Significance of language morphology in active learning aided systematic reviews

Mathijs van der Kroft, B.Eng.

Supervisor: Jelle Tijema, M.Sc.

First examiner: Prof. dr. Rens van de Schoot

Second examiner: Prof. dr. Marco van Leeuwen

Abstract

Active learning aided abstract screening can alleviate the labour-intensive process of systematic reviewing. In such a learning cycle, a machine learning model suggests the next abstract to be reviewed, and a researcher classifies the abstract as relevant or irrelevant. A systematic review should include all relevant studies, regardless of the language it is conducted in. Machine translation of abstracts helps here, but it is unknown how classification performance changes when abstracts are translated. This study simulates the active learning process with English datasets, and with the same datasets that were machine-translated to German, Spanish and Turkish. A key step in the active learning pipeline is the generation of a vector representation of the text, using a feature extractor. The feature extraction methods tf-idf, Doc2Vec, FastText and SBERT were compared on their classification performance for all languages. The results show that no consistent disadvantage to translation can be found for the selected datasets, except for FastText.

CONTENTS

1 INTRODUCTION AND MOTIVATION

Active learning aided abstract screening can alleviate the labour-intensive process of systematic reviewing. In such a learning cycle, a machine learning model suggests the next abstract to be reviewed, and a researcher classifies the abstract as relevant or irrelevant. Systematic reviews should include all valid articles, no matter the language they are written in [1]. Machine translation is a valid option for researchers who wish to include foreign languages in their study [2]. However, how active learning models comparatively perform on translated texts is unclear. This study aims to provide insight into the change in classification performance introduced by translation into linguistically more complex languages than English. The study simulates the active learning process with English datasets, and with the same datasets that were machine-translated to German, Spanish and Turkish. To generate a better understanding of the research question this study aims to answer, a general overview of the relevant literature is provided.

2 LITERATURE REVIEW

2.1 SYSTEMATIC REVIEWS

Systematic reviews are a form of literature review that aim to synthesize the results of many related research studies into a single overview. The studies to be reviewed should be gathered in an unbiased, comprehensive, transparent and reproducible manner [1]. Once this search has been performed, the relevant studies need to be identified, based on a predetermined set of inclusion and exclusion criteria. An initial search can easily identify thousands of potentially relevant studies [3]. A first step in selecting relevant studies from this search is often abstract screening. Here, the abstract of every search result is weighed against the inclusion/exclusion criteria. Human abstract reviewers mark abstracts as relevant or irrelevant with error rates of about 1 in 9 abstracts [4], [5]. Abstract reviewing is therefore often conducted by a team of reviewers, which makes it an even more labour-intensive process.

2.2 ACTIVE LEARNING AIDED ABSTRACT SCREENING

Active learning aided abstract screening aims to reduce the workload of human abstract screening. In active learning, a machine learning model picks the next instance(s) it will learn from. With human-in-the-loop machine learning, a human then labels these instances [6]. In the case of abstract screening, a human reviewer can label an abstract that is proposed by the model for reviewing as relevant or irrelevant for the study they are conducting. After being trained on this added information, the model will make new predictions on the unlabelled data and make the next selection for labelling. This process ends when all instances are labelled, or when the human stops labelling earlier. A workload reduction in abstract screening can of course only be achieved if the researcher stops labelling before the last abstract is reviewed.

2.3 ASREVIEW

ASReview is an open-source software platform developed at Utrecht University [6]. It provides an offline pipeline for active learning aided reviewing. As input, it requires a set of records (e.g. abstracts), and at least one record pre-labelled as relevant, and one as irrelevant. Its output is the subset of relevant and irrelevant records, labelled by the human-in-the-loop. The four basic elements of this pipeline are feature extraction, classification, query strategy and balance strategy. [7]

2.3.1 Feature extraction

The feature extraction method defines how the input records are transformed into a feature matrix, in which each document is represented as a vector. This transformation of records needs to be done only once before the active learning cycle starts. The default method is tf-idf (term frequency-inverse document frequency) [8], which weights a bag-of-words vector by the inverse frequency of the words as they appear in the entire corpus. Among other available methods are Doc2Vec [9] and Sentence-BERT (SBERT) [10], which are respectively simple and complex neural network feature extractors, which try to embed the semantics of a document within a vector. In the ASReview implementation, SBERT uses a pre-trained model, while Doc2Vec trains on the corpus of abstracts.

2.3.2 Classification

Classification is the method by which a probability is calculated for the unlabelled records to belong to the 'relevant' class, given the labelled records. ASReview provides implementations ranging from statistical methods like Multinomial Naive Bayes (default) and Logistic Regression to Neural Network-based methods.

2.3.3 Balance strategy

In a systematic review, usually, the irrelevant records far outnumber the relevant ones. This class imbalance can lead to a model that has high accuracy, but only because it has a high true negative rate and most cases are indeed negative. The balance strategy resamples the training data to account for the imbalance.

2.3.4 Query strategy

The query strategy defines the method by which the relevance probabilities that are calculated by the classifier lead to the next recommended record to be reviewed. When the goal is to find the next most similar record, the record with the highest probability must be recommended. However, a random or uncertainty-based query strategy may find new clusters of relevant records sooner.

2.3.5 Simulation

ASReview implements a method to simulate human-in-the-loop active learning. In such a use case, a fully labelled dataset of abstracts is provided as input. Several prelabelled datasets are available in the ASReview repository. A set of relevant and irrelevant labels is provided as prior knowledge for the classifier to train on. The classifier makes a recommendation for the next record to review. The classifier is then retrained with the added knowledge of the label of the new recommendation it made, as if a human would assign the label at that moment. The simulation continues until a stopping rule is reached (e.g. all relevant records found).

2.4 Language Morphology

Words are composed of the smallest semantic units in a language, called morphemes. Morphology is the study of words in a language; how they are built up from morphemes, and how their composition depends on the linguistical context in which they appear [11, p. 2]. In morphologically rich languages (MRLs), like Turkish, the combination of morphemes into words is varied and complex. Therefore, MRLs have a larger vocabulary than analytical (morphologically poor) languages, like English, given the same corpus size. The larger vocabulary makes the MRLs in general perform worse than English on language processing tasks that implement statistical feature extraction methods and that do not account for additional morphological rules [12]. The larger vocabulary increases the sparseness of the feature matrix for tf-idf, since the chance that a term occurs in multiple documents decreases.

Models trained on multiple languages can suffer from the curse of multilinguality, decreasing in performance with the more languages they are trained on [13]. This could lead to a decrease in classification performance when implementing a feature extraction model that has been pre-trained to recognize multiple languages. Multilingual SBERT, one of the feature extractors that will be tested in RQ1, could have lower classification performance than an SBERT model trained on English only, due to the multilingual nature of its pretrained vectors.

3 RESEARCH QUESTION

This study investigates the following: *Does the classification performance of active learning aided systematic review simulations depend on the morphological richness of the input language?*

This research question has been divided into two sub-questions, which can be methodologically investigated:

RQ1: What is the classification performance of current feature extractors implemented in ASReview for languages with differing morphological richness?

RQ2: Does FastText as a feature extractor increase classification performance over the selected languages and feature extractors in RQ1?

Results from this study could be used to inform researchers on the possibility of including translations of abstracts of studies that were written in a foreign language in their active learning aided systematic review pipeline.

3.1 RQ1

This study explores how classification performance depends on the morphological richness of a language. In RQ1 it is investigated if this dependence is different for the feature extraction methods tf-idf, Doc2Vec and SBERT, which are already implemented in ASReview. For td-idf, the link between increased vocabulary size and feature matrix sparseness is direct (see Table 1). Tf-idf cannot use sparse words to find similar documents, since they only occur in one document. Therefore, tf-idf is expected to perform worse with an increase in morphological richness. For Doc2Vec it is less obvious how increased vocabulary size will impact the vector training, since the model is fitted to vectors of a fixed dimension. SBERT does not fit a model to the dataset at all, but uses a pre-trained model to transform documents to vectors of a fixed dimension. The hypothesis for RQ1 is: *All feature extraction methods will see a significant decrease in classification performance due to machine translation into a language with a higher morphological richness than English.*

3.2 RQ2

FastText [14] is a word embedding skip-gram model like Word2Vec (on which Doc2Vec is based). FastText extends this method by comprising each word embedding of a bag of n-gram sub-word vectors. The n-grams typically are 3 to 6 characters long. This way, even out-of-vocabulary (OOV) words can be transformed into a vector representation [13]. Vector generation for OOV words could be an important feature for text classification of MRLs, because they have a high OOV rate due to their linguistic complexity. In one example, the OOV rate for Turkish was 8% versus English 1% at the same vocabulary size (60k) [15]. For RQ2, a FastText feature extractor will be implemented and subjected to the same setups as the feature extractors in RQ1. The hypothesis for RQ2 is: *FastText will not see a significant decrease in classification performance due to machine translation into a language with a higher morphological richness than English.*

4 DATA AND METHODS

Three languages were selected to be compared to the classification performance in English. English is mostly an analytical language. English words are not often agglutinated or transformed based on their context. Therefore, it is considered morphologically poor. Selection criteria for comparison languages were: 1) Different language families; 2) Different morphological typology; 3) Reasonable number of native speakers (+50 million). Based on these criteria, the following languages were chosen:

- German; it has higher derivational synthesis than English, agglutinating morphemes more often than English. It stems from the Germanic language family [16].
- Spanish; it has higher relational synthesis than English, adding bounding morphologies to root words to create new grammatical meaning. It stems from the Latin language family [16].
- Turkish; it is a highly relational synthetic language, with complex grammatical rules like vowel harmony that contribute to a large vocabulary. Statistical language processing approaches perform poorly due to the sparseness of words within a corpus [15]. It stems from the Turkic language family.

The increase in vocabulary size and sparse words (words that occur in only a single document) after machine translation will be used as a proxy for the increase in morphological richness introduced by the translation into a linguistically more complex language.

The six datasets that were used to estimate classification performance are the same as the ones used in Ferdinands et al. [17]. These prelabeled datasets were selected on basis of their diversity in research fields and availability. They are part of the ASReview `systematic-review-datasets` package and have been published under an open license. In this report, the investigated datasets are referred to as: ACE, Nudging, PTSD, Software, Virus and Wilson. They respectively cover systematic reviews on the following research topics: Angiotensin-Converting Enzyme Inhibitors [18]; Nudging Health Care Professionals [19]; Post Traumatic Stress Disorder Trajectories [20]; Software Fault Prediction [21]; Virus Metagenomics [22]; Wilson's disease [23].

The abstracts and titles in the datasets were translated using Google Translate v3 machine translation. In comparative studies, DeepL slightly outperforms Google Translate on standard translation metrics [24]. However, Google Translate can perform document translation of excel files, which decreases the complexity of the translation pipeline. Google Translate also supports more languages than DeepL. Google translate has been deemed a viable, accurate tool for translation for medical systematic reviews [2].

For **RQ1**, the datasets were classified in all languages on all selected feature extractors: tf-idf, Doc2Vec and multilingual SBERT. During Doc2Vec pre-processing, stop word removal was performed in the applicable language. For SBERT, the pretrained multilingual model 'distiluse-base-multilingual-cased-v2' [10] was used, which supports 50+ languages, including the ones under investigation. In this report, the multilingual SBERT feature extractor will be referred to as simply 'SBERT'.

For **RQ2**, FastText was implemented as a feature extractor. For each language, pre-trained word vectors trained on the Common Crawl corpus of that language were used [25]. These word vectors have 300 dimensions and are comprised of 5-character long n-grams. For each document in a dataset, all word vectors were calculated. These word vectors were then transformed into a sentence vector using the `get_sentence_vector()`function from the `fasttext` package. This function calculates the L2 norm of each separate dimension over all word vectors in the document.[1]

For *classification*, the Logistic Regression classifier was used instead of the default Multinomial Naive Bayes classifier, since it is able to classify the negative vector components of Doc2Vec and SBERT. The default *balance* and *query strategies* were used.

For each *simulation setup* of a language and feature extraction combination, a simulation was run 15 times. Each simulation was initiated with one randomly picked relevant and one randomly picked irrelevant record as prior knowledge. To decrease the runtime of the simulations, the feature extraction step of SBERT was performed only once per setup, and the resulting feature matrix was used for all 15 simulations. This method is not different to generating the feature matrix every single time.

The *recall curve* visualises the progression of the labelling process (see Figure 1). It plots the number of records that have been reviewed as a proportion of the total number of records in the dataset on the x-axis. This can be understood as a progression through time, where every step on the x-axis is a newly labelled record. This is plotted against the number of relevant records that have been labelled as a proportion of all relevant records in the dataset on the y-axis. Therefore, this plot can only be drawn once all relevant records are known.

The *Relevant Records Found* (RRF) is the proportion of relevant records that have been found after reviewing a certain proportion of all records. When read from the recall curve, the RRF is the y-value at a given x-value. In the results, the RRF after reviewing 10% of all records is reported. If one would stop reviewing after only a certain portion (e.g. 10%) of all records were reviewed, and label all remaining records as ‘irrelevant’, the RRF would be equal to the recall of the experiment.

If the next record to be reviewed was picked completely at random (simple sampling), the averaged recall curve would approach a straight line with intercept = 0 and slope = 1. The *Work Saved over Sampling* (WSS) is the difference in the number of records that need to be reviewed to reach a certain RRF (e.g 95%) between the actual recall curve and the simple sampling line. In the results, the WSS after finding 95% of all relevant records is reported.

The *Average Time to Discovery* (ATD) was introduced by Ferdinands et al. [17] as a metric for classification performance. It is the average number of document reviews it takes to find a relevant document, as a percentage of all documents in the dataset. The closer the ATD is to zero, the better the classification performance. Another way to interpret the ATD is as the area above the recall curve. For each setup, the estimated mean and standard error of the mean (SEM) of the ATD over 15 runs were calculated.

With Student’s t-test, the ATD of each machine-translated setup was compared to the same setup with the original (not translated) dataset, to estimate the statistical significance of the sample

[1] For RQ2, the Sent2Vec [26] FastText implementation was considered as a feature extraction method. This would have allowed for unsupervised sentence vector training, much like the current Doc2Vec method. However, the available package `sent2vec` is only supported for Linux and macOS distributions. Therefore this method was not further investigated.

difference. The resulting p-value reports the probability that both sets of 15 samples were drawn from the same t-distribution. In other words, the p-value represents the probability that the null hypothesis “The machine translation had no impact on classification performance” is true.

