High-Recall NLP Screening for Systematic Reviews: Integrating TF-IDF, SMOTE, and Domain-Specific Criteria

By Christopher Leu, Melody Yacoubian and Vasvi Jain

Abstract:

Systematic reviews play a critical role in evidence-based medicine but are hindered by the labor-intensive process of manually screening thousands of research papers for relevance. To address this bottleneck, we present an NLP-based classification pipeline that automates the initial screening phase. Using a dataset of 2,175 research papers' abstracts and titles from PubMed labeled for inclusion in a brain arteriovenous malformation (AVM) review, we evaluate multiple models including SVM, logistic regression, Complement Naive Bayes, and cosine similarity. Our best-performing model – SVM trained on bigram TF-IDF features with SMOTE balancing – achieved an F_1 score of 0.7083 and a ROC-AUC of 0.9543. We further incorporate domain-specific rule-based features, such as patient age and sample size thresholds, which improved interpretability and boosted performance in edge cases. Through extensive grid search, threshold calibration, and high-recall tuning, our tool demonstrates the ability to reduce reviewer burden without sacrificing sensitivity, providing a scalable and transparent solution for early-stage systematic review screening.

1. Introduction:

Evidence-based medicine, shaping clinical guidelines, informing treatment decisions and underpinning healthcare policies rely heavily on systematic reviews and meta-analyses. While it is an essential process, the initial phase of the reviews requires manually screening thousands of publications for relevance. This reviewing process is time-consuming and could be prone to errors (see Appendix 1 for a visual representation of the process). Researchers typically use databases such as PubMed, Scopus, and Google Scholar to retrieve candidate studies using keyword combinations and Boolean logic. While effective at surfacing literature, these tools often return large volumes of results, many of which are irrelevant or fail to meet key inclusion criteria. Researchers must then manually sift through each title and abstract to identify eligible studies – a task that can take weeks or months depending on the scope.

This manual task creates a bottleneck imposing significant limitations. First, systematic reviews are expensive, requiring substantial human labor, often from domain experts. Second, they are prone to becoming outdated quickly, as the slow pace of screening cannot keep up with the rapid growth of new research publications. Finally, the time cost scales linearly with the amount of literature – there is no reusability or compounding efficiency in human labor as the field expands. This makes it increasingly difficult to maintain comprehensive and timely reviews.

To address these challenges, our project introduces a natural language processing (NLP) tool designed to automate the initial stage of the systematic review process: research paper classification. Given a citation list exported from a search engine and a set of user-defined inclusion criteria, the tool classifies each citation as either relevant or irrelevant. These criteria may include factors like sample size thresholds, population characteristics, or study design

constraints – elements that are difficult or impossible to specify using standard search syntax. Our goal is to augment them – helping reduce tedious manual work, improve consistency in inclusion decisions, and speed up the overall review timeline. By automating the most repetitive and low-level judgments, researchers can focus on interpreting findings and assessing study quality.

2. Related Work:

Norman et al. (2020) develop a logistic regression ranker trained on TF-IDF features with bag-of-n-gram inputs to automate the abstract screening phase of systematic reviews. They evaluate the effect of using training labels drawn from different stages of the review process, and specifically, whether labels from the title and abstract stage are as effective as those from full-text review. Their results indicate minimal performance loss when relying on the just titles and abstracts, suggesting that coarse early-stage labels are sufficient for model training. Their best model uses trigram features and achieves an AUC of 0.839. They also investigate the role of ambiguous labels, showing that incorporating "maybe" decisions as positive training examples can effectively supplement the training set, especially when data is limited.

Goldfarb-Tarrant et al. (2020) present an end-to-end pipeline for automating systematic reviews in veterinary science. Their work spans all 3 steps of systematic reviews: document scraping, binary classification, and data extraction. Specifically, for the classification step, they use a Support Vector Machine (SVM) trained on TF-IDF vectors. With only two weeks of medical annotations used as training data, their system achieves 88% accuracy while reducing review time to 20 minutes per 100 documents. They further observe that even with a small training set of around 100 documents, the model achieves high recall early on (above 80%), and that additional training data primarily improves the precision-recall balance rather than overall recall.

