024
		UR-Margot	0.28	0.31	0.125	0.42	0.43	0.032
RF	SBERT	UR	0.28	0.27	0.108	0.31	0.37	0.062
		UR-Margot	0.15	0.21	0.081	0.46	0.38	0.041
RF	TFIDF	UR	0.38	0.41	0.055	0.46	0.39	0.077
		UR-Margot	0.41	0.36	0.057	0.42	0.45	0.024
SVM	Doc2Vec	UR	0.23	0.19	0.041	0.16	0.27	0.033
		UR-Margot	0.21	0.20	0.050	0.46	0.40	0.032
SVM	SBERT	UR	0.21	0.19	0.077	0.36	0.33	0.059
		UR-Margot	0.08	0.20	0.062	0.45	0.42	0.049
SVM	TFIDF	UR	0.10	0.38	0.040	0.48	0.44	0.038
		UR-Margot	0.38	0.33	0.073	0.43	0.43	0.029

Table 5.4: Analysis of model stability on Urinary Incontinence
The *Recall* and *WSS* columns stand for the result obtained from the initial simulation in terms of Recall@10 and WSS@95, while the *mean* and *standard deviation* are derived to verify the stability of Recall@10 and WSS@95 based on 15 simulations.

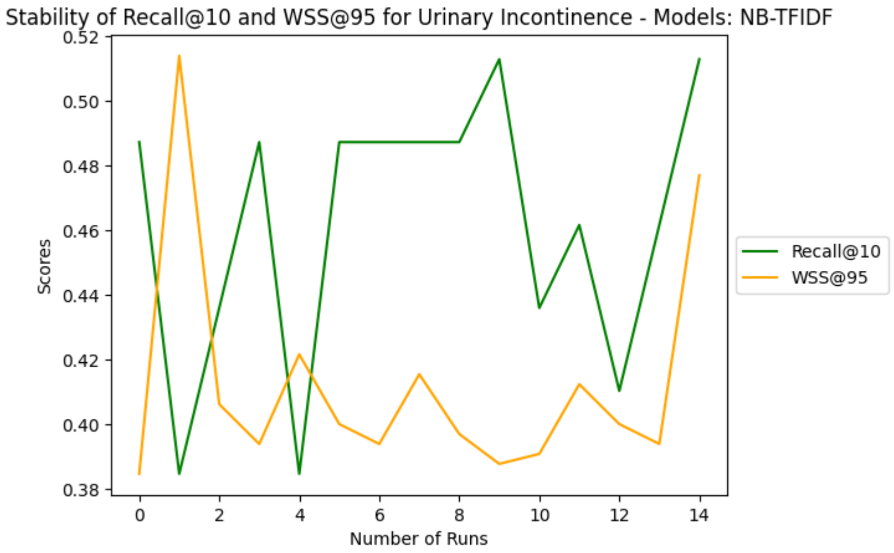


Figure 5.4: Stability of the best model of Urinary Incontinence
This figure illustrates the stability of the best model, NB-TFIDF, for the full-abstracts version of Urinary Incontinence across 15 simulations

5.5 Stability of Opioids

To check stability, I consider the best performing version of this dataset: the one where the mislabelled relevant records as well as the missing abstracts and the records with the negative ERF score are removed.

- RF-Doc2Vec is top performing model with the AM data and it remains as such across simulations. The second best model is now SVM-TFIDF with the full-abstracts version of the data.
- The dataset shows excellent performance with very low values of standard deviation both in term of Recall@10 and WSS@95.

Model		Dataset	Recall@10			WSS@95		
Classifier	**Features Extractor**		**Recall**	**Mean**	**Std**	**WSS**	**Mean**	**Std**
LR	Doc2Vec	OP	0.82	0.78	0.056	0.76	0.72	0.025
		AM-OP	0.91	0.88	0.070	0.78	0.77	0.011
LR	SBERT	OP	0.64	0.69	0.093	0.65	0.67	0.034
		AM-OP	0.64	0.59	0.099	0.72	0.69	0.037
LR	TFIDF	OP	1.0	0.98	0.062	0.78	0.77	0.01
		AM-OP	0.82	0.83	0.031	0.73	0.73	0.023
NB	TFIDF	OP	0.82	0.84	0.062	0.75	0.76	0.02
		AM-OP	0.82	0.84	0.049	0.74	0.73	0.025
RF	Doc2Vec	OP	0.73	0.73	0.179	0.73	0.74	0.018
		AM-OP	**1.0**	**0.99**	**0.045**	**0.82**	**0.80**	**0.017**
RF	SBERT	OP	0.73	0.58	0.199	0.71	0.60	0.056
		AM-OP	0.73	0.67	0.159	0.75	0.67	0.061
RF	TFIDF	OP	0.82	0.88	0.052	0.71	0.77	0.023
		AM-OP	0.73	0.78	0.087	0.74	0.74	0.030
SVM	Doc2Vec	OP	0.82	0.78	0.056	0.74	0.74	0.018
		AM-OP	0.82	0.81	0.07	0.74	0.75	0.018
SVM	SBERT	OP	0.82	0.74	0.093	0.64	0.65	0.06
		AM-OP	0.73	0.77	0.056	0.76	0.73	0.029
SVM	TFIDF	OP	1.0	0.98	0.062	0.77	0.78	0.02
		AM-OP	0.82	0.84	0.059	0.74	0.76	0.016

Table 5.5: Analysis of model stability on Opioids
The *Recall* and *WSS* columns stand for the result obtained from the initial simulation in terms of Recall@10 and WSS@95, while the *mean* and *standard deviation* are derived to verify the stability of Recall@10 and WSS@95 based on 15 simulations.

Figure 5.5 displays the stability of the top-performing models combination. There's almost perfect stability in terms of Recall@10, with an exception during run 3, meaning that the selected prior knowledge might not be as meaningful as in the other runs.