5 RESULTS

5.1 VOCABULARY SIZE

Table 1 compares the word counts of the machine translation for German, Spanish and Turkish (DE, ES, TR) to the original English (EN) PTSD dataset. These translations were also translated back to English (EN from XX). The vocabulary size is the total number of unique words in the dataset. The sparse words occur in only one document in the dataset. It can be seen that the vocabulary size and the number of sparse words increase like: EN < ES < DE < TR. This is in line with the claim that Turkish has a high morphological richness. Note that when the translated datasets are translated back to English, the word counts drop below the original dataset. This implies that a generalization of the vocabulary occurs, due to the machine translations. The ratio of vocabulary size to sparse words stays roughly equal over all versions of the dataset.

Table 1: Sparse words in the PTSD dataset

Language	Vocabulary size	Sparse Words	Sparse Words to Vocabulary ratio
EN (original)	22334	9103	0.41
DE	37170	17002	0.46
ES	27702	11134	0.40
TR	45657	20047	0.44
EN (from DE)	20791	8184	0.39
EN (from ES)	20744	8119	0.39
EN (from TR)	20405	7875	0.39

5.2 RECALL CURVES

For the sake of brevity, only the recall curves of the PTSD dataset are displayed here. All dataset recall curves can be viewed at full scale in the Tables and Figures section. Each line is comprised of the average of 15 samples. The width of the curve represents the standard error of the mean RRF at a given point in time. It can be noted that the WSS@95% for the German SBERT setup overestimates its deviation from the other setups. The RFF@10% is not a good performance measure for the tf-idf and Doc2Vec, because most samples have reached 100% RRF by then. Both WSS and RRF represent a single point in time and are not representing classification performance over the whole review process. The ADT is reported further on as the classification performance measure of choice, since it summarises the performance over the whole simulated review.

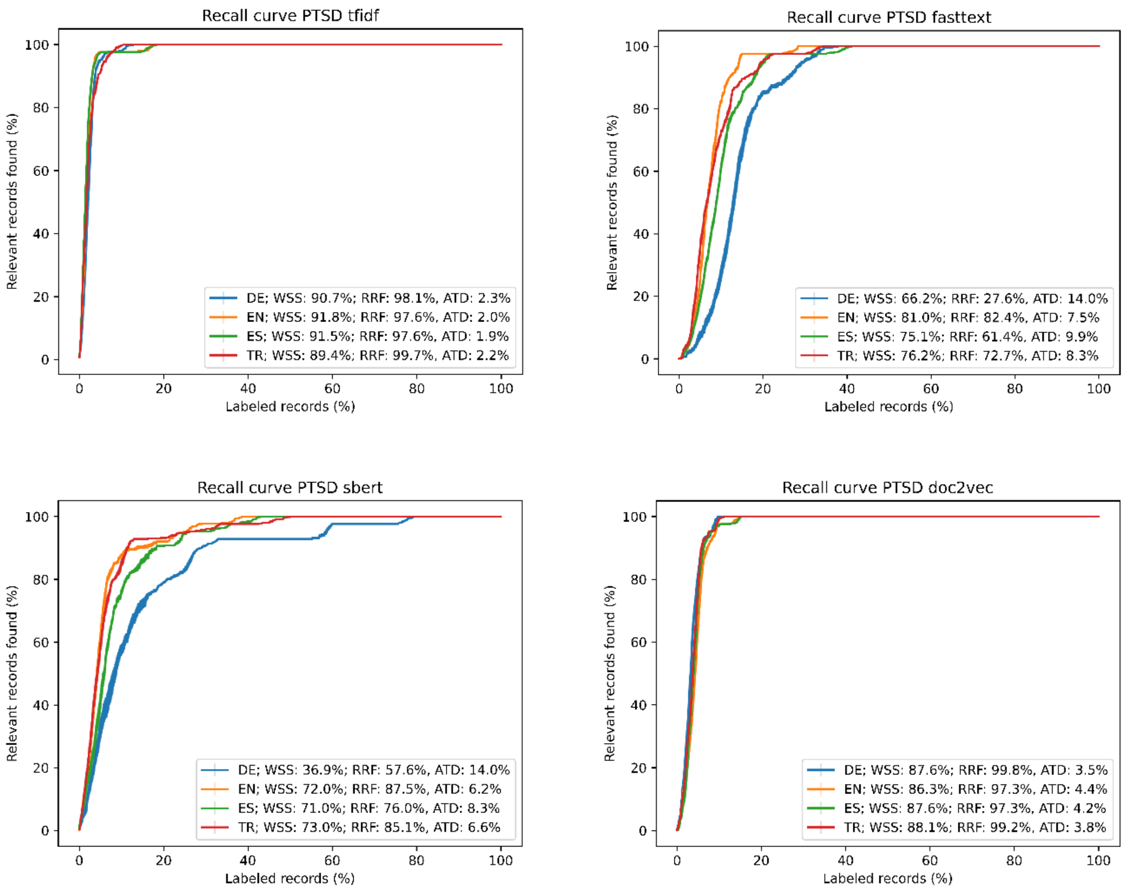


Figure 1: Recall curves of the PTSD dataset. RRF at 10% labelled records, WSS at 95% RRF and ATD are included in the legend

5.3 ATD COMPARISON

In Figure 2 and Figure 3, all the simulation results are presented. Each unique combination of feature extractor, translation language and dataset is represented as a point. On the x-axis, the ATD of a translated dataset setup is shown. On the y-axis, the ATD of the same setup is shown, except for using the original English dataset. The reported ADT for a setup is the average of 15 simulations. The position of each point shows the relation between the classification performance of a translated setup and the same setup with the original dataset. The closer a point is to the origin, the better the classification performance is in general.

The dashed grey line represents the null hypothesis that machine translation has no impact on classification performance, since in that case ATD Translated would be equal to ATD Original. If a point appears above the grey line, this means that classification performance increased after translation. Conversely, classification performance decreased after translation if a point appears below the grey line. If the deviation from the grey line is not significant (p-value ≥ 0.05 for a t-test on the ATD of 15 original and 15 translated simulations), then the point is drawn as a cross.

The simulation results are split over two figures, to account for the performance difference per dataset. The ATD for the “Software” and “PTSD” datasets is much lower (i.e. better) than for the other datasets. To visually represent the points properly, they were plotted in a separate figure, with shorter axis ranges. As a consequence, Figure 2 and Figure 3 are not one-to-one visually comparable.

All plotted results can be viewed as well in Table 2 in the Tables and Figures section.

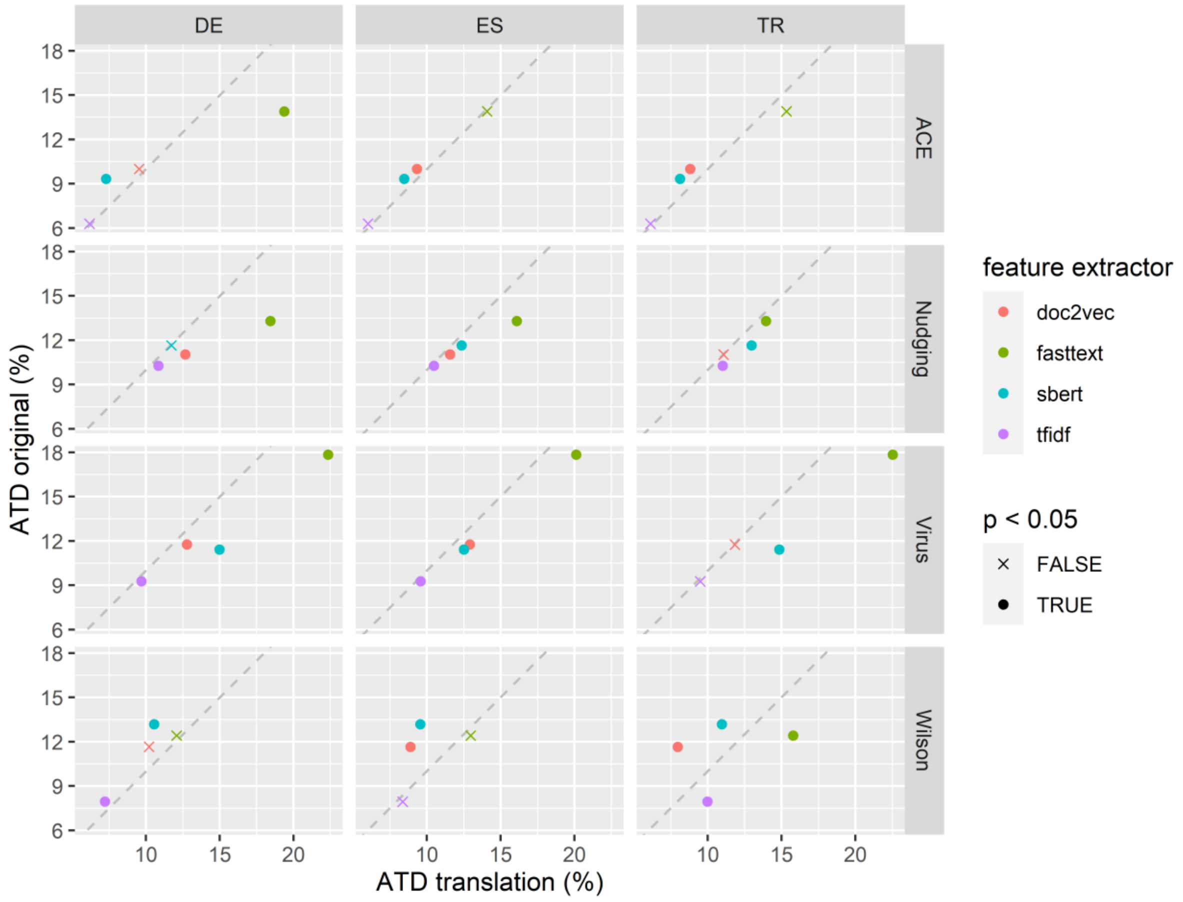


Figure 2: ATD comparison for ACE, Nudging, Virus and Wilson datasets

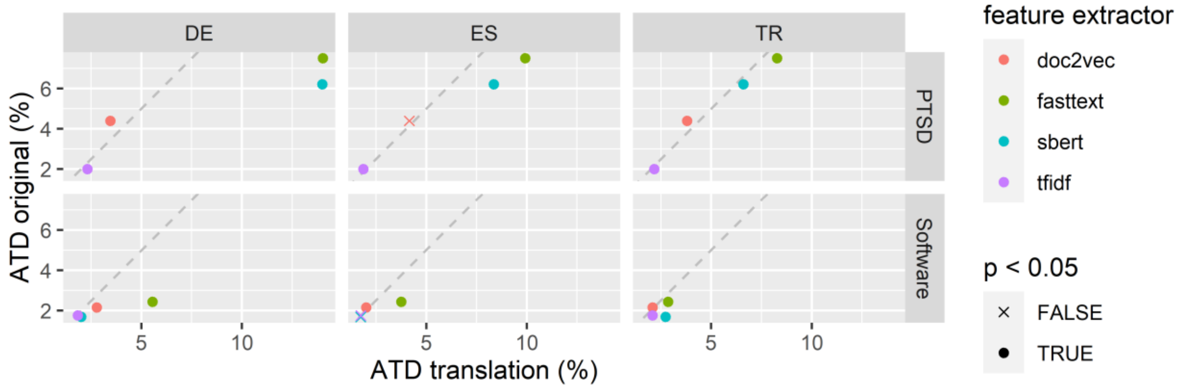


Figure 3: ATD comparison for PTSD and Software datasets. Please take note that the ATD-axes ranges are shorter and closer to zero

5.4 Results RQ1

Concerning **RQ1**, it was found that for tf-idf, Doc2Vec and SBERT, there exist setups that perform significantly better on the translation than on the original dataset. Therefore the hypothesis that all feature extractors will perform significantly worse on translation to languages with a higher morphological richness than English, must be rejected.

For Doc2Vec and SBERT, the change in ATD after translation seems to depend on the dataset. Doc2Vec's ATD improved for ACE, Wilson and PTSD, and worsened for Nudging and Virus. SBERT's ATD improved for ACE and Wilson, and worsened for Nudging, Virus, PTSD and Software. Noteworthy is SBERT's relatively poor performance on the German PTSD dataset (ATD original: 6.2%; ATD translation: 14.0%).

Tf-idf seems resilient to the translations in its classification performance. Most setups do not show a large deviation from the null line. A slight performance loss for tf-idf can be seen for all significant Turkish setups, which is in line with the expectation that a high morphological richness will decrease tf-idf performance. Especially on the Turkish Wilson set, tf-idf performs relatively poorly. This also is the only setup where tf-idf is clearly outperformed by another feature extractor, Doc2Vec in this case. SBERT performs slightly better than tf-idf on the original Software dataset.

5.5 Results RQ2

Concerning **RQ2**, the performance of the current implementation of FastText as a sentence embedding algorithm has the worst overall classification performance of all investigated feature extractors. The only setup where FastText performs better than another feature extractor is on the original Wilson dataset, where it slightly outperforms SBERT.

Furthermore, the classification performance of the translations relative to their original datasets is significantly worse for most setups. Especially in the German translations, FastText performs poorly. Therefore, the hypothesis that FastText will *not* see a significant decrease in classification performance due to machine translation, must be rejected as well.

6 Discussion

This study intended to find a relation between the morphological richness of a language, and the classification performance of different feature extractors. The most surprising and unintuitive finding is that translation into a morphologically richer language than English does not impact the classification performance in a consistently negative way (with exception of the current FastText implementation). This was observed even for an MRL like Turkish, which had double the vocabulary size of the original language.

6.1 Limitations

The results from this study are foremost limited by the low number of datasets and languages that were used. From the sample space of three languages, performance in other languages is hard to infer. Languages that are not written in the Latin alphabet were not included in this study, and remain to be investigated. The six different datasets showed large variability among the performance of different feature extractors. Also here it is difficult to infer how different sets of abstracts will respond to translation, let alone use cases other than systematic reviewing.

So why is there no clear impact caused by translation? It could well be that the words and document features that account for the most classification performance are so specific within the research field that translating them does not transform them in an impactful way. Going a step further, it is possible that the machine translation step generalizes terms in some datasets in such a way that the similarity of relevant documents increases. Therefore, it is difficult to extrapolate the findings of this study to datasets of documents that were originally written in another language than English.

An explanation for the stable performance of tf-idf is that, although the number of sparse words increases during translation, the vocabulary size also increases, so their ratio stays roughly the same. Tf-idf cannot use the sparse words to find similar documents, since they only occur in one document. However, the increase in vocabulary size might compensate for the increase in sparse words for tf-idf, since the total number of non-sparse terms actually increases.