Frunza et al. (2010) propose a per-question classification strategy that reflects how human reviewers make inclusion decisions in systematic reviews. Rather than assigning a single global relevance label, their system models the screening process by training a separate classifier for each inclusion/exclusion question used in the review protocol. For each question, they train a Complement Naive Bayes (CNB) classifier on a targeted dataset. This approach is applied to a highly imbalanced corpus with a 1:5.6 relevant-to-irrelevant ratio, where CNB is selected for its superior handling of skewed distributions. Although they experimented with other classifiers, CNB consistently produced better results. The per-question method achieved extremely high recall, peaking at 99.2% using a four-vote ensemble, although precision in that configuration was comparatively low at 15.6%.

Rezapour et al. (2020) develop a classifier for detecting the societal impact of publicly funded research projects using supervised learning models trained on annotated project reports. Their system uses TF-IDF vectorization with unigram, bigram, and trigram features, achieving the best F_1 scores of 78.81% using trigram features. Initial baseline models using only unigrams performed notably lower at 52.95%. This indicates that expanding to higher-order n-grams significantly improved performance, particularly recall. Combining lexical, syntactic, and domain-specific features yielded the best overall performance, with ROC scores reaching over 80.8%. They also use Synthetic Minority Over–sampling TEchnique (SMOTE) to enhance performance by addressing the data imbalance in the input data for classification training.

Gutiérrez et al. (2020) benchmark several models for multi-label classification on the LitCovid corpus. Traditional classifiers, including logistic regression and linear SVMs, are trained on TF-IDF weighted bag-of-words features, and compared to pre-trained models such as BioBERT. Despite the dominance of deep contextual models (BioBERT achieving 86% micro-F_1), logistic regression and SMV remain strong, with F_1 scores of 81.1% and 83.4% respectively. The authors highlight that these traditional models remain a strong, computationally efficient baseline when training data is scarce.

While prior research shows that TF-IDF-based models like logistic regression and SVM can effectively automate the abstract screening stage of systematic reviews, existing systems tend to focus on isolated components and often omit key strategies for domain adaptation, performance tuning, and recall-sensitive design. Few integrate domain-specific inclusion criteria, robust handling of data imbalance, or optimized feature selection beyond basic n-grams. We address this gap by blending the most effective elements across the literature: from Norman et al., we adopt trigram-based TF-IDF features and logistic regression on titles and abstracts; from Goldfarb-Tarrant et al., we incorporate SVMs and anticipate the importance of managing the precision-recall tradeoff; from Frunza, we take inspiration from exclusion-based logic and use Complement Naive Bayes for skewed data; from Rezapour, we implement SMOTE and combine lexical with domain-specific features; and from Gutiérrez et al., we confirm that linear models remain effective under limited data. Our system brings these insights together into an NLP screening tool for systematic reviews that aims to support user-defined criteria, balance performance and interpretability, and improve overall recall without sacrificing practical usability.

3. Dataset

The initial set of publications was generated by running Boolean search queries tailored to the specific topic of the systematic review: “Review of methods used to report brain arteriovenous malformation obliteration rates post-stereotactic radiosurgery in adults.”.

The dataset consists of 2,175 PubMed publications, each manually labeled as either relevant or irrelevant by a medical student. To ensure consistency in labeling, one of our team members also reviewed the publications. and we discussed edge cases to reach consensus. For borderline cases, we chose to label them as relevant, following the approach suggested by Norman et al. (2020), who found that including ambiguous examples as positives can improve recall and supplement limited training data.

The binary labels of “relevant” or “irrelevant” are intended to determine whether a given medical publication is relevant or irrelevant to the medical researcher’s systematic review study. More specifically, the annotators were classifying a given publication suggested by the search engine based on two factors. (1) Relevance or irrelevance based on the domain-specific keyword queries. For example, if the search words “Radiosurgery” or “Intracranial arteriovenous malformation” were present in the title or the abstract, this indicated relevance. (2) Relevance or irrelevance based on real-world inclusion and exclusion criteria as delineated by our medical collaborator. For instance, any study with

fewer than 35 patients was marked irrelevant. Similarly, studies involving patients under 18 years old were automatically excluded.