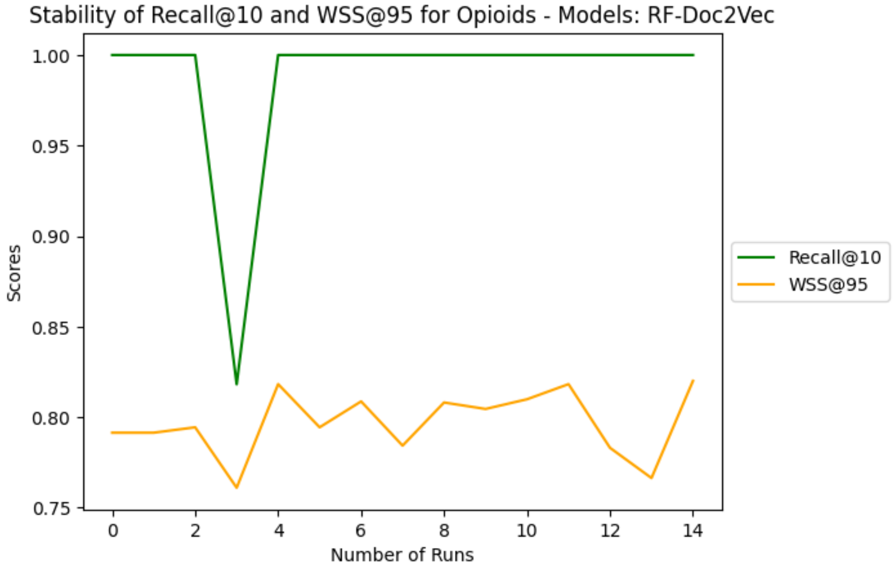


Figure 5.5: Stability of the best model of AM-Opioids
This figure illustrates the stability of the best model, RF-Doc2Vec, for the AM version of Opioids across 15 simulations

5.6 Stability of ADHD

Also for this dataset, the version considered is the one without wrongly labelled relevant records.

- The LR-TFIDF models performed the best on the original dataset but on average SVM-TFIDF perform slightly better and has more stability in terms of Recall@10.
- In general, this dataset is quite stable, but some models show more instability than others. For instance, all models using Doc2Vec as a feature extractor for both data versions, LR, SVM with SBERT and RF with TFIDF and SBERT for full-abstracts data are more unstable.
- Overall, the scores are highly satisfactory, and the model performance remained fairly stable during simulations.

SVM-TFIDF exhibited low standard deviation (especially in terms of Recall@10), confirming its stability during simulations, as shown in Figure 5.6.

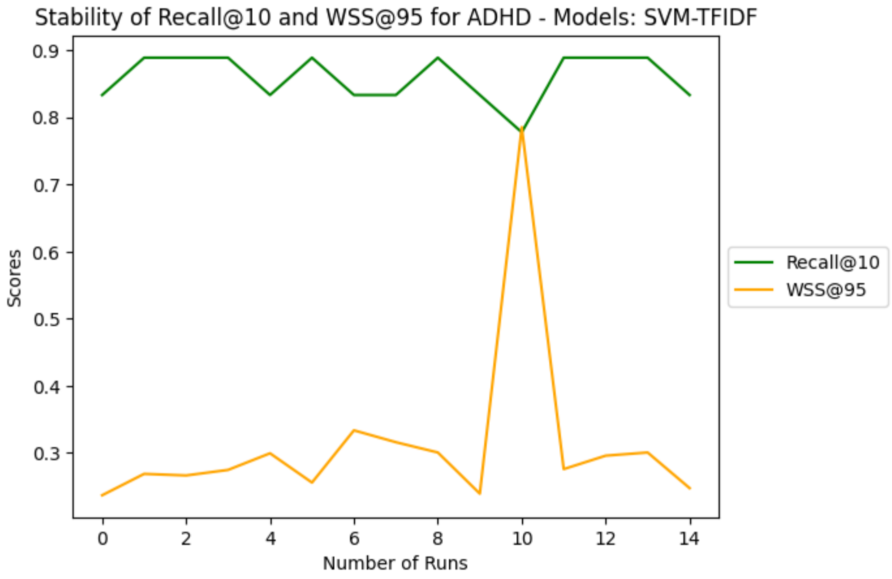


Figure 5.6: Stability of the best model of ADHD
This figure illustrates the stability of the best model, SVM-TFIDF, for the full-abstracts version of the ADHD dataset across 15 simulations

Model		Dataset	Recall@10			WSS@95		
Classifier	**Features Extractor**		**Recall**	**Mean**	**Std**	**WSS**	**Mean**	**Std**
LR	Doc2Vec	ADHD-C	0.61	0.57	0.172	0.67	0.66	0.036
		AM-ADHD-C	0.56	0.43	0.162	0.52	0.53	0.056
LR	SBERT	ADHD-C	0.72	0.77	0.106	0.77	0.77	0.025
		AM-ADHD-C	0.67	0.66	0.095	0.76	0.77	0.018
LR	TFIDF	ADHD-C	0.89	0.85	0.059	0.52	0.54	0.069
		AM-ADHD-C	0.78	0.80	0.044	0.60	0.60	0.053
NB	TFIDF	ADHD-C	0.83	0.84	0.025	0.47	0.48	0.076
		AM-ADHD-C	0.83	0.84	0.019	0.58	0.60	0.053
RF	Doc2Vec	ADHD-C	0.67	0.66	0.194	0.76	0.74	0.034
		AM-ADHD-C	0.50	0.51	0.139	0.29	0.47	0.088
RF	SBERT	ADHD-C	0.72	0.76	0.14	0.77	0.76	0.021
		AM-ADHD-C	0.78	0.82	0.059	0.77	0.78	0.027
RF	TFIDF	ADHD-C	0.61	0.48	0.121	0.32	0.37	0.132
		AM-ADHD-C	0.78	0.65	0.193	0.68	0.63	0.074
SVM	Doc2Vec	ADHD-C	0.61	0.58	0.152	0.65	0.65	0.032
		AM-ADHD-C	0.50	0.36	0.154	0.50	0.50	0.052
SVM	SBERT	ADHD-C	0.72	0.72	0.113	0.78	0.74	0.037
		AM-ADHD-C	0.56	0.64	0.078	0.70	0.74	0.052
SVM	TFIDF	**ADHD-C**	**0.83**	**0.86**	**0.034**	**0.27**	**0.31**	**0.129**
		AM-ADHD-C	0.83	0.84	0.049	0.54	0.49	0.086

Table 5.6: Analysis of model stability on ADHD
The Recall and WSS columns stand for the result obtained from the initial simulation in terms of Recall@10 and WSS@95, while the mean and standard deviation are derived to verify the stability of Recall@10 and WSS@95 based on 15 simulations.