Doc2Vec creates vectors using a skip-gram model, so it trains the vectors of words that surround a word. In this way, it may circumvent the issue of sparse words partly, since sparse words will occur in a similar context to words that mean almost the same.

SBERT and FastText are pre-trained on very large datasets, so they should not suffer from increased sparseness. Based on the results, there is no evidence to suggest that multilingual SBERT performs better in English than in the other investigated languages.

The FastText model seems to perform better in English than in the other languages. This could mean that the training quality of the vectors in the non-English models is lower than in the English model. However, it could also be due to the way FastText was implemented. Taking the L2-norm of all word vectors in a document is not an effective way to extract semantics from a document, since the semantics encoded in the order of the words is lost. This study was not designed to estimate the vector quality difference of pre-trained models on different languages in an unambiguous way.

6.2 Further research

Further research could include implementing a FastText unsupervised sentence vector training algorithm like Sent2Vec [26]. However, the benefit of FastText, namely being able to represent OOV words, may be very slim compared to other pretrained libraries like SBERT, since these pre-trained libraries have a vast vocabulary. The OOV rate of SBERT on the English and the translated datasets has not yet been investigated. Multilingual models like multilingual SBERT should also be able to classify datasets containing *mixed language* documents. The classification performance of such a setup remains to be researched.

This study only investigated datasets translated from English. However, datasets originating in other languages that are translated into English were not investigated. This may be a valid follow-up study, since it would provide further evidence for researchers who wish to use active learning to review documents from multilingual sources that are translated to English.

Tf-idf is still by far the best-performing feature extractor that is implemented in ASReview. Given the large number of sparse terms (~40% of the vocabulary) that are ignored by tf-idf, it might be beneficial to transform these terms. A pre-trained word embedding library like BERT, Word2Vec or FastText could be used for this purpose as a pre-processing step for tf-idf. This pre-processor would convert a sparse word into a similar word that occurs in at least one other document if the similarity between these words is above a certain threshold.

6.3 Conclusion

This study finds no clear and structurally significant disadvantage to classification in active learning aided systematic reviews when the investigated datasets are translated to German, Spanish or Turkish. An exception is the discussed FastText implementation. These are promising results for researchers who hope to use an active learning pipeline with (partially) translated documents. However, further research is needed to find out if these results apply to different types of setups.

7 References

[1] E. Aromataris and A. Pearson, ‘The Systematic Review: An Overview’, *AJN Am. J. Nurs.*, vol. 114, no. 3, pp. 53–58, Mar. 2014, doi: 10.1097/01.NAJ.0000444496.24228.2c.

[2] J. L. Jackson *et al.*, ‘The Accuracy of Google Translate for Abstracting Data From Non–English-Language Trials for Systematic Reviews’, *Ann. Intern. Med.*, vol. 171, no. 9, p. 677, Nov. 2019, doi: 10.7326/M19-0891.

[3] J. R. Polanin, T. D. Pigott, D. L. Espelage, and J. K. Grotpeter, ‘Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses’, *Res. Synth. Methods*, vol. 10, no. 3, pp. 330–342, Sep. 2019, doi: 10.1002/jrsm.1354.

[4] Z. Wang, T. Nayfeh, J. Tetzlaff, P. O’Blenis, and M. H. Murad, ‘Error rates of human reviewers during abstract screening in systematic reviews’, *PLOS ONE*, vol. 15, no. 1, p. e0227742, Jan. 2020, doi: 10.1371/journal.pone.0227742.

[5] G. Gartlehner *et al.*, ‘Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial’, *J. Clin. Epidemiol.*, vol. 121, pp. 20–28, May 2020, doi: 10.1016/j.jclinepi.2020.01.005.

[6] R. van de Schoot *et al.*, ‘An open source machine learning framework for efficient and transparent systematic reviews’, *Nat. Mach. Intell.*, vol. 3, no. 2, pp. 125–133, Feb. 2021, doi: 10.1038/s42256-020-00287-7.

[7] De Bruin, Jonathan *et al.*, ‘ASReview Software Documentation’, Dec. 2021, doi: 10.5281/ZENODO.5565351.

[8] J. E. Ramos, ‘Using tf-idf to determine word relevance in document queries’, *Proc. First Instr. Conf. Mach. Learn.*, vol. 242, pp. 133–142, 2003.

[9] Q. V. Le and T. Mikolov, ‘Distributed Representations of Sentences and Documents’. arXiv, May 22, 2014. Accessed: Jun. 15, 2022. [Online]. Available: http://arxiv.org/abs/1405.4053

[10] N. Reimers and I. Gurevych, ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks’. arXiv, Aug. 27, 2019. Accessed: May 25, 2022. [Online]. Available: http://arxiv.org/abs/1908.10084

[11] M. Aronoff and K. A. Fudeman, *What is morphology?*, 2nd ed. Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell, 2011.

[12] R. Tsarfaty, D. Seddah, S. Kübler, and J. Nivre, ‘Parsing Morphologically Rich Languages: Introduction to the Special Issue’, *Comput. Linguist.*, vol. 39, no. 1, pp. 15–22, Mar. 2013, doi: 10.1162/COLI_a_00133.

[13] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, ‘A Survey on Text Classification Algorithms: From Text to Predictions’, *Information*, vol. 13, no. 2, p. 83, Feb. 2022, doi: 10.3390/info13020083.

[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, ‘Enriching Word Vectors with Subword Information’, 2016, doi: 10.48550/ARXIV.1607.04606.

[15] K. Oflazer, ‘Turkish and its challenges for language processing’, *Lang. Resour. Eval.*, vol. 48, no. 4, pp. 639–653, Dec. 2014, doi: 10.1007/s10579-014-9267-2.

[16] E. Sapir, *Language: an introduction to the study of speech*. New York: Harcourt, Brace Jovanovich, 1970.

[17] G. Ferdinands *et al.*, ‘Active learning for screening prioritization in systematic reviews - A simulation study’, Open Science Framework, preprint, Sep. 2020. doi: 10.31219/osf.io/w6qbg.

[18] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, 'Reducing Workload in Systematic Review Preparation Using Automated Citation Classification', *J. Am. Med. Inform. Assoc.*, vol. 13, no. 2, pp. 206–219, Mar. 2006, doi: 10.1197/jamia.M1929.

[19] R. Nagtegaal, L. Tummers, M. Noordegraaf, and V. Bekkers, 'Nudging healthcare professionals towards evidence-based medicine: A systematic scoping review', *J. Behav. Public Adm.*, vol. 2, no. 2, Oct. 2019, doi: 10.30636/jbpa.22.71.

[20] R. van de Schoot, M. Sijbrandij, S. D. Winter, S. Depaoli, and J. K. Vermunt, 'The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies', *Struct. Equ. Model. Multidiscip. J.*, vol. 24, no. 3, pp. 451–467, May 2017, doi: 10.1080/10705511.2016.1247646.

[21] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, 'A Systematic Literature Review on Fault Prediction Performance in Software Engineering', *IEEE Trans. Softw. Eng.*, vol. 38, no. 6, pp. 1276–1304, Nov. 2012, doi: 10.1109/TSE.2011.103.

[22] K. T. T. Kwok, D. F. Nieuwenhuijse, M. V. T. Phan, and M. P. G. Koopmans, 'Virus Metagenomics in Farm Animals: A Systematic Review', *Viruses*, vol. 12, no. 1, p. 107, Jan. 2020, doi: 10.3390/v12010107.

[23] C. Appenzeller-Herzog, T. Mathes, M. L. S. Heeres, K. H. Weiss, R. H. J. Houwen, and H. Ewald, 'Comparative effectiveness of common therapies for Wilson disease: A systematic review and meta-analysis of controlled studies', *Liver Int.*, vol. 39, no. 11, pp. 2136–2152, Nov. 2019, doi: 10.1111/liv.14179.

[24] G. Sarti, A. Bisazza, A. G. Arenas, and A. Toral, 'DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages', 2022, doi: 10.48550/ARXIV.2205.12215.

[25] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, 'Learning Word Vectors for 157 Languages'. arXiv, Mar. 28, 2018. Accessed: Jun. 26, 2022. [Online]. Available: http://arxiv.org/abs/1802.06893

[26] P. Gupta, M. Pagliardini, and M. Jaggi, 'Better Word Embeddings by Disentangling Contextual n-Gram Information', in *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota, 2019, pp. 933–939. doi: 10.18653/v1/N19-1098.

8 DECLARATIONS

8.1 ETHICAL CONSIDERATIONS

This study did not deal with personal information in any way. For this study, only data that has been published under an open licence was used. Google Translate is free for anybody to use under the Google Services Terms of Service

8.2 OPEN SCIENCE

This study has been made publicly available at https://github.com/mathijsvanderkroft/ASReview_Language_Study where all used scripts and generated data have been made freely available under the Apache-2.0 licence. The generated state_files have not been uploaded since their combined size (277GB) prevents this.

9 TABLES AND FIGURES

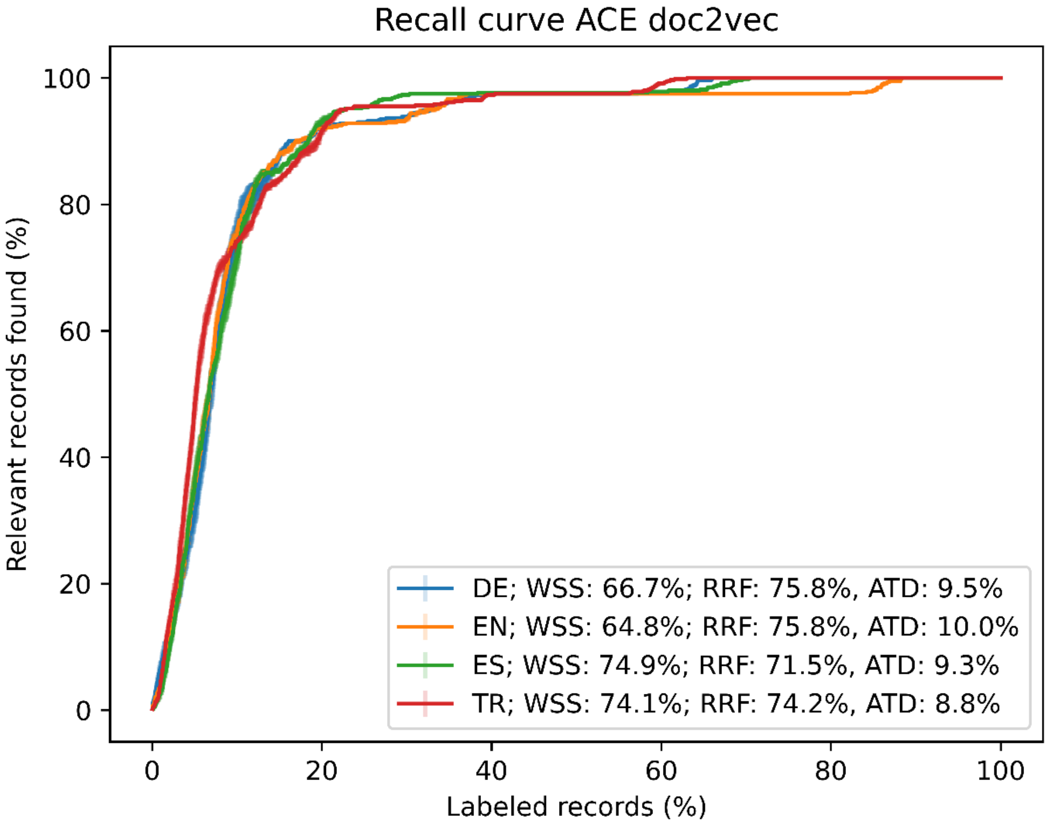
9.1 ATD COMPARISON TABLE

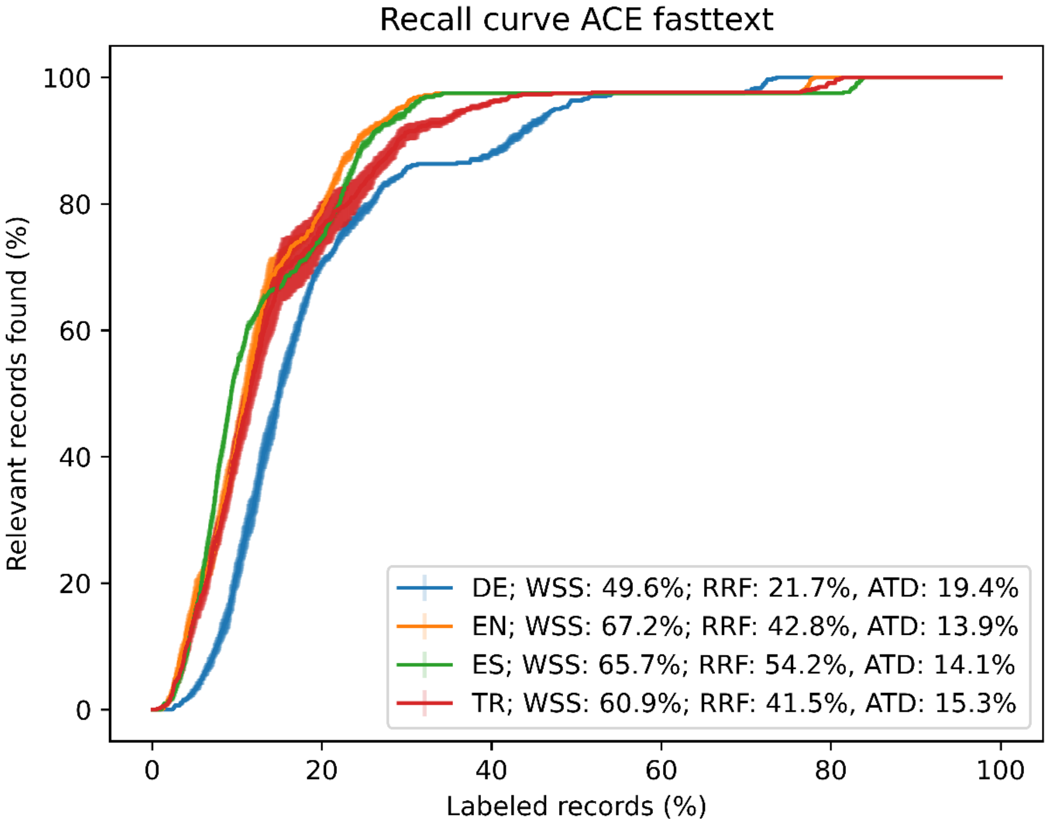
Table 2: Simulation results, each ATD value is the mean of 15 samples. The standard error is the standard error of the mean for the translated datasets.