Thus, the articles labeled "relevant" not only matched the initial search terms, but also passed through a second layer of judgment-based filtering that captures what PubMed alone cannot. This means our model's task isn't just about identifying keywords, rather it's about replicating the nuanced decision-making a human expert would use to screen each article individually.

4. Methodology & Implementation

We designed a high-recall classification pipeline for screening systematic review abstracts, combining n-gram-based lexical modeling with expert-informed feature extraction and rigorous threshold calibration. The pipeline is built using modular scikit-learn components, allowing easy substitution and parameter tuning across preprocessing, feature representation, and classification.

4.1 Pipeline Architecture

Our core architecture consists of a Pipeline that merges the title and abstract fields using a custom TextCombiner, applies TF-IDF vectorization, and fits a linear SVM classifier. TF-IDF encodes relative term importance while reducing the influence of common tokens, and bigrams were used to capture short multi-word expressions indicative of relevance (e.g., "case report").

To identify the best-performing configuration, we conducted a comprehensive grid search over both TF-IDF feature parameters and multiple classifier types, including SVM, logistic regression, Complement Naive Bayes (CNB), and cosine similarity. Each classifier was paired with its own hyperparameters. For SVM, this included regularization strength ($C \in \{0.01, 0.1, 1, 10, 100\}$), kernel type, and gamma for RBF kernels. The TF-IDF parameter space included vocabulary size (max_features: 5000, 10000, 20000), n-gram ranges (ngram_range: (1,2) and (1,3)), and document frequency thresholds (min_df: 2, 3, 5; max_df: 0.85, 0.9, 0.95).

This yielded 120 TF-IDF combinations. To account for differences in model complexity and learning mechanisms, we defined classifier-specific hyperparameter grids. For instance, the SVM grid included regularization strength ($C \in \{0.01, 0.1, 1, 10, 100\}$), kernel type (linear, rbf), and gamma (scale, auto) for RBF kernels. These hyperparameters were chosen based on common practice and values reported in prior studies (Norman et al., 2020; Gutiérrez et al., 2020).

Each classifier was evaluated across all 120 feature configurations and its own hyperparameter grid, resulting in approximately 600 model configurations per classifier and over 3000 model fits in total, each validated using 5-fold stratified cross-validation. This is consistent with protocols used in prior studies on systematic review automation (e.g., Norman et al., 2020; Rezapour et al., 2020). We optimized all models for validation F_1 score, which is standard in biomedical and NLP classification tasks (e.g., Gutierrez et al., 2020; Cohen et al., 2006).

SVM with a linear kernel and bigram TF-IDF ultimately outperformed the alternatives under high-recall constraints and was selected as the final model. The most computationally intensive configuration, SVM with stemming taking approximately 24 minutes to complete. All models

were evaluated using a stratified 80-10-10 train-validation-test split, maintaining class balance across partitions.

4.2 Class Imbalance and Normalization

To mitigate the effect of class imbalance, where irrelevant abstracts significantly outnumber relevant ones, we used SMOTE (Synthetic Minority Over-sampling Technique) during training (Rezapour et al., 2020). This was combined with class_weight='balanced' in the SVM to ensure robust decision boundaries.

We compared three normalization modes: raw text, stemming, and lemmatization. Each normalization strategy was evaluated independently through grid search to avoid confounded effects.

4.3 Threshold Calibration for High Recall

Because our application requires minimizing false negatives, we calibrated the model's decision threshold to guarantee a minimum recall of 95% on the validation set. Using predicted probabilities from the classifier, we computed the precision-recall curve and selected the lowest threshold that achieved ≥95% recall. Among those candidates, we selected the one that also maximized precision. If no threshold reached the target, we defaulted to the lowest threshold available to maximize sensitivity. This approach corrected earlier issues with cross-validation based thresholding and ensured alignment with best practices for high-recall biomedical applications.

Because our application requires minimizing false negatives, we calibrated the model's decision threshold to ensure a minimum recall of 95% on the validation set. We computed the precision-recall curve from classifier probabilities and selected the lowest threshold that satisfied the recall constraint. Among qualifying thresholds, we chose the one that maximized precision. If no threshold met the requirement, we defaulted to the lowest threshold available. This threshold was then fixed and used for all test set evaluations.