5.7 Stability of Estrogens

Also for Estrogens, it is considered the version without wrongly labelled records. Considering the maximum level of Recall@10 attainable for this dataset is approximately 47%, I manage to achieve pretty good performance.

- The best model is NB-TFIDF and the second best is LR-TFIDF with the original data. Despite extremely similar performance, the best model combination is more stable in terms of both Recall@10 and WSS@95.

Figure 5.7 shows the stability of NB-TFIDF on the full-abstracts version of the data.

Model		Dataset	Recall@10			WSS@95		
Classifier	**Features Extractor**		**Recall**	**Mean**	**Std**	**WSS**	**Mean**	**Std**
LR	Doc2Vec	ES-C AM-ES-C	0.11 0.14	0.13 0.15	0.042 0.034	0.17 0.26	0.20 0.26	0.051 0.025
LR	SBERT	ES-C AM-ES-C	0.21 0.20	0.21 0.21	0.044 0.058	0.40 0.42	0.41 0.42	0.021 0.016
LR	TFIDF	ES-C AM-ES-C	0.26 0.18	0.28 0.23	0.040 0.036	0.43 0.33	0.42 0.37	0.030 0.030
NB	TFIDF	**ES-C** AM-ES-C	**0.29** 0.16	**0.28** 0.24	**0.037** 0.035	**0.44** 0.40	**0.43** 0.35	**0.013** 0.017
RF	Doc2Vec	ES-C AM-ES-C	0.26 0.12	0.20 0.19	0.050 0.037	0.35 0.29	0.30 0.26	0.039 0.051
RF	SBERT	ES-C AM-ES-C	0.21 0.20	0.23 0.23	0.028 0.036	0.52 0.50	0.49 0.47	0.024 0.032
RF	TFIDF	ES-C AM-ES-C	0.25 0.22	0.24 0.21	0.042 0.021	0.41 0.35	0.38 0.34	0.035 0.030
SVM	Doc2Vec	ES-C AM-ES-C	0.09 0.12	0.12 0.13	0.021 0.025	0.18 0.25	0.15 0.24	0.056 0.023
SVM	SBERT	ES-C AM-ES-C	0.20 0.12	0.18 0.17	0.032 0.042	0.40 0.36	0.35 0.35	0.041 0.026
SVM	TFIDF	ES-C AM-ES-C	0.22 0.13	0.25 0.23	0.035 0.037	0.45 0.33	0.40 0.36	0.038 0.025

Table 5.7: Analysis of model stability on Estrogens
The *Recall* and *WSS* columns stand for the result obtained from the initial simulation in terms of Recall@10 and WSS@95, while the *mean* and *standard deviation* are derived to verify the stability of Recall@10 and WSS@95 based on 15 simulations.

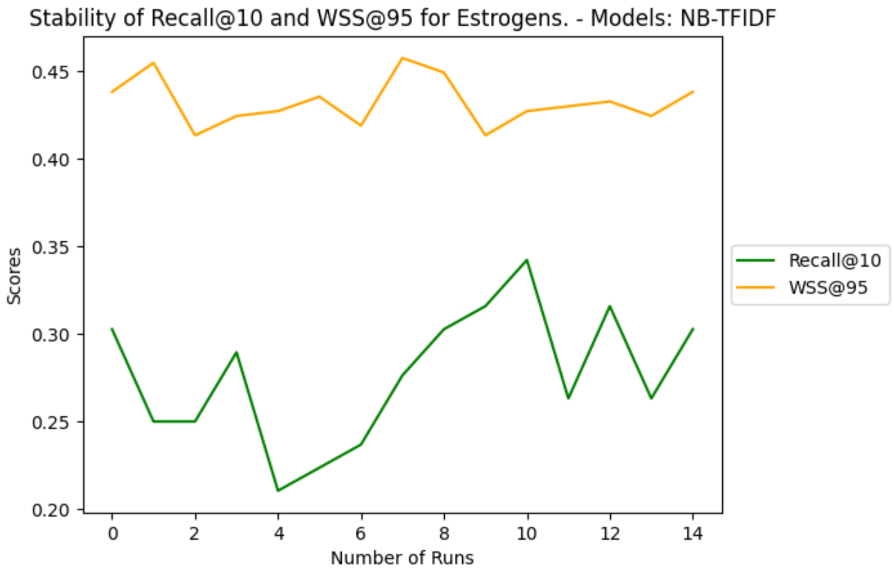


Figure 5.7: Stability of the best model of Estrogens
This figure illustrates the stability of the best model, NB-TFIDF, for the full-abstracts version of Estrogens across 15 simulations

6. Results Across the Simulations

Figure 6.1 shows, for each dataset, the best overall average performance across 15 simulations. It is worth to notice that MARGOT improves the overall performance for three different set of data, out of a total of seven.

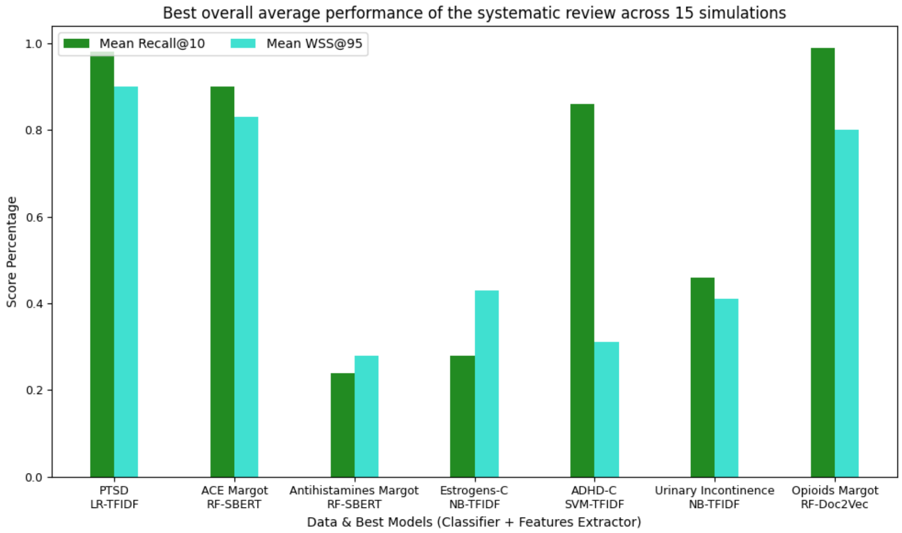


Figure 6.1: Best overall performance of the systematic review for each dataset

Table 6.1 gives a detailed analysis of the average performance of Recall@10 and WSS@95 for each model across all datasets based on the 15 simulations. The accompanying confidence intervals shed light on the stability and reliability of each model's performance. The model that uses LR as classifier and TFIDF as feature extractor achieves the highest Recall@10 when trained on the original data. This is the best performance across all models in terms of Recall@10. Meanwhile, Table 6.2 aggregates the results derived from the models presented in Table 6.1, showing their averaged outcomes (including also the ATD). This facilitates a more straightforward comparison between the overall performance derived from AM and full-abstracts data. On average, the use of full-abstracts is slightly better than that of AM abstracts in terms of Recall@10 (+ 4.77%), but for WSS@95 it is exactly the same.