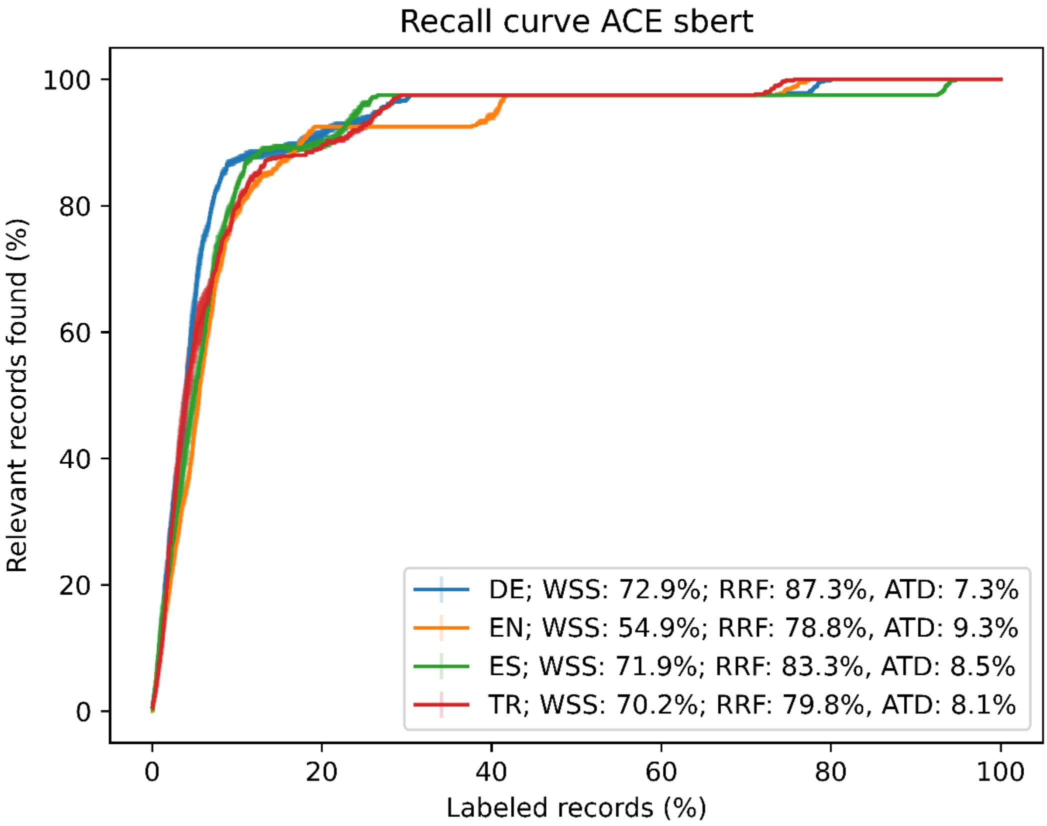
dataset	language	feature extractor	ATD (%)	ATD (%) English	Standard error	p-value	p < 0.05
Software	ES	sbert	1.73	1.69	0.0214	0.068494	FALSE
Software	ES	tfidf	1.74	1.76	0.0279	0.626554	FALSE
Software	DE	tfidf	1.84	1.76	0.0224	0.000474	TRUE
PTSD	ES	tfidf	1.86	1.99	0.0413	0.006237	TRUE
Software	DE	sbert	2	1.69	0.0196	2.58E-15	TRUE
Software	ES	doc2vec	2.01	2.16	0.0341	0.000402	TRUE
Software	TR	doc2vec	2.08	2.16	0.0356	0.038852	TRUE
Software	TR	tfidf	2.08	1.76	0.024	4.01E-13	TRUE
PTSD	TR	tfidf	2.16	1.99	0.0743	0.029058	TRUE
PTSD	DE	tfidf	2.31	1.99	0.091	0.002369	TRUE
Software	TR	sbert	2.73	1.69	0.0236	2.41E-25	TRUE
Software	DE	doc2vec	2.78	2.16	0.0372	5.21E-15	TRUE
Software	TR	fasttext	2.86	2.44	0.0216	1.61E-16	TRUE
PTSD	DE	doc2vec	3.46	4.39	0.1134	2.13E-08	TRUE
Software	ES	fasttext	3.75	2.44	0.0557	1.55E-14	TRUE
PTSD	TR	doc2vec	3.79	4.39	0.1327	0.000113	TRUE
PTSD	ES	doc2vec	4.15	4.39	0.1603	0.153669	FALSE
Software	DE	fasttext	5.55	2.44	0.0491	8.50E-23	TRUE
ACE	ES	tfidf	6.01	6.29	0.3864	0.465009	FALSE
ACE	TR	tfidf	6.11	6.29	0.381	0.629021	FALSE
ACE	DE	tfidf	6.17	6.29	0.4391	0.770739	FALSE
PTSD	TR	sbert	6.6	6.21	0.1549	0.016888	TRUE
Wilson	DE	tfidf	7.21	7.95	0.1401	2.92E-05	TRUE
ACE	DE	sbert	7.29	9.34	0.1415	2.36E-12	TRUE
Wilson	TR	doc2vec	7.96	11.66	0.6874	6.31E-05	TRUE
ACE	TR	sbert	8.12	9.34	0.2032	1.49E-05	TRUE
PTSD	TR	fasttext	8.28	7.5	0.0885	1.73E-09	TRUE
PTSD	ES	sbert	8.35	6.21	0.1647	2.73E-13	TRUE
Wilson	ES	tfidf	8.36	7.95	0.836	0.637312	FALSE
ACE	ES	sbert	8.47	9.34	0.1551	2.02E-05	TRUE
ACE	TR	doc2vec	8.8	10	0.1926	1.56E-06	TRUE
Wilson	ES	doc2vec	8.88	11.66	0.6974	0.000976	TRUE
ACE	ES	doc2vec	9.32	10	0.2088	0.002903	TRUE
Virus	TR	tfidf	9.49	9.27	0.2123	0.313163	FALSE
ACE	DE	doc2vec	9.54	10	0.3294	0.182915	FALSE
Wilson	ES	sbert	9.55	13.18	0.8481	0.000427	TRUE
Virus	ES	tfidf	9.57	9.27	0.1178	0.018212	TRUE
Virus	DE	tfidf	9.7	9.27	0.1666	0.017732	TRUE

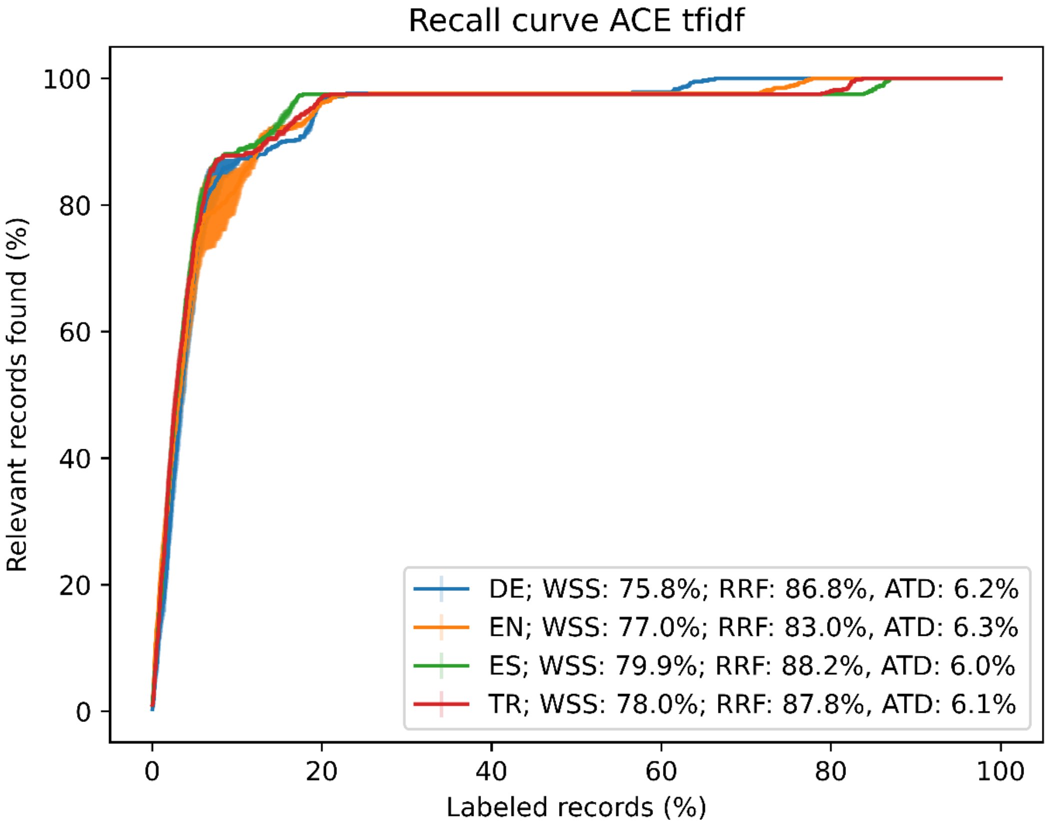
PTSD	ES	fasttext	9.92	7.5	0.0968	8.61E-20	TRUE
Wilson	TR	tfidf	9.98	7.95	0.1828	1.07E-11	TRUE
Wilson	DE	doc2vec	10.21	11.66	0.7412	0.064245	FALSE
Nudging	ES	tfidf	10.47	10.27	0.0891	0.033487	TRUE
Wilson	DE	sbert	10.55	13.18	0.7904	0.004764	TRUE
Nudging	DE	tfidf	10.84	10.27	0.106	2.02E-05	TRUE
Wilson	TR	sbert	10.96	13.18	0.8209	0.015123	TRUE
Nudging	TR	tfidf	11.01	10.27	0.1106	1.17E-06	TRUE
Nudging	TR	doc2vec	11.07	11.03	0.2199	0.862334	FALSE
Nudging	ES	doc2vec	11.56	11.03	0.232	0.032565	TRUE
Nudging	DE	sbert	11.72	11.65	0.2691	0.82178	FALSE
Virus	TR	doc2vec	11.83	11.76	0.1298	0.562183	FALSE
Wilson	DE	fasttext	12.08	12.42	0.5136	0.51586	FALSE
Nudging	ES	sbert	12.36	11.65	0.2995	0.026062	TRUE
Virus	ES	sbert	12.51	11.43	0.1343	1.35E-07	TRUE
Nudging	DE	doc2vec	12.66	11.03	0.2495	7.83E-07	TRUE
Virus	DE	doc2vec	12.78	11.76	0.2059	8.81E-05	TRUE
Virus	ES	doc2vec	12.91	11.76	0.1428	2.13E-08	TRUE
Wilson	ES	fasttext	12.97	12.42	1.2745	0.670627	FALSE
Nudging	TR	sbert	12.97	11.65	0.2618	2.52E-05	TRUE
Nudging	TR	fasttext	13.95	13.3	0.108	2.40E-06	TRUE
PTSD	DE	sbert	14	6.21	0.3062	1.39E-15	TRUE
PTSD	DE	fasttext	14.03	7.5	0.2565	4.21E-14	TRUE
ACE	ES	fasttext	14.08	13.9	0.4147	0.67029	FALSE
Virus	TR	sbert	14.84	11.43	0.1482	3.62E-18	TRUE
Virus	DE	sbert	14.99	11.43	0.2636	7.27E-12	TRUE
ACE	TR	fasttext	15.33	13.9	0.9311	0.141765	FALSE
Wilson	TR	fasttext	15.78	12.42	0.7015	5.60E-05	TRUE
Nudging	ES	fasttext	16.09	13.3	0.1449	1.75E-16	TRUE
Nudging	DE	fasttext	18.43	13.3	0.112	1.06E-27	TRUE
ACE	DE	fasttext	19.38	13.9	0.5076	3.36E-11	TRUE
Virus	ES	fasttext	20.11	17.85	0.1534	2.93E-12	TRUE
Virus	DE	fasttext	22.35	17.85	0.1504	2.52E-18	TRUE
Virus	TR	fasttext	22.52	17.85	0.1707	1.26E-16	TRUE