Although our pipeline targets high recall operationally, we selected models using F_1 to balance sensitivity and precision during tuning. This mirrors prior work where fixed-recall metrics are supplemented with F-measure to estimate overall utility (Cohen et al., 2006).

4.4 Domain-Specific Feature Extraction

In addition to lexical patterns, we engineered a set of binary domain-specific features based on inclusion and exclusion criteria defined by a medical researcher. These were extracted using a custom InclusionExclusionTransformer that applies regular expression patterns to detect: population constraints (e.g., pediatric or pregnancy mentions, age under 18; study designs (e.g., case studies, RCTs, meta-analyses); and malformation types and interventions (e.g., gamma knife, occlusion rate); and sample size indicators (e.g., studies with <30 or <10 patients). See Appendix 3 for a comprehensive list of what was used.

Each matched criterion is converted into a binary feature, yielding a 12-dimensional feature vector. These features were concatenated with the TF-IDF vectors and served as input to the final classifier. Features can also function as broad filters (supporting model input) or hard exclusion rules, depending on user configuration.

4.5 Model Interpretability

To support interpretability in clinical workflows, we extract the top-weighted features from the final linear SVM model. Using the learned coefficients (.coef_), we identify terms most predictive of relevance and irrelevance. These are visualized using bar plots, allowing researchers to inspect which textual cues influence predictions (see Appendix 2 for an example). This ensures the system is not only performing as desired, but also transparent.

5. Evaluation & Results:

We evaluated our system through a structured series of experiments comparing multiple classifiers, normalization strategies, feature configurations, and threshold calibration schemes. We evaluated using our test dataset which contained 218 research papers. All models were assessed in two operational modes: a balanced model, optimized for F_1 score, and a high-recall model, optimized to ensure a minimum of 95% recall on the validation set. Evaluation metrics included precision, recall, F_1, F_2, ROC-AUC, and WSS@95 (work saved over random sampling at 95% recall). These metrics allow us to assess both general screening performance and utility in recall-sensitive contexts such as medical literature review.

5.1 Classifier Comparison

We evaluated four classifiers – Support Vector Machine (SVM), Logistic Regression, Complement Naive Bayes (CNB), and Cosine Similarity – on their ability to distinguish relevant from irrelevant research papers in a balanced screening configuration. For each model, we extracted TF-IDF features over raw text (i.e., titles and abstracts) and ran a 5-fold cross-validation grid search over key hyperparameters, including n-gram range (unigrams only; unigrams+bigrams; and unigrams+bigrams+trigrams), document-frequency thresholds, and model-specific settings.We selected the settings that maximized balanced F_1.

Among all models, SVM demonstrated the strongest overall performance, achieving the highest F_1 score (0.6780), recall (0.8333), and ROC-AUC (0.9565). This indicates that SVM strikes the best balance between sensitivity and precision, making it the most effective model for general-purpose systematic review screening.

Logistic Regression performed slightly worse in terms of F_1 (0.6296) but achieved the highest WSS@95 score (0.8124), suggesting it is especially well-suited for reducing manual screening effort at high-recall thresholds. While it lags slightly in recall (0.7083), its strong ranking quality makes it a viable alternative in practical workflows.

Complement Naive Bayes, consistent with prior findings, delivered the highest raw recall (0.9167) but suffered from low precision (0.3333), resulting in the lowest F_1 (0.4889). This

makes it appropriate only in recall-critical settings where false positives are tolerable or can be filtered post hoc.

Cosine Similarity served as a non-parametric baseline. While it performed better than CNB in terms of F_1 (0.5424), it was outperformed by both linear classifiers in all major metrics, indicating the advantage of supervised learning even with relatively simple feature sets.

Interestingly, we found that restricting TF-IDF to unigrams + bigrams (1–2 grams) yielded roughly a 1 percentage-point boost in balanced F_1 for SVM, Logistic Regression, and Cosine Similarity, whereas extending to trigrams (1–3 grams) was most beneficial for Complement Naive Bayes – giving a 0.9 pp gain over the bigram setting.

In summary, SVM emerged as the most effective and reliable model, especially when balanced performance across all metrics is required. Logistic regression may offer better screening efficiency, and CNB can serve recall-sensitive use cases, but with the caveat of low precision.