Data	Model		Mean Metrics		Confidence Interval	
	Classifier	**Features Extractor**	**Recall@10**	**WSS@95**	**Recall@10**	**WSS@95**
Full-Abs.	LR	Doc2Vec	0.51	0.52	(0.45, 0.58)	(0.47, 0.57)
AM-Abs.			0.48	0.52	(0.42, 0.55)	(0.47, 0.57)
Full-Abs.	LR	SBERT	0.56	0.58	(0.50, 0.63)	(0.53, 0.62)
AM-Abs.			0.53	0.60	(0.46, 0.59)	(0.56, 0.64)
Full-Abs	LR	TFIDF	**0.65**	**0.58**	**(0.59, 0.71)**	**(0.54, 0.63)**
AM-Abs.			0.60	0.57	(0.54, 0.66)	(0.52, 0.61)
Full-Abs	NB	TFIDF	0.64	0.57	(0.58, 0.70)	(0.52, 0.61)
AM-Abs.			0.62	0.56	(0.56, 0.68)	(0.52, 0.60)
Full-Abs	RF	Doc2Vec	0.55	0.56	(0.48, 0.61)	(0.52, 0.60)
AM-Abs.			0.55	0.53	(0.49, 0.62)	(0.48, 0.58)
Full-Abs	RF	TFIDF	0.58	0.54	(0.52, 0.64)	(0.50, 0.59)
AM-Abs.			0.54	0.53	(0.48, 0.59)	(0.48, 0.57)
Full-Abs	RF	SBERT	0.54	0.58	(0.48, 0.60)	(0.54, 0.62)
AM-Abs.			0.55	0.60	(0.49, 0.62)	(0.56, 0.64)
Full-Abs	SVM	Doc2Vec	0.50	0.52	(0.43, 0.56)	(0.47, 0.57)
AM-Abs.			0.45	0.50	(0.39, 0.51)	(0.45, 0.55)
Full-Abs	SVM	SBERT	0.54	0.55	(0.48, 0.61)	(0.51, 0.60)
AM-Abs.			0.53	0.59	(0.47, 0.59)	(0.55, 0.63)
Full-Abs	SVM	TFIDF	0.64	0.55	(0.58, 0.71)	(0.50, 0.60)
AM-Abs.			0.60	0.54	(0.54, 0.66)	(0.50, 0.58)

Table 6.1: Performance metrics of the model combinations
This table is based on 15 simulations. It differentiates between 'Full-Abs' (full-abstracts data) and 'AM-Abs' (AM abstracts data). Metrics include the mean values of Recall@10 and WSS@95, alongside their respective confidence intervals for assessing stability and reliability.

Data	Recall@10	WSS@95	ATD.
Original	0.571	0.55	118.14
MARGOT	0.545	0.55	125.07

Table 6.2: Overall average of among models
This table shows the overall average among all models calculated from the 15 simulations, in terms of Recall, WSS@95 and ATD.

The minimal difference between AM abstracts and full-text abstracts, underscores their comparable efficacy when employed in systematic reviews using active learning. Although the information is approximately halved, MARGOT demonstrates its ability to retain only the relevant content. In addition, feature extractors like SBERT can only process a maximum of number of tokens (512) [46], thereby truncating any additional information. MARGOT mitigates this limitation by preserving the most pertinent information. Moreover, these findings demonstrate that the differences in performance between the models highlight the importance of selecting the right combination of classifier and feature extractors and, this choice, depends primarily on the data at hand. The Figures below show a detailed graphical overview of the obtained outcomes.