9.2 ACE Recall Curves

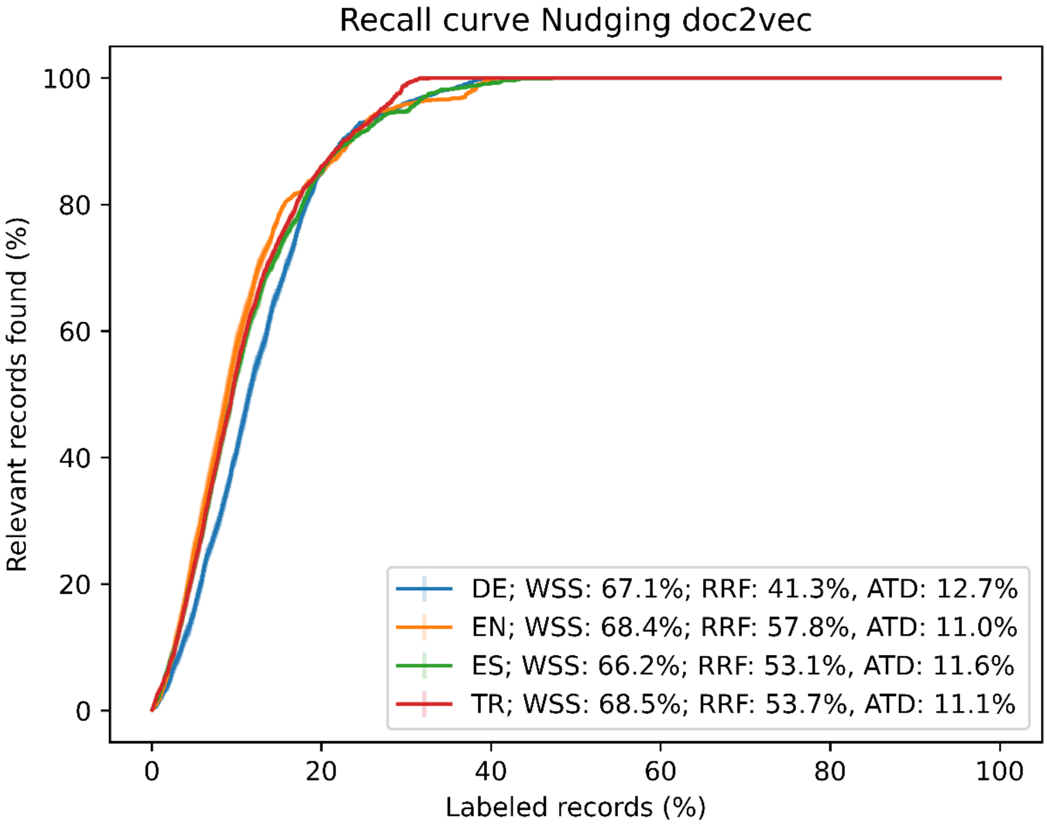


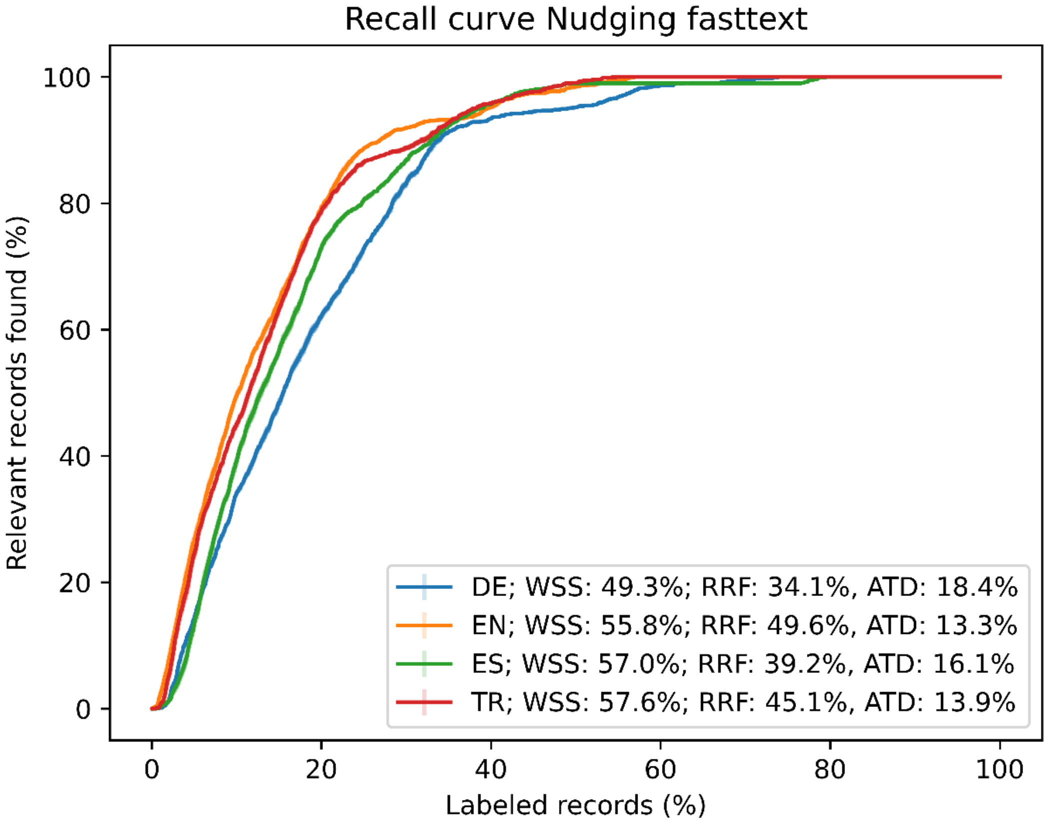


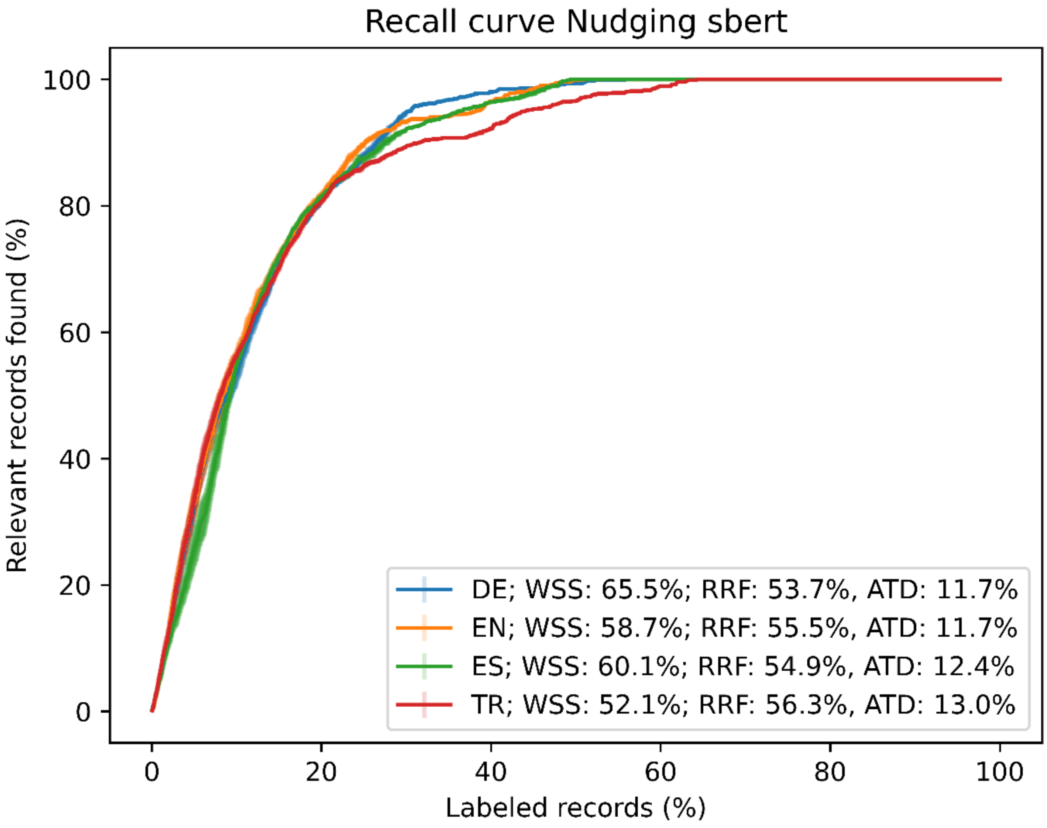

Recall curve ACE sbert
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 72.9%; RRF: 87.3%, ATD: 7.3%
EN; WSS: 54.9%; RRF: 78.8%, ATD: 9.3%
ES; WSS: 71.9%; RRF: 83.3%, ATD: 8.5%
TR; WSS: 70.2%; RRF: 79.8%, ATD: 8.1%

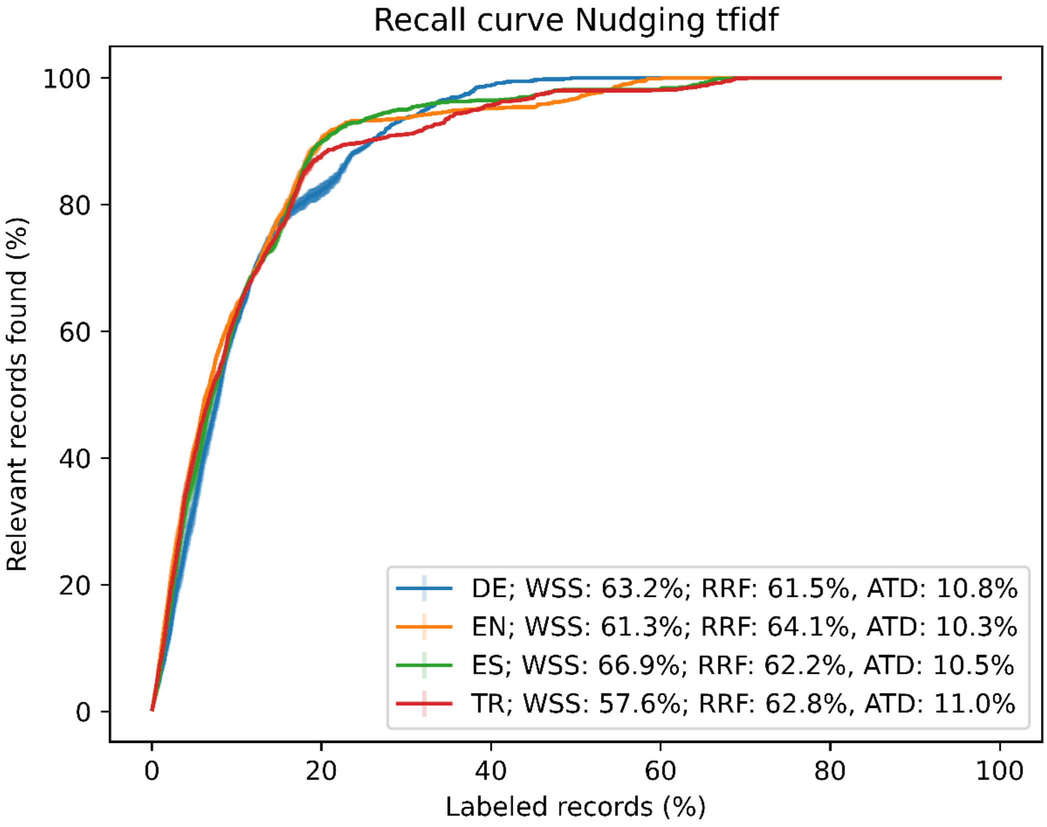

Recall curve ACE tfidf
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 75.8%; RRF: 86.8%, ATD: 6.2%
EN; WSS: 77.0%; RRF: 83.0%, ATD: 6.3%
ES; WSS: 79.9%; RRF: 88.2%, ATD: 6.0%
TR; WSS: 78.0%; RRF: 87.8%, ATD: 6.1%