5.2 Text Normalization Strategies

To assess the effect of text normalization, we evaluated three preprocessing modes – raw text, lemmatization, and stemming. We fixed our SVM to the optimal hyperparameters discovered in the initial grid search (TF-IDF vectorization with ngram_range=(1,2), max_df=0.85, and max_features=5000). Each variant was evaluated independently in both balanced and high-recall modes, with performance compared using F_1 score, ROC-AUC, and WSS@95.

The results show that raw text consistently outperformed both normalization techniques in the balanced configuration, achieving an F_1 of 0.6780 and the highest ROC-AUC of 0.9565 (see Appendix 4). When SMOTE was added, performance further improved, with SVM + SMOTE yielding the best balanced F_1 (0.7083) and highest WSS@95 (0.8399), while ROC-AUC remained relatively equal at 0.954, confirming its robustness in class-imbalanced settings (see Appendix 5)..

Under high-recall constraints, stemming offered a slight advantage over other normalization strategies. SVM with stemming achieved the highest high-recall F_1 (0.6216) and WSS@95 (0.7206), indicating its potential value in recall-sensitive applications. However, this came at the cost of lower balanced F_1 (0.6333) compared to raw text.

Lemmatization underperformed in both balanced and high-recall modes, producing lower F_1 scores and WSS@95 than either raw or stemmed input. These results suggest that in this domain, simpler raw text performs best overall, and stemming may only offer marginal gains when optimizing for high recall.

These results indicate that, once optimal hyperparameters are applied, raw unigrams+bigrams remain the most reliable overall, while stemming can be employed when maximizing recall is paramount.

5.3 Domain-Specific Filtering

To incorporate expert criteria, we transformed each exclusion criterion into binary features, such as study design flags, age thresholds, and intervention types. These were implemented as a 12-dimensional vector and combined with TF-IDF features to enhance model specificity. We tested three configurations: SVM with criteria features, SVM with stemmed text and criteria, and SVM with both criteria and MeSH-derived terms.

In the balanced setting, adding criteria features improved the baseline SVM's F_1 score from 0.6780 to 0.7547 and ROC AUC from 0.9565 to 0.9570, confirming that even sparse rule-based features can offer meaningful gains. The SVM with stemmed text and criteria further improved recall-specific metrics, while maintaining strong F_1 (0.7308) and AUC (0.9592). However, the best ROC AUC in this setup was achieved by the full model incorporating text + criteria + MeSH terms, reaching 0.9680, although this configuration underperformed on WSS@95 (0.7015), reflecting diminished ranking efficiency.

In the high-recall mode, SVM with criteria and stemmed text achieved a F_1 score of 0.3966 and an F_2 score of 0.6117, indicating better tradeoffs between sensitivity and review burden than the baseline. Notably, adding MeSH features boosted recall stability but did not dramatically increase F_1, reinforcing that such features act more as soft filters rather than major classifiers. This improvement may also reflect the baseline model's ability to capture key exclusionary terms, such as "pediatrics" or "children," which frequently appeared among the most strongly negative coefficients in the final linear model (see Appendix 2).

Ultimately, while criteria features were sparsely activated, their inclusion nudged the model toward more consistent exclusion decisions, especially in borderline cases. There was a steady increase in ROC, implying that the model is performing better at ranking documents, but also that the model does not translate to better binary cutoffs for the F1 score and recall metrics. However, the impact of the domain-specific criteria was limited due to their small relative weight in the feature space, which was still dominated by thousands of TF-IDF dimensions. The clearest gains were observed in hybrid configurations (SVM + stemming + criteria), which best balanced interpretability, sensitivity, and recall-specific efficiency.

6. Conclusion

This project systematically examines and evaluates the classifier architectures and hyperparameter choices most commonly reported in the systematic‑review literature: spanning n-gram ranges in TF-IDF, regularization settings for SVM, oversampling with SMOTE, and expert‑driven exclusion criteria. We used a novel corpus of 2175 medical titles and abstracts. In response to the lack of consensus across the literature, we explored a comprehensive grid of lexical, normalization, sampling, and domain‑feature combinations.