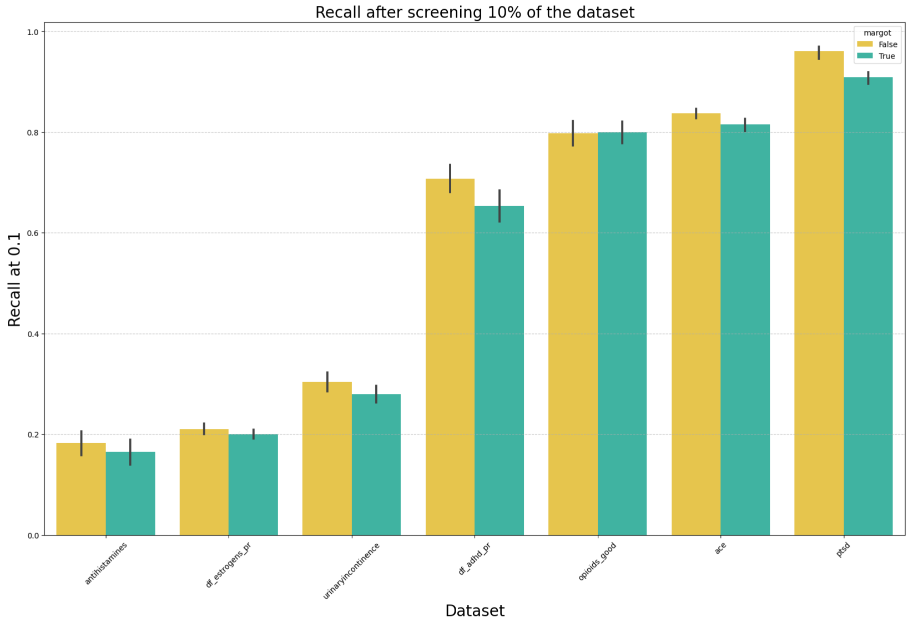


Figure 6.2: Average Recall@10 for each dataset across the 15 simulations

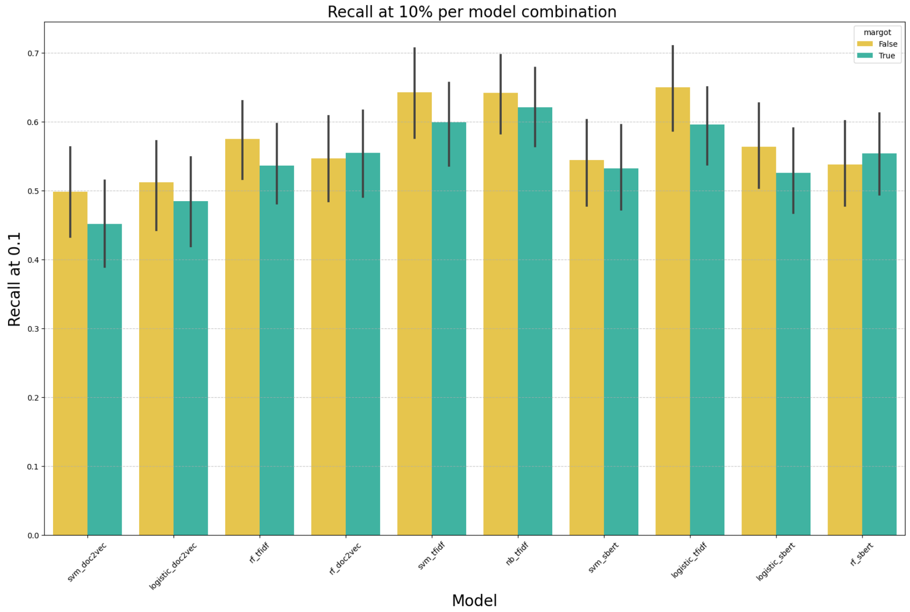


Figure 6.3: Average Recall@10 of each model for each dataset across the 15 simulations.

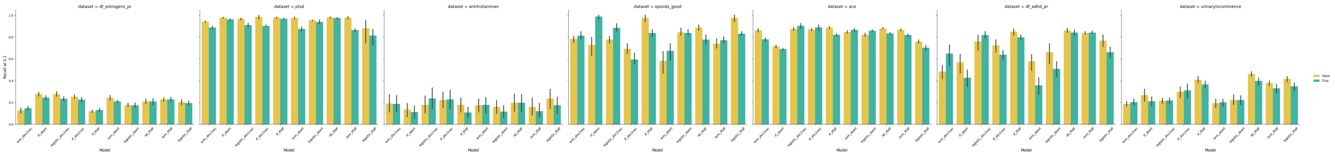


Figure 6.4: Average Recall@10 for each model of each dataset across the 15 simulations

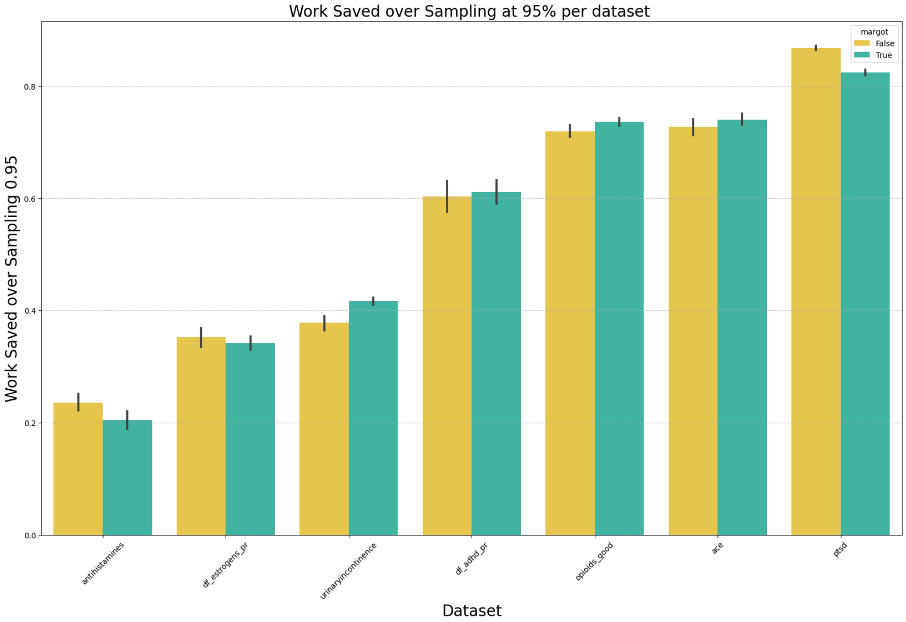


Figure 6.5: Average WSS@95 for each dataset across the 15 simulations

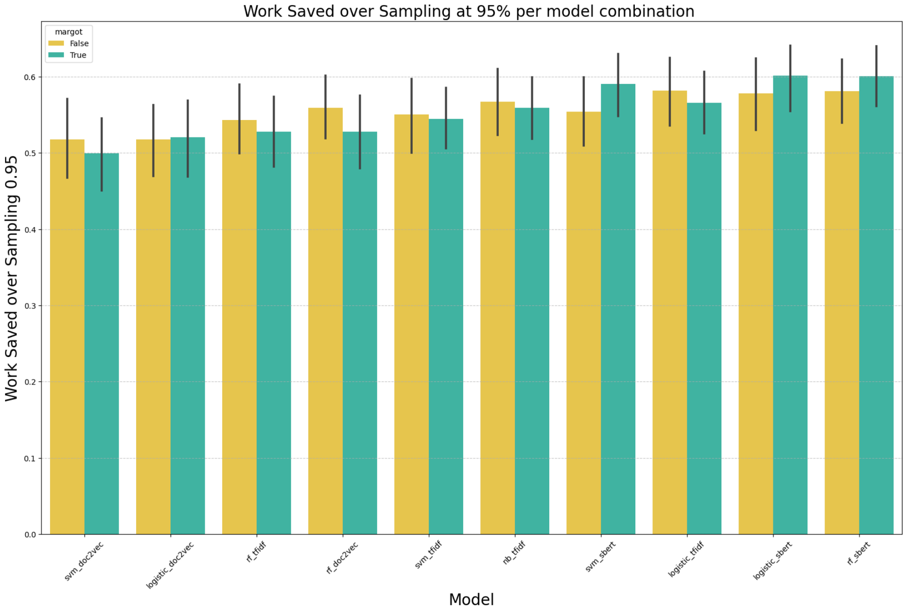


Figure 6.6: Average WSS@95 of each model for each dataset across the 15 simulations