9.3 NUDGING RECALL CURVES

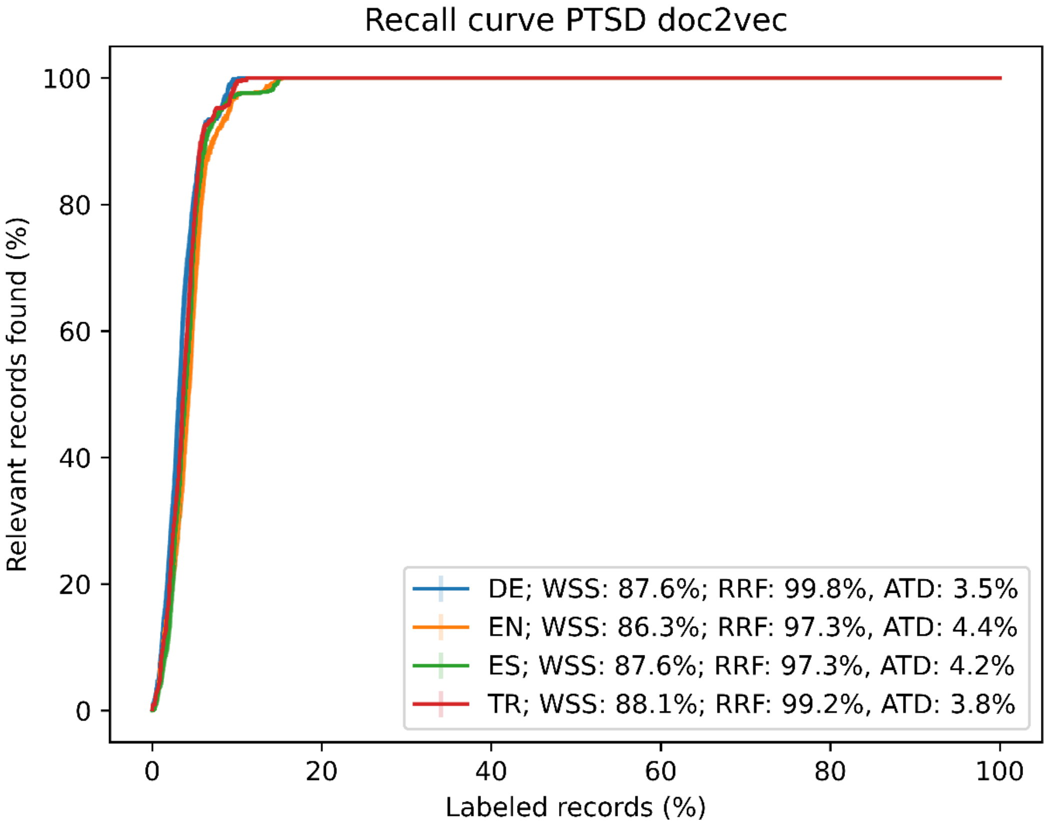


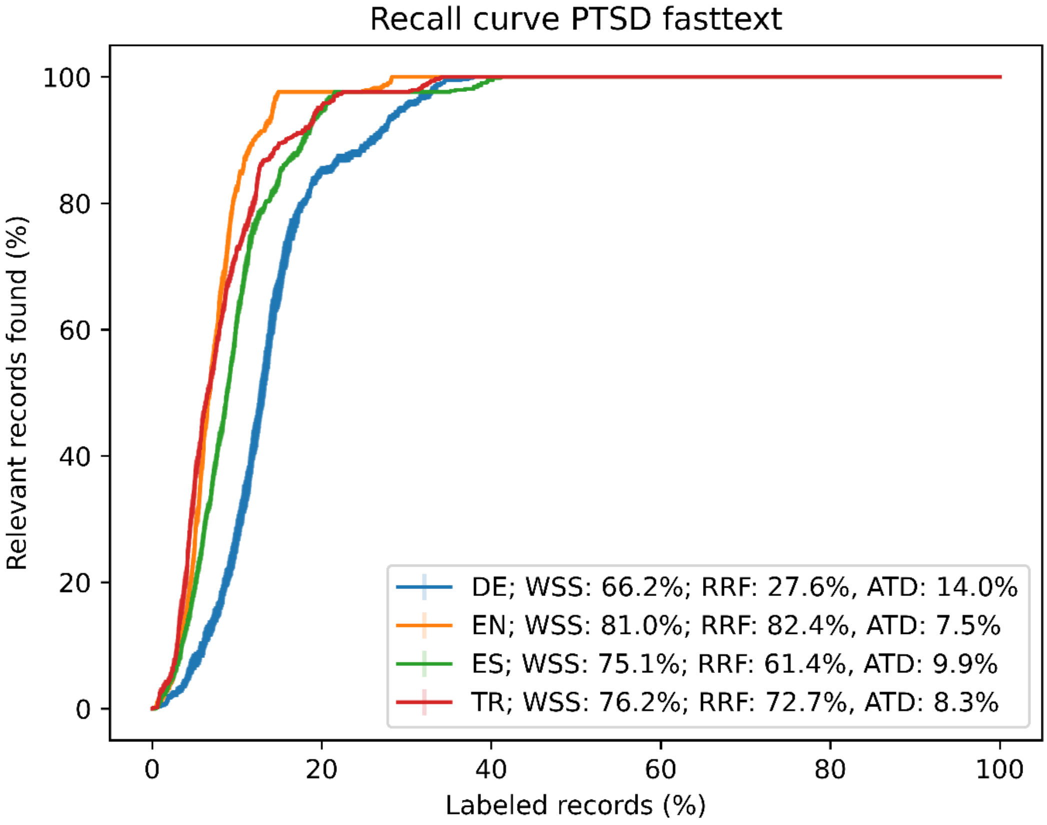


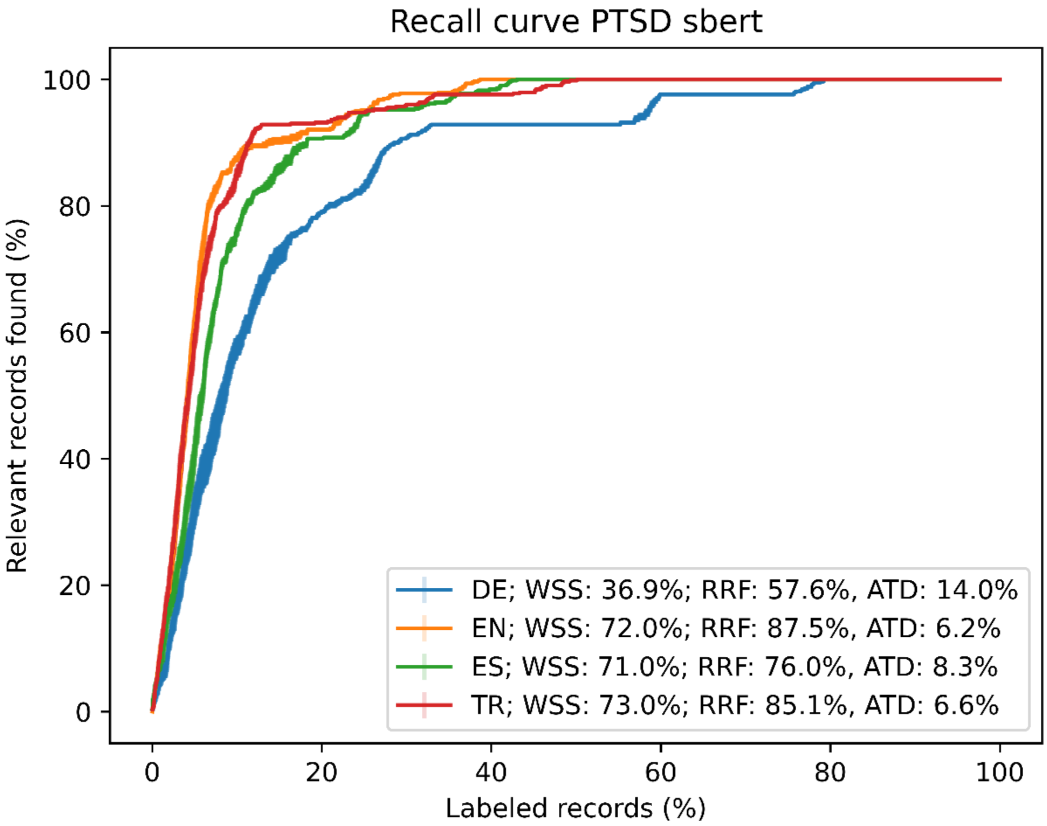

Recall curve Nudging sbert
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 65.5%; RRF: 53.7%, ATD: 11.7%
EN; WSS: 58.7%; RRF: 55.5%, ATD: 11.7%
ES; WSS: 60.1%; RRF: 54.9%, ATD: 12.4%
TR; WSS: 52.1%; RRF: 56.3%, ATD: 13.0%

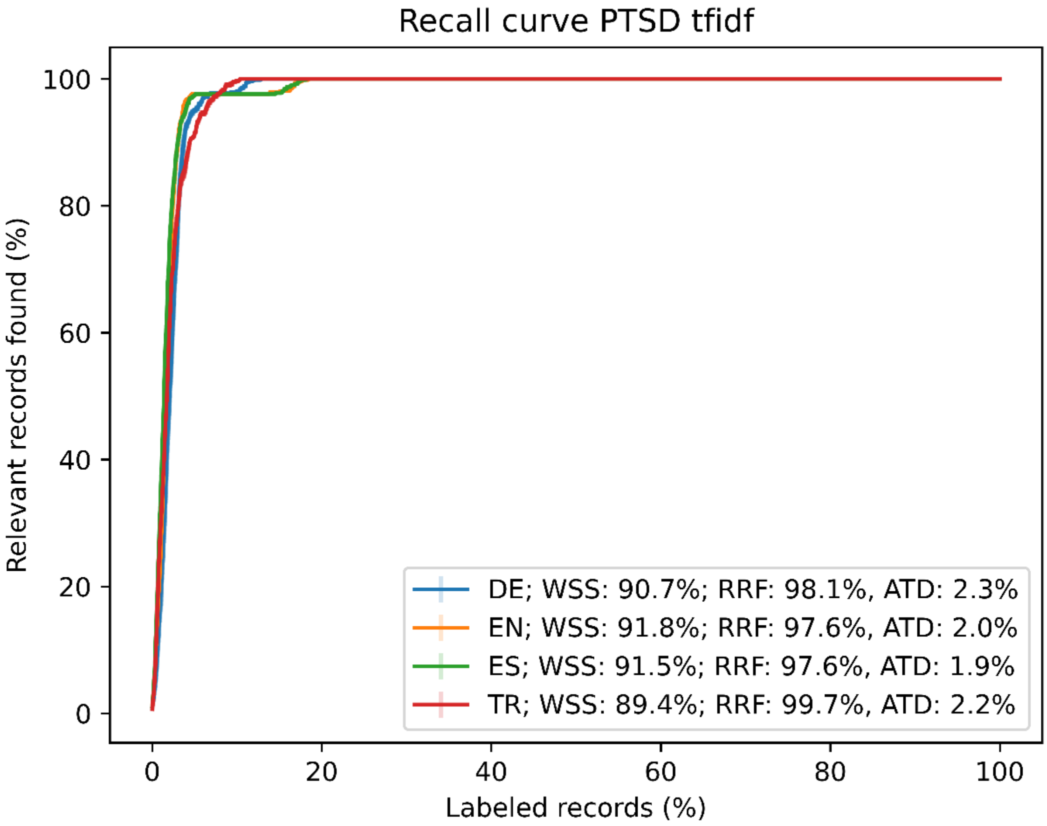

Recall curve Nudging tfidf
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 63.2%; RRF: 61.5%, ATD: 10.8%
EN; WSS: 61.3%; RRF: 64.1%, ATD: 10.3%
ES; WSS: 66.9%; RRF: 62.2%, ATD: 10.5%
TR; WSS: 57.6%; RRF: 62.8%, ATD: 11.0%

9.4 PTSD Recall Curves

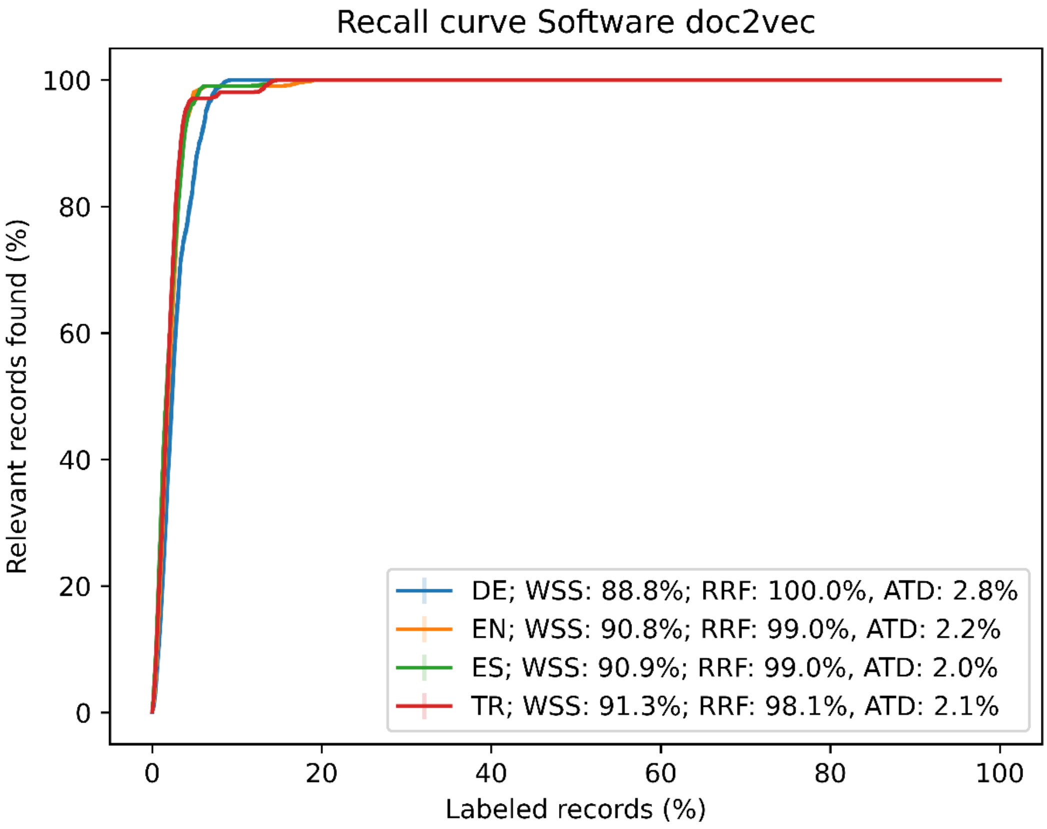


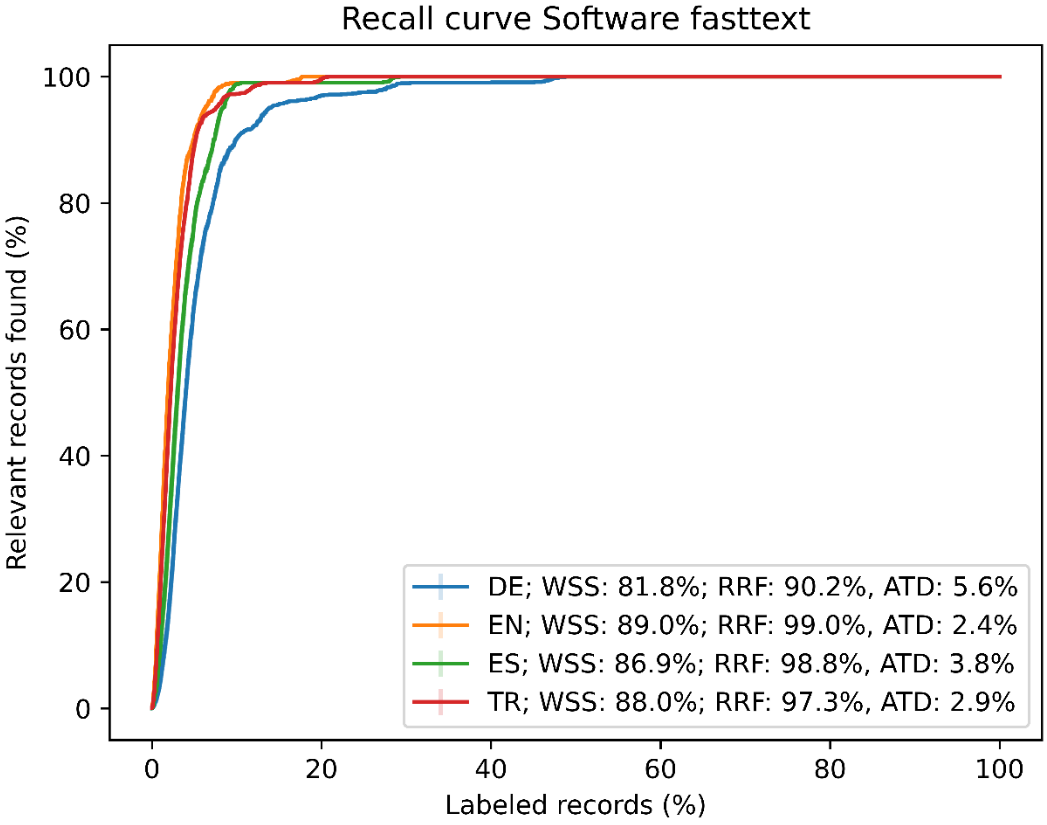


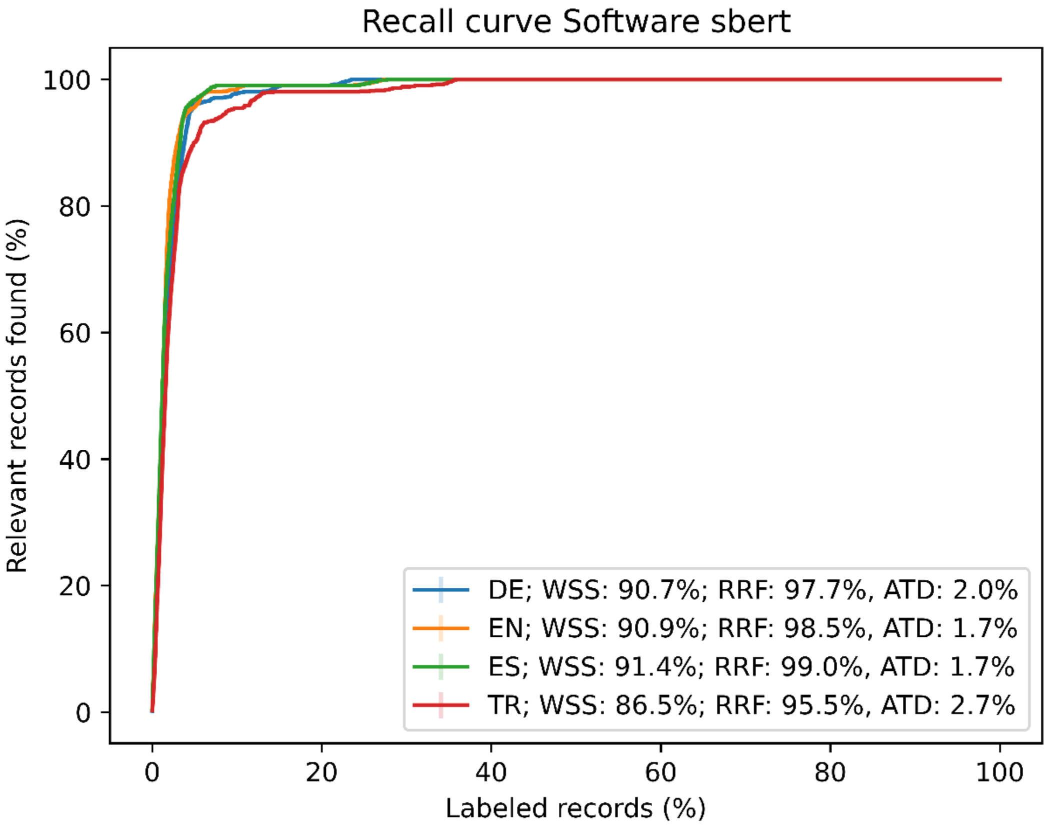

Recall curve PTSD sbert
Relevant records found (%)
100
80
60
40
20
0
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 36.9%; RRF: 57.6%, ATD: 14.0%
EN; WSS: 72.0%; RRF: 87.5%, ATD: 6.2%
ES; WSS: 71.0%; RRF: 76.0%, ATD: 8.3%
TR; WSS: 73.0%; RRF: 85.1%, ATD: 6.6%