Our final pipeline blends TF-IDF lexical modeling (optimized on unigrams + bigrams), a linear SVM classifier, SMOTE for class imbalance, and lightweight exclusion-criteria flags, yielding balanced F1 of 0.7083 and sustained recall ≥95% in high‑recall mode. SVM emerged as the most robust model, Logistic Regression offered superior screening efficiency (WSS@95), and

Complement Naive Bayes remains a viable option when extreme recall is required. These results demonstrate that simple, interpretable models, when tuned and combined thoughtfully, can rival more complex alternatives while remaining modular enough for further extension (e.g., by integrating ontology features or semantic embeddings alongside TF-IDF). Ultimately, this project supports the broader goal of making evidence synthesis more efficient, reproducible, and scalable in modern medicine.

8. Future Work

While our current system effectively automates the initial screening phase of systematic reviews using lexical and domain-specific features, several directions remain for future development.

First, integrating deep contextualized language models such as BioBERT or PubMedBERT could improve classification performance, especially in cases where relevance depends on subtle semantic cues not captured by n-gram-based models. However, adopting these architectures would require balancing gains in accuracy against increased computational cost and reduced interpretability – particularly important in clinical research contexts where transparency is essential.

Second, we plan to extend the current rule-based filtering layer into a fully configurable module, enabling researchers to define custom inclusion and exclusion criteria (such as population characteristics, study designs, and sample size thresholds) through a structured, user-editable schema. This would increase flexibility across research topics and reduce the need for bespoke code edits.

Third, to address class imbalance more effectively, future work will explore advanced data augmentation strategies beyond SMOTE, including generative data synthesis and weak supervision approaches that leverage unlabeled or noisily labeled data.

Finally, we envision incorporating human-in-the-loop feedback to support active learning and iterative retraining. As researchers interact with the system, their feedback could be used to refine the model, improving performance over time and aligning the tool more closely with real-world systematic review workflows.

9. Acknowledgments

We thank Mikaël Devletian for producing the annotated data for the large dataset collection.We are also grateful to Abhishek Upadhayaya and Professor Adam Meyers for their continuous support and guidance throughout the course of this project.

10. References

Bernal Jimenez Gutiérrez, Jucheng Zeng, Dongdong Zhang, Ping Zhang, and Yu Su. 2020. Document Classification for COVID-19 Literature. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3715–3722, Online. Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.332/.

Christopher Norman, Mariska Leeflang, Pierre Zweigenbaum, and Aurélie Névéol. 2018. Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1582/.

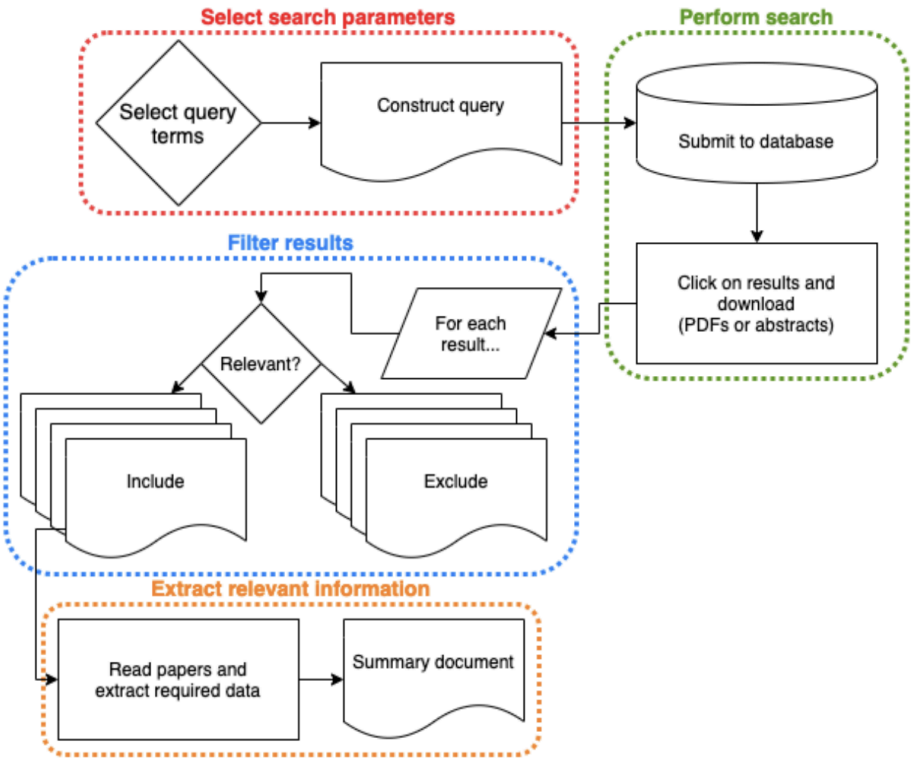
Oana Frunza, Diana Inkpen, and Stan Matwin. 2010. Building Systematic Reviews Using Automatic Text Classification Techniques. In *Coling 2010: Posters*, pages 303–311, Beijing, China. Coling 2010 Organizing Committee. https://aclanthology.org/C10-2035/.