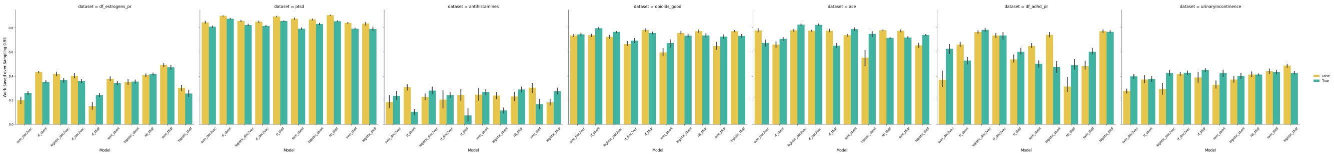


Figure 6.7: Average WSS@95 for each model of each dataset across the 15 simulations

6.1 Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank Test is applied to the Recall@10 values from the MARGOT data and the Recall@10 values from the original data, gathered over 15 simulations. The test is equally applied to the WSS@95.

A key assumption underlying this test is that the differences between paired observations are symmetrically distributed [54].

For the Recall@10 paired samples between MARGOT and the original data, the results are:

```
V = 115680, p-value < 2.2e-16 alternative hypothesis: true location shift
is not equal to 0
```

The V value represents the sum of the ranks of the positive differences between paired observations [57]. The further the V-value is from zero, the greater the divergence between the medians [54]. In this case, the p-value is essentially zero, thereby providing a strong evidence against the null hypothesis. In the context of the Wilcoxon Signed-Rank Test, the null hypothesis typically posits that the median difference between paired observations is zero, implying no significant distinction between the samples [58]. Therefore, the results suggest a significant median difference, or ”location shift,” between the two samples that is not equal to zero.

For the paired WSS@95 samples, the results were:

```
V = 263781, p-value = 0.3058 alternative
hypothesis: true location shift is not equal to 0
```

The comparatively high p-value in this test indicates that the distinction between the WSS@95 samples — whether using MARGOT or not — is not statistically meaningful, at least not at the conventional alpha level of 0.05. In conclusion, even if the alternative hypothesis for both tests is the same, the main distinctions lie in the data being tested and the results obtained. In summary, the first Wilcoxon Signed-Rank Test with the Recall@10 samples provides very strong evidence that there is a significant difference between the paired samples whereas the second test with the WSS@95 samples doesn’t provide enough evidence to say there’s a significant difference between the paired samples.

7. Discussion

The percentage of relevant records, as well as the size of the data, have no influence on the systematic reviews. On the other hand, the presence or absence of abstracts distinctly affected the outcomes. For instance, while Opioids exhibits enhanced performance upon eliminating missing abstracts, PTSD displays an opposite trend. This discrepancy may be topic-related as for certain topics titles may be more informative than abstracts and vice versa. Extension of this study could be consider also different topics, domains and delve into the missingness of abstracts. This dissertation has also shown that even with approximately halved data, the results remain commendably close to their full-text counterparts. Therefore, a major influence in the process of systematic review seems to be the representativeness of the text. That is, the content present in the text and the actual relevance and irrelevance of the records in the data. MARGOT have proven to be particularly good at capturing the most important information in the text. This result can have promising implications for large-scale data processing in systematic reviews. Hence, it would be beneficial to explore and evaluate the merits of AM processing on longer texts beyond traditional abstracts. To conclude, future research should focus on leveraging richer datasets to improve performance and further investigate the potential broad-scale replacement of traditional abstracts with their AM counterpart.

8. Conclusion

Systematic reviews play a key role in synthesising a large amount of literature. Through the integration of Argument Mining into systematic reviews using AL, the aim is to refine the screening process across different scenarios. This research analyses how the AM text affects the dynamics of ASReview LAB, an AL tool designed for systematic reviews. Furthermore, I investigated the potential of AM abstracts to replace traditional ones across various topics and domains, offering a multi-faceted perspective to this study. Argument Mining recognises the challenges of information loss and missing content, reinforcing the importance of optimal text representation and data quality. The results demonstrated that the inclusion of AM abstracts can positively influence AL systematic reviews, as they improved the overall performance in several scenarios. In a general overview, as Table 6.2 shows, full-text abstracts still maintain a slight edge. Yet, when directly compared across datasets, AM data outperform the original ones in three out of seven datasets. Thus, neither method dominates consistently in efficacy.

These findings emphasise the effectiveness of MARGOT in distilling pertinent information from abstracts. Also, considering that the content of each abstracts is approximately halved, remarkable results are obtained. It is worth observing that there's a statistically significant difference between the Recall@10 values of MARGOT and the original data. In contrast, there is no significant difference in terms of WSS@95 between the two pairs of samples. These distinctions further shed light on the different impacts of using MARGOT versus traditional data in different contexts.

Additionally, the performance of a systematic review appears to be influenced by multiple factors, from data characteristics to technical challenges, including missing data or inconsistent labels, as highlighted in [33]. While longer texts might introduce more noise, the observations made indicate that full-abstracts aren't necessarily always superior to their AM counterparts. Consequently, future performance gains are likely to be based on the acquisition of richer datasets, rather than trying to optimise them to reduce noise. I also explored the erroneous labelling of relevant records. Noticeable improvements were observed when removing mislabelled records from the data that had at least one. The impact of missing abstract fields was also explored: while PTSD faltered in the absence of missing abstracts, Opioids thrived. This divergence could be attributed to the inherent domain differences: abstracts may hold paramount significance in PTSD, whereas in Opioids, the titles might be of greater relevance. This finding highlights that, the complexities of labelling, the correlation between relevant study as well as text representation play a determining role in the reliability of the data. An added advantage of AM abstracts lies in their lower memory requirements, making them more suitable for certain simulations. In this specific context, integrating SBERT with the AM abstract vector within the 32 GB memory was feasible, whereas this was not achievable with the full-abstracts as, having $\approx$ 2 GB of memory for loading the vector and SBERT, caused some simulations to fail. This proves MARGOT's value, especially when working with BERT-based models that have token limitations. In conclusion, it would be an oversimplification to declare a definitive ”best” model as the models are highly data-dependent and, vice versa, the data are highly model-dependent. Moreover, the merit of AM abstracts is not confined to performance metrics alone, but present practical advantages. Their streamlined processing and scalability make them a viable option for researchers dealing with long text data. Hence, AM abstracts can potentially offer less dependence on full-text sources and simplify information management. Even if the universal replacement of full-text abstracts with AM abstracts isn't imminent, MARGOT has proven its effectiveness in certain scenarios, making the systematic review process more efficient.