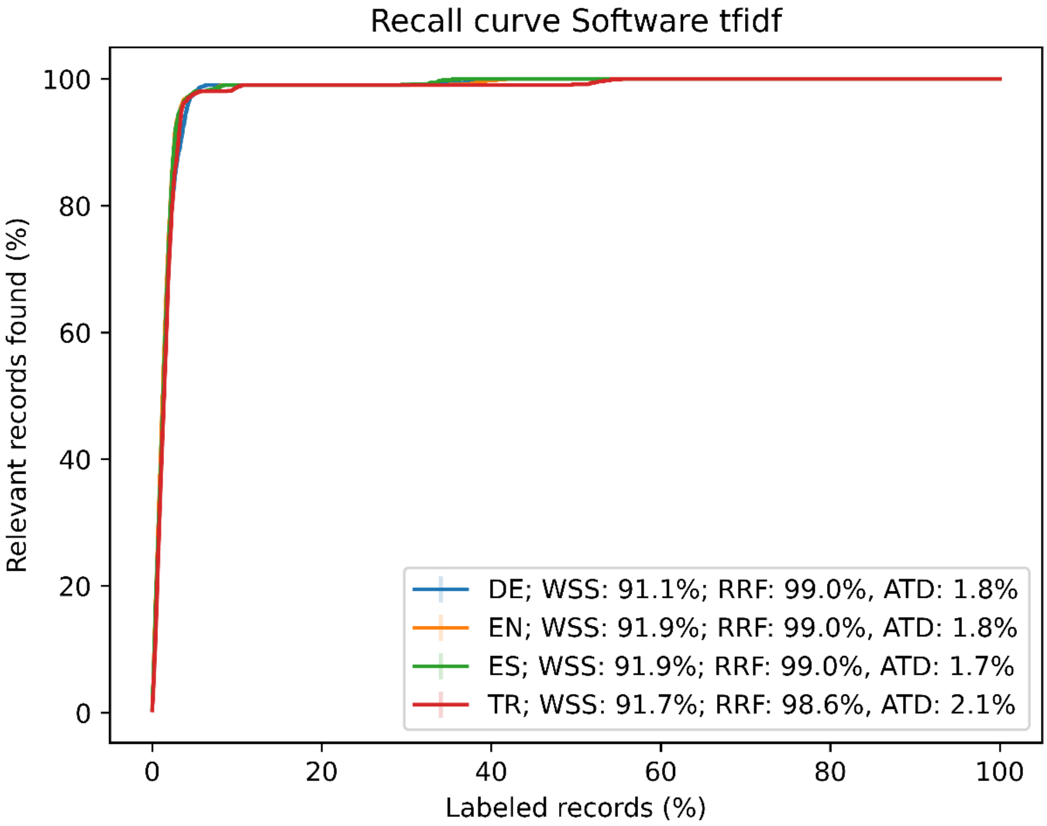

Recall curve PTSD tfidf
Relevant records found (%)
100
80
60
40
20
0
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 90.7%; RRF: 98.1%, ATD: 2.3%
EN; WSS: 91.8%; RRF: 97.6%, ATD: 2.0%
ES; WSS: 91.5%; RRF: 97.6%, ATD: 1.9%
TR; WSS: 89.4%; RRF: 99.7%, ATD: 2.2%

9.5 Software Recall Curves

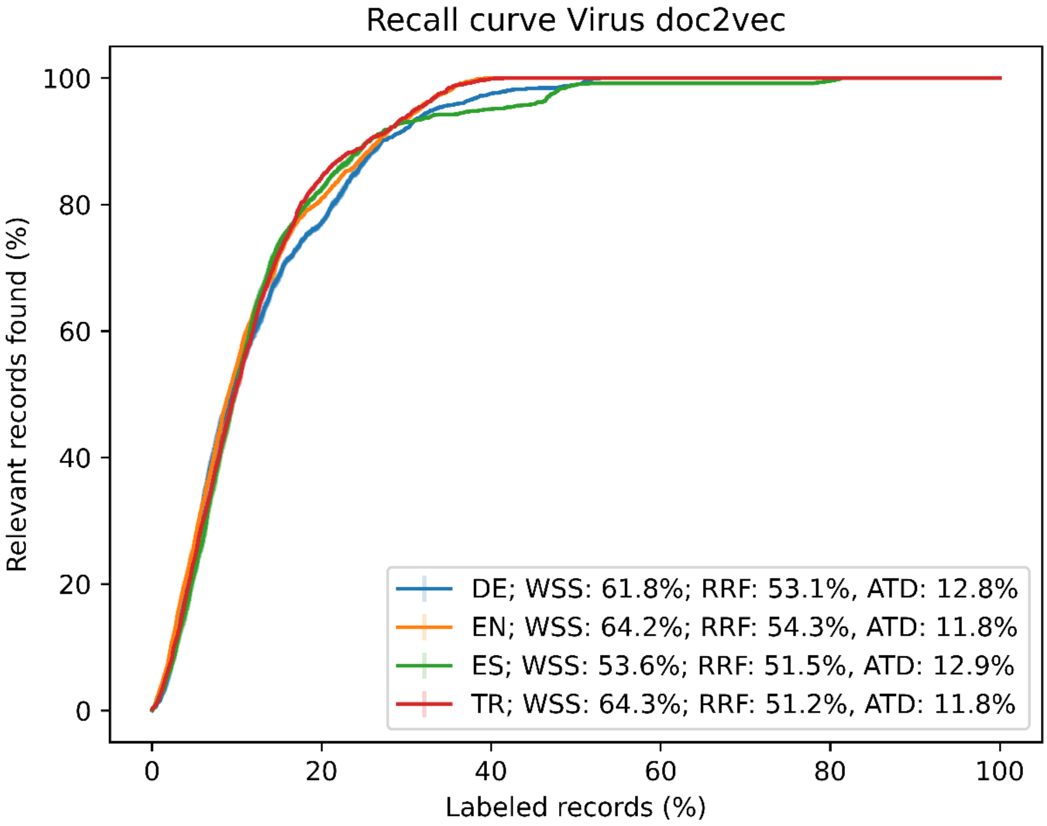


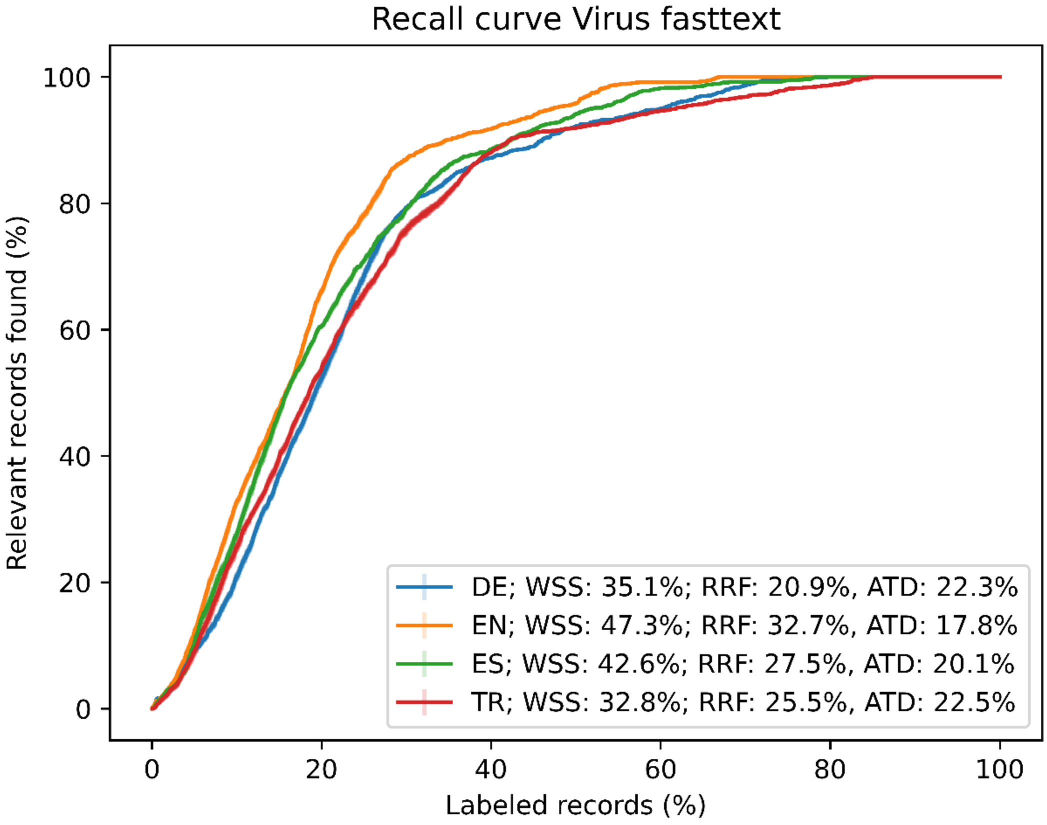


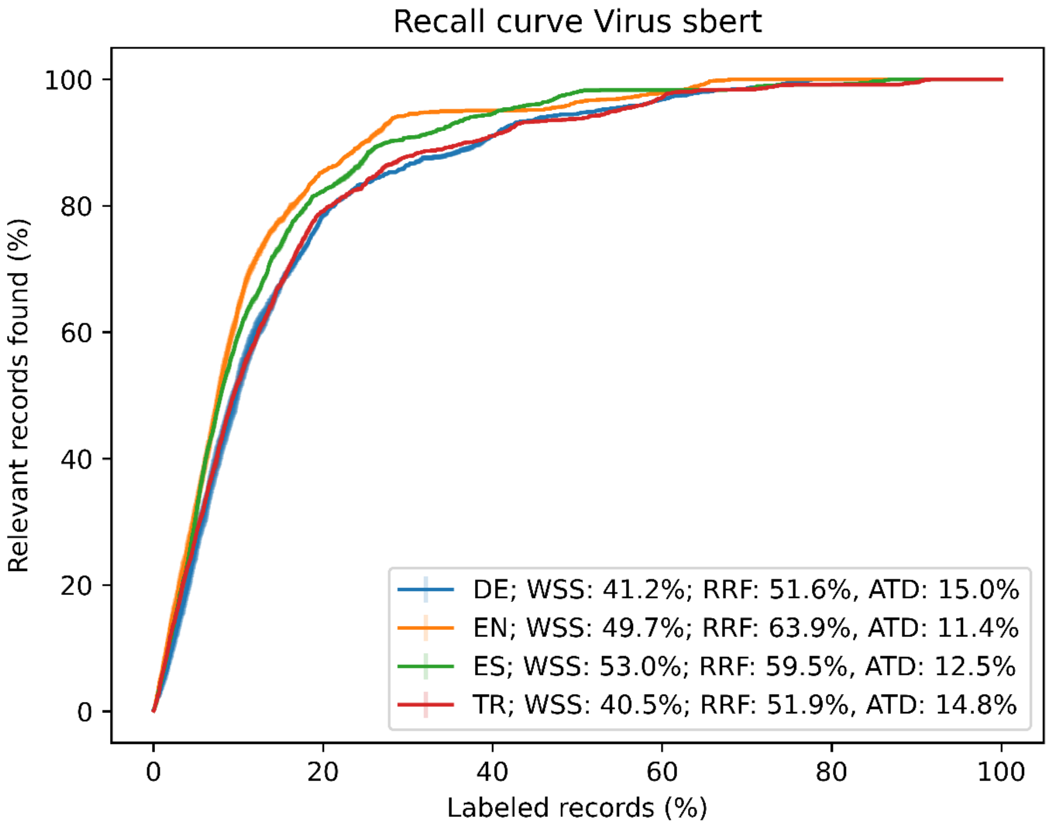

Recall curve Software sbert
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 90.7%; RRF: 97.7%, ATD: 2.0%
EN; WSS: 90.9%; RRF: 98.5%, ATD: 1.7%
ES; WSS: 91.4%; RRF: 99.0%, ATD: 1.7%
TR; WSS: 86.5%; RRF: 95.5%, ATD: 2.7%