Rezvaneh Rezapour, Jutta Bopp, Norman Fiedler, Diana Steffen, Andreas Witt, and Jana Diesner. 2020. Beyond Citations: Corpus-based Methods for Detecting the Impact of Research Outcomes on Society. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6777–6785, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.837/.

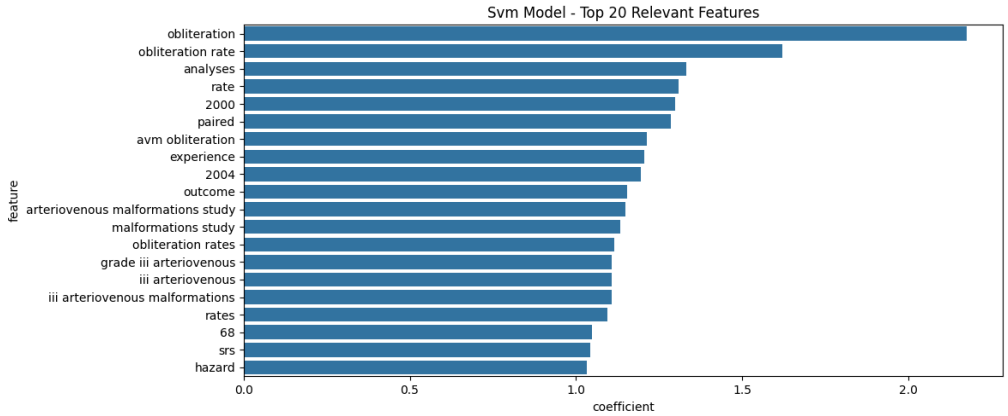
Seraphina Goldfarb-Tarrant, Alexander Robertson, Jasmina Lazic, Theodora Tsouloufi, Louise Donnison, and Karen Smyth. 2020. Scaling Systematic Literature Reviews with Machine Learning Pipelines. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 184–195, Online. Association for Computational Linguistics. https://aclanthology.org/2020.sdp-1.21/.

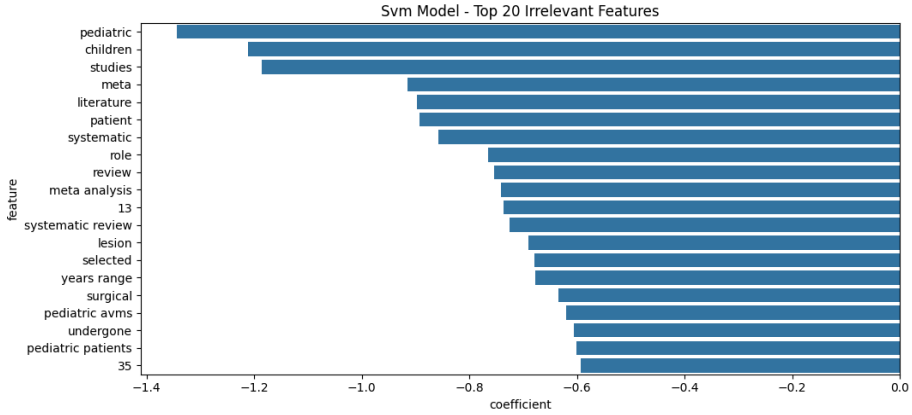
Appendices

Appendix 1: Human-based systematic review pipeline (Goldfarb-Tarrant et al., 2020)



Appendix 2: Barplots for the SVM model (baseline)