Bibliography

[1] G. S and G. P, "Systematic reviews and meta-analysis: Understanding the best evidence in primary healthcare," *J Family Med Prim Care*, vol. 2, pp. 9–14, 2013. DOI: 10.4103/2249-4863.109934.

[2] U. LS., "Systematic reviews and meta-analyses.," *J Can Acad Child Adolesc Psychiatry.*, pp. 57–9, Feb. 2011.

[3] "Systematic reviews." (2023), [Online]. Available: https://utica.libguides.com/c.php?g=960363&p=6934092.

[4] J. J. Teijema, S. Seuren, D. Anadria, A. Bagheri, and R. van de Schoot, *Simulation-based active learning for systematic reviews: A systematic review of the literature*, 2023. DOI: 10.31234/osf.io/67zmt. [Online]. Available: https://doi.org/10.31234/osf.io/67zmt.

[5] S. Fortunato, C. T. Bergstrom, K. Börner, *et al.*, "Science of science," *Science*, vol. 359, no. 6379, eaao0185, 2018. DOI: 10.1126/science.aao0185. eprint: https://www.science.org/doi/pdf/10.1126/science.aao0185. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aao0185.

[6] "Treadwell's service for systematic, scoping and other reviews." (2023), [Online]. Available: https://libguides.massgeneral.org/c.php?g=651077&p=4565528#:~:text=Systematic%20reviews%20are%20work%20and,%2D18%20months%20(Source).

[7] K. Nguyen-Trung, A. K. Saeri, and S. Kaufman, "Applying chatgpt and ai-powered tools to accelerate evidence reviews.," *OSF Preprints*, Apr. 2023. DOI: 10.31219/osf.io/pcrqf.

[8] S. A. G. E. Van den Brand and R. van de Schoot, *A systematic review on studies evaluating the performance of active learning compared to human reading for systematic review data*, Sep. 2021. DOI: 10.17605/OSF.IO/T9HGM.

[9] R. van de Schoot, *Saving time and sanity: Using active learning for systematic reviews and meta-analyses*, Aug. 2023. DOI: 10.23668/psycharchives.13047. [Online]. Available: https://psycharchives.org/en/item/fe3f1dbb-7bc2-4e7a-b2ba-120000356c1c.

[10] "Systematic reviews in the health sciences." (2023), [Online]. Available: https://libguides.rutgers.edu/c.php?g=337288&p=2269575.

[11] G. Ferdinands, R. Schram, J. de Bruin, *et al.*, "Active learning for screening prioritization in systematic reviews - A simulation study," Center for Open Science, OSF Preprints w6qbg, Sep. 2020. DOI: 10.31219/osf.io/w6qbg. [Online]. Available: https://ideas.repec.org/p/osf/osfxxx/w6qbg.html.

[12] G. Ferdinands, R. Schram, J. de Bruin, *et al.*, "Performance of active learning models for screening prioritization in systematic reviews: A simulation study into the Average Time to Discover relevant records," *Systematic Reviews*, vol. 12, no. 1, p. 100, 2023. DOI: 10.1186/s13643-023-02257-7.

[13] "Active learning in machine learning [guide and examples]." (2023), [Online]. Available: https://www.v7labs.com/blog/active-learning-guide.

[14] R. Schoot, J. de Bruin, R. Schram, *et al.*, "An open source machine learning framework for efficient and transparent systematic reviews," *Nature Machine Intelligence*, vol. 3, pp. 1–9, Feb. 2021. DOI: 10.1038/s42256-020-00287-7.

[15] J. J. Teijema, L. Hofstee, M. Brouwer, *et al.*, "Active learning-based systematic reviewing using switching classification models: The case of the onset, maintenance, and relapse of depressive disorders," *Frontiers in Research Metrics and Analytics*, vol. 8, p. 1 178 181, 2023. DOI: 10.3389/frma.2023.1178181.

[16] J. J. Teijema, S. Seuren, D. Anadria, A. Bagheri, and R. van de Schoot. "Simulation-based active learning for systematic reviews: A systematic review of the literature." (Jun. 2023), [Online]. Available: https://doi.org/10.31234/osf.io/67zmt.

[17] I. Montani and M. Honnibal, *Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models*. [Online]. Available: https://prodi.gy/.

[18] M. R. Berthold, N. Cebron, F. Dill, *et al.*, "Knime - the konstanz information miner: Version 2.0 and beyond," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 26–31, Nov. 2009, ISSN: 1931-0145. DOI: 10.1145/1656274.1656280. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656280.

[19] S. Jarl, L. Aronsson, S. Rahrovani, and M. H. Chehreghani, "Active learning of driving scenario trajectories," *Engineering Applications of Artificial Intelligence*, vol. 113, p. 104 972, 2022, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2022.104972. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622001750.

[20] J. Polanin, T. Pigott, D. Espelage, and J. Grotpeter, "Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses," *Research Synthesis Methods*, vol. 10, May 2019. DOI: 10.1002/jrsm.1354.

[21] C. Hamel, M. Hersi, S. Kelly, *et al.*, "Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses," *BMC Medical Research Methodology*, vol. 21, p. 285, 2021. DOI: 10.1186/s12874-021-01451-2.

[22] "Asreview documentation." (2023), [Online]. Available: https://asreview.readthedocs.io/en/latest/.

[23] J. Teijema, R. Van de Schoot, G. Ferdinands, P. Lombaers, and J. De Bruin, *ASReview Makita: a workflow generator for simulation studies using the command line interface of ASReview LAB*. [Online]. Available: https://github.com/asreview/asreview-makita.