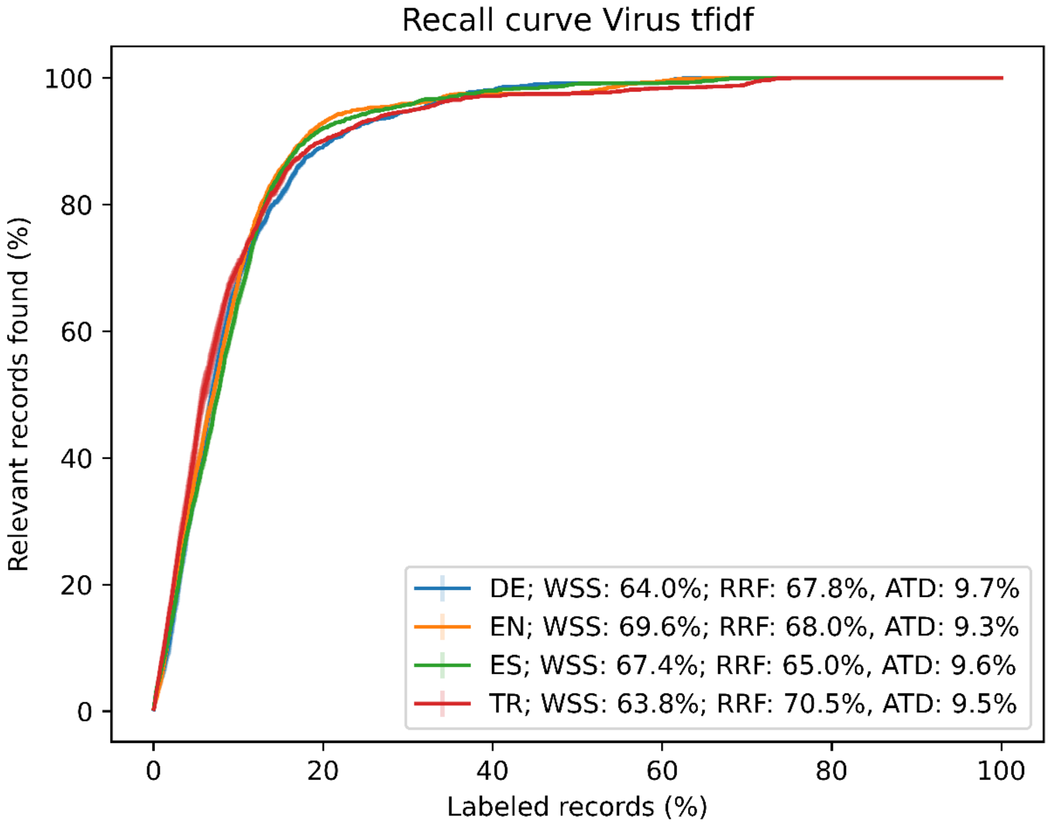

Recall curve Software tfidf
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 91.1%; RRF: 99.0%, ATD: 1.8%
EN; WSS: 91.9%; RRF: 99.0%, ATD: 1.8%
ES; WSS: 91.9%; RRF: 99.0%, ATD: 1.7%
TR; WSS: 91.7%; RRF: 98.6%, ATD: 2.1%

9.6 Virus Recall Curves

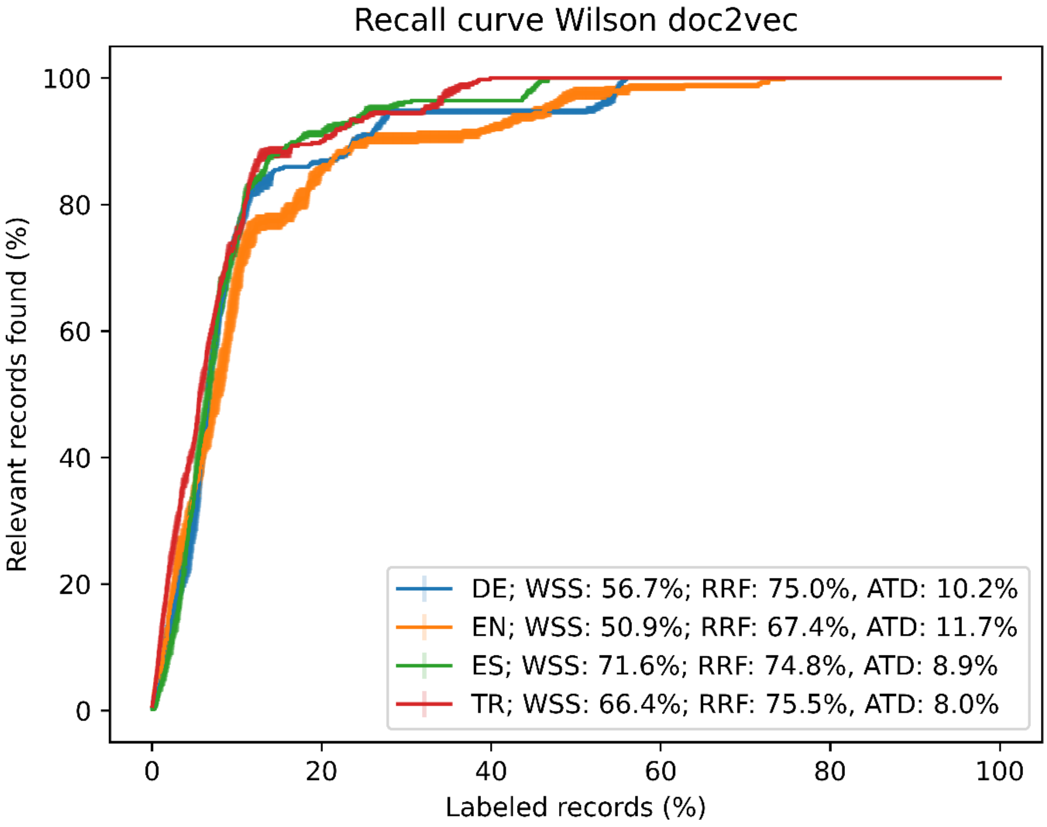


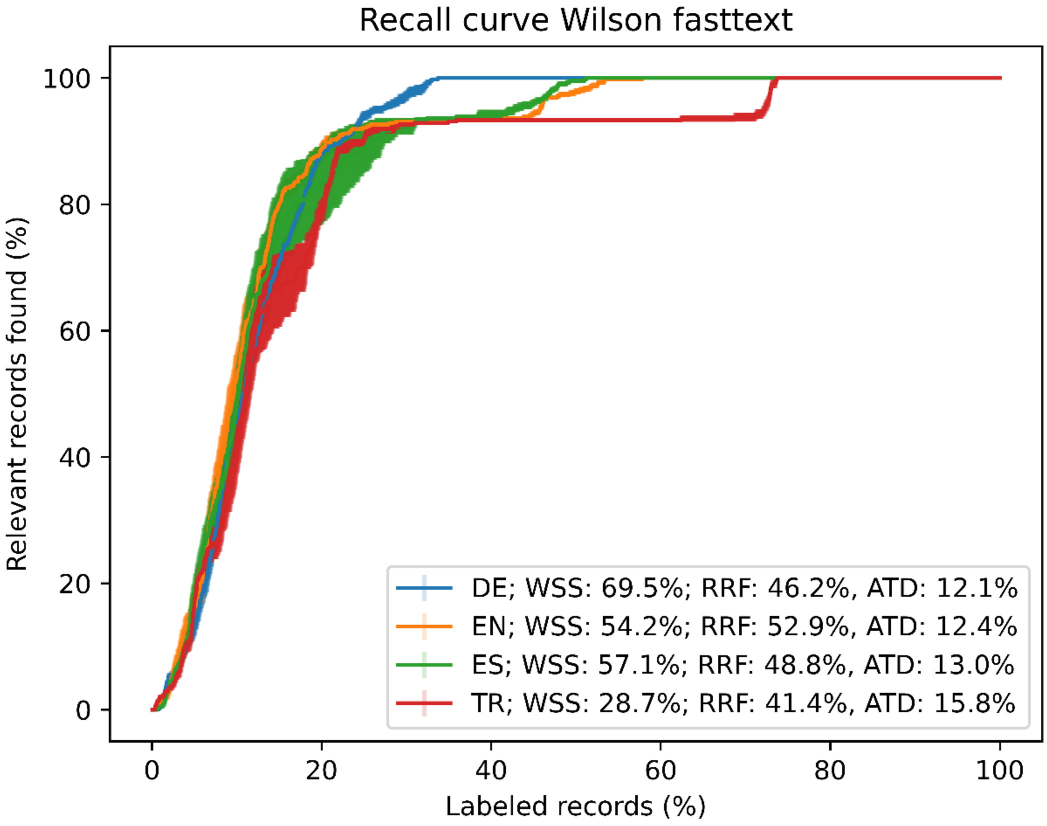


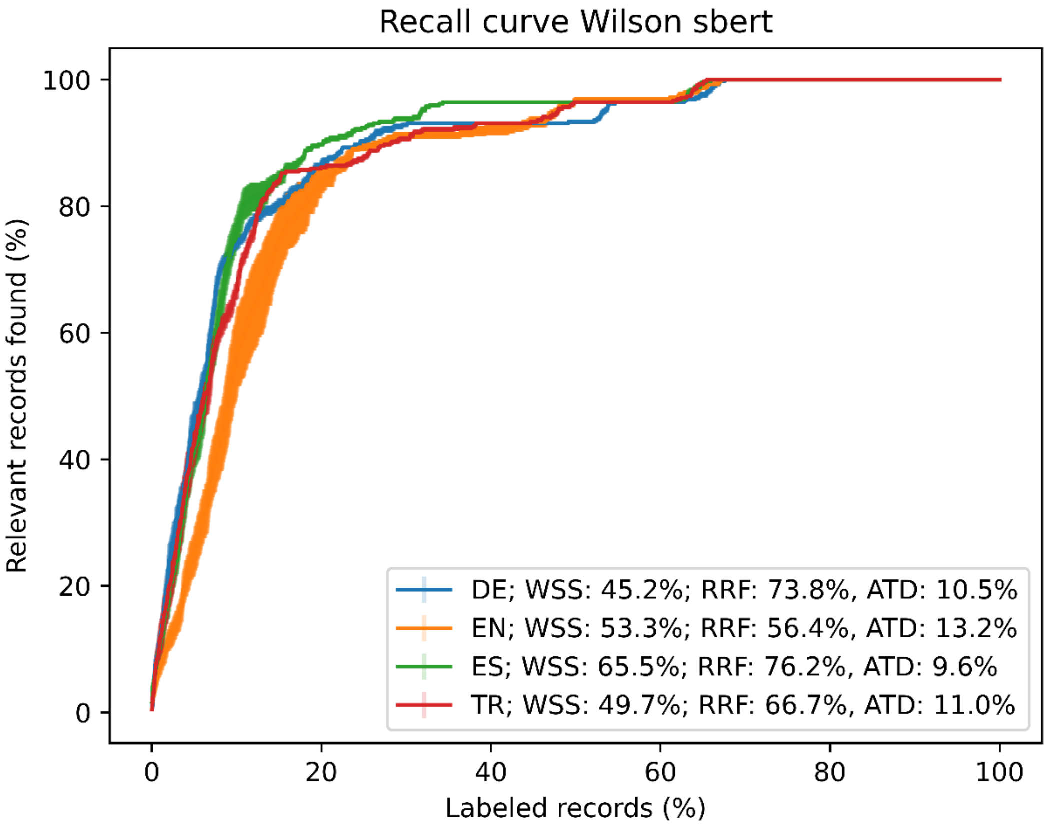

Recall curve Virus sbert
Relevant records found (%)
Labeled records (%)
0
20
40
60
80
100
0
20
40
60
80
100
DE; WSS: 41.2%; RRF: 51.6%, ATD: 15.0%
EN; WSS: 49.7%; RRF: 63.9%, ATD: 11.4%
ES; WSS: 53.0%; RRF: 59.5%, ATD: 12.5%
TR; WSS: 40.5%; RRF: 51.9%, ATD: 14.8%

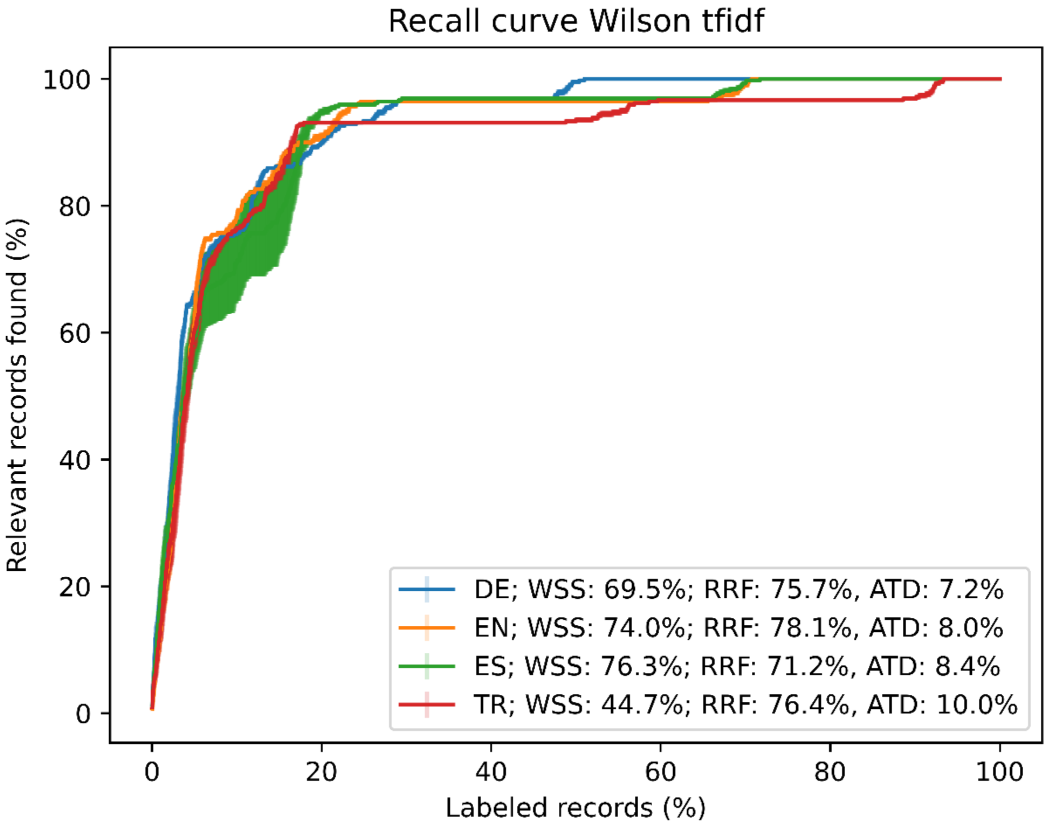

Recall curve Virus tfidf
Relevant records found (%)
Labeled records (%)
0
20
40
60
80
100
0
20
40
60
80
100
DE; WSS: 64.0%; RRF: 67.8%, ATD: 9.7%
EN; WSS: 69.6%; RRF: 68.0%, ATD: 9.3%
ES; WSS: 67.4%; RRF: 65.0%, ATD: 9.6%
TR; WSS: 63.8%; RRF: 70.5%, ATD: 9.5%

9.7 Wilson Recall Curves






Recall curve Wilson sbert
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 45.2%; RRF: 73.8%, ATD: 10.5%
EN; WSS: 53.3%; RRF: 56.4%, ATD: 13.2%
ES; WSS: 65.5%; RRF: 76.2%, ATD: 9.6%
TR; WSS: 49.7%; RRF: 66.7%, ATD: 11.0%


Recall curve Wilson tfidf
Relevant records found (%)
0
20
40
60
80
100
Labeled records (%)
0
20
40
60
80
100
DE; WSS: 69.5%; RRF: 75.7%, ATD: 7.2%
EN; WSS: 74.0%; RRF: 78.1%, ATD: 8.0%
ES; WSS: 76.3%; RRF: 71.2%, ATD: 8.4%
TR; WSS: 44.7%; RRF: 76.4%, ATD: 10.0%