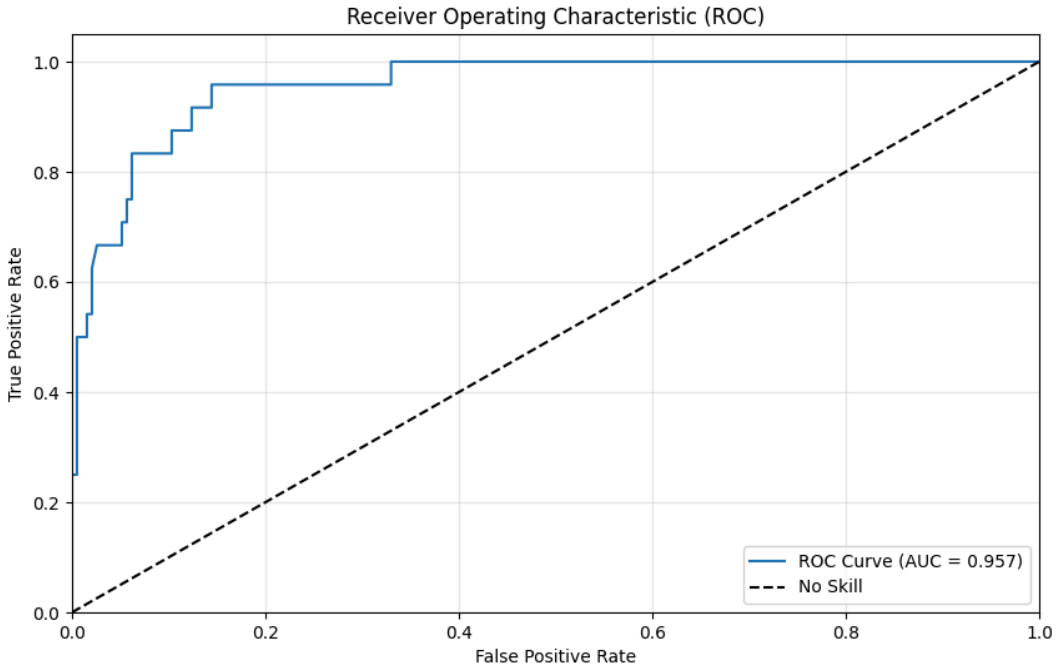




Appendix 3: Inclusion and exclusion criteria implemented in Feature Extraction


```
{
  "publication_year": {
    "include": "Published in 2000 or later",
    "exclude": "Published before 2000"
  },
  "language": {
    "include": "English or French",
    "exclude": "Non-English and non-French articles"
  },
  "population": {
    "include": "All patients are ≥18 years old with brain AVMs",
    "exclude": [
      "Only pediatric populations",
      "Pregnant patients",
      "Patients with dural arteriovenous fistulas",
      "Patients with pial arteriovenous fistulas",
      "Patients with vein of Galen malformations",
      "Patients with cavernous malformations"
    ]
  },
  "age_mention_in_text": {
    "include": "No mention of patients under 18",
    "exclude": "Any mention of patients under 18 (e.g., 'age range 4–98')"
  },
  "study_design": {
    "include": [
      "Randomized controlled trials",
      "Clinical trials",
      "Cohort studies",
      "Case series",
      "Systematic reviews"
    ],
    "exclude": [
      "Case studies with fewer than 10 patients",
      "Meta-analyses",
      "Literature reviews"
    ]
  },
  "outcome_reporting": {
    "include": "Mentions methodology for AVM occlusion rate reporting",
    "exclude": "No mention of AVM occlusion rates"
  },
  "intervention_exposure": {
    "include": [
      "Gamma Knife",
      "CyberKnife",
      "Novalis",
      "Linear accelerator-based radiosurgery"
    ],
    "exclude": "Treatment with other methods"
  },
  "automatic_exclusion_keywords": {
    "exclude": [
      "hypofractionated",
      "proton beam therapy",
      "fractionated stereotactic radiotherapy",
      "fractionated stereotactic surgery",
      "tomotherapy"
    ]
  },
  "patient_minimum": {
    "include": "No patient number specified → keep",
    "exclude": "If specified and <30 patients → exclude"
  }
}
```

Appendix 4: ROC graph for SVM Baseline



Appendix 5: ROC graph for SVM+SMOTE