[24] J. Boetje and R. van de Schoot, "The safe procedure: A practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses," *PsyArXiv*, Jul. 2023. DOI: 10.31234/osf.io/c93gq.

[25] G. V. Cormack and M. R. Grossman, "Engineering quality and reliability in technology-assisted review," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '16, Pisa, Italy: Association for Computing Machinery, 2016, pp. 75–84, ISBN: 9781450340694. DOI: 10.1145/2911451.2911510. [Online]. Available: https://doi.org/10.1145/2911451.2911510.

[26] "Asreview discussions." (2023), [Online]. Available: https://github.com/asreview/asreview/discussions/557.

[27] M. Lippi and P. Torroni, "Argumentation mining," *ACM Transactions on Internet Technology*, vol. 16, pp. 1–25, Mar. 2016. DOI: 10.1145/2850417.

[28] M. Lippi and P. Torroni, "Margot: A web server for argumentation mining," *Expert Systems with Applications*, vol. 65, pp. 292–303, 2016, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2016.08.050. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417416304493.

[29] E. Aharoni, A. Polnarov, T. Lavee, *et al.*, "A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics," in *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 64–68. DOI: 10.3115/v1/W14-2109. [Online]. Available: https://aclanthology.org/W14-2109.

[30] M. Lippi and P. Torroni, "Context-independent claim detection for argument mining," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15, Buenos Aires, Argentina: AAAI Press, 2015, pp. 185–191, ISBN: 9781577357384.

[31] A. Moschitti, "State-of-the-art kernels for natural language processing," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, p. 2. [Online]. Available: https://aclanthology.org/P12-4002.

[32] M. Lippi, F. Antici, G. Brambilla, *et al.*, "Amica: An argumentative search engine for covid-19 literature," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, vol. 22, 2022.

[33] R. Garud. "Simulation study on ensemble classifiers." (), [Online]. Available: https://github.com/asreview/asreview/discussions/1334.

[34] J. De Bruin, Y. Ma, G. Ferdinands, J. Teijema, and R. Van de Schoot, *SYNERGY - Open machine learning dataset on study selection in systematic reviews*, version V1, 2023. DOI: 10.34894/HE6NAQ. [Online]. Available: https://doi.org/10.34894/HE6NAQ.

[35] R. e. a. van de Schoot, "A systematic review of bayesian articles in psychology: The last 25 years.," *Psychological methods vol. 22,2 (2017): 217-239.*, vol. 3, pp. 1–9, Feb. 2021. DOI: 10.1037/met0000100.

[36] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 206–219, Mar. 2006, ISSN: 1067-5027. DOI: 10.1197/jamia.M1929. eprint: https://academic.oup.com/jamia/article-pdf/13/2/206/2178574/13-2-206.pdf. [Online]. Available: https://doi.org/10.1197/jamia.M1929.

[37] D. Jurafsky and J. H. Martin, "Speech and language processing," in 2023, ch. 5, Draft of January 7, 2023.

[38] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society*, vol. 53, pp. 370–418, 1763, By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. DOI: 10.1098/rstl.1763.0053.

[39] Vikramkumar, V. B, and Trilochan, *Bayes and naive bayes classifier*, 2014. arXiv: 1404.0933 [cs.LG].

[40] T. T. S. Nguyen and P. M. T. Do, "Classification optimization for training a large dataset with naïve bayes," *Journal of Combinatorial Optimization*, vol. 40, no. 1, pp. 141–169, 2020. DOI: 10.1007/s10878-020-00578-0. [Online]. Available: https://doi.org/10.1007/s10878-020-00578-0.

[41] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[42] Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015. DOI: 10.11919/j.issn.1002-0829.215044.

[43] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning* (Adaptive Computation and Machine Learning), 2nd ed. Cambridge, MA: MIT Press, 2018, 504 pp., ISBN: 978-0-262-03940-6.

[44] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *CoRR*, vol. abs/1405.4053, 2014. arXiv: 1405.4053. [Online]. Available: http://arxiv.org/abs/1405.4053.

[45] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, Apr. 2001. DOI: 10.1145/505282.505283.

[46] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. arXiv: 1908.10084. [Online]. Available: http://arxiv.org/abs/1908.10084.

[47] X. Sun, Y. Meng, X. Ao, *et al.*, "Sentence Similarity Based on Contexts," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 573–588, May 2022, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00477. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00477/2022948/tacl_a_00477.pdf. [Online]. Available: https://doi.org/10.1162/tacl%5C_a%5C_00477.

[48] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[49] A. L. developers, *ASReview Insights - Insights and plotting tool for the ASReview project*. [Online]. Available: https://github.com/asreview/asreview-insights.

[50] Wikipedia, *Precision and recall - wikipedia, the free encyclopedia*, [Online; accessed 09/23], 2023. [Online]. Available: https://en.wikipedia.org/wiki/Precision_and_recall.

[51] G. Ferdinands, R. Schram, J. de Bruin, *et al.*, "Performance of active learning models for screening prioritization in systematic reviews: A simulation study into the average time to discover relevant records," *Systematic Reviews*, vol. 12, Jun. 2023. DOI: 10.1186/s13643-023-02257-7.

[52] A. C., "How to write a good abstract for a scientific paper or conference presentation.," *Indian J Psychiatry. 2011 Apr;53(2):172-5.*, DOI: 10.4103/0019-5545.82558.

[53] B. Rosner, R. J. Glynn, and M. L. Lee, "The wilcoxon signed rank test for paired comparisons of clustered data," *Biometrics*, vol. 62, no. 1, pp. 185–192, 2006. DOI: 10.1111/j.1541-0420.2005.00389.x.

[54] B. Derrick, a. broad antonia, D. Toher, and P. White, "The impact of an extreme observation in a paired samples design," *Metodološki Zvezki - Advances in Methodology and Statistics*, vol. 14, Oct. 2017. DOI: 10.51936/ktch6909.

[55] OpenAI. "What are tokens and how to count them." (2023), [Online]. Available: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them.

[56] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.10.118. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220307153.

[57] "Wilcoxon signed-rank test." (2023), [Online]. Available: https://www.sciencedirect.com/topics/mathematics/wilcoxon-signed-rank-test.

[58] J. McDonald, *Handbook of Biological Statistics*, 3rd ed. Baltimore, Maryland: Sparky House Publishing, 2014, pp. 186–189.