Classify Sentences from Scientific Research Articles by Rhetorical Categories

Yuanxi Fu

IS567TM Fall 2020

1. Introduction

Research articles in experimental science often follow specific rhetorical structures. The authors start by providing a summary of the background for the research. They then state their objectives, outline their methods, report their results, discuss the results and finally, state their conclusions. Therefore, articles can often be clearly sectioned and labeled according to the rhetorical functions. Structured abstracts in PubMed also adopted similar section headings (e.g., background, objective, method, results, or conclusions).

I am interested in developing a classification pipeline that can tell whether a sentence is about background, objective, method, results, or conclusions. Being able to do so will bring fundamental changes to how we use scientific literature. For example, methods often limit how well we can trust the conclusions. While researchers want to know the conclusions, they are equally interested in the methods. Also, in a study I did with Prof. Jodi Schneider, we found that citations in sentences about methods and discussions can play a critical role in supporting the arguments of the paper [1]. We named those citation “keystone citations,” because if the cited papers lost validity, citing article’s arguments would crumble. A sentence-based classification pipeline will function as the first pass on screening “keystone citations,” helping construct a validity chain among research papers.

Several studies have attempted to develop such a classification pipeline, and many used section headings as a proxy to the actual rhetorical categories [2]–[5] because they can be automatically generated. Yet such proxy labels may not be accurate at the sentence level. Fortunately, I got hold of a manually labeled dataset from Prof. Halil Kilicoglu, and this project is also an opportunity to evaluate whether models trained on manual labels will outperform models trained on automatically generated labels.

For this project, I hope to answer to the following research questions:

RQ1: Build my own classification models and understand factors that can influence their performance.

RQ2: Will models trained on manual labels outperform models trained by automatically generated labels?

RQ3: Can the models trained on abstract sentences be applied to citation context sentences? How bad will that be?

2. Related Work

Yamamoto & Takagi built supported vector machine classifiers for each of the five rhetorical categories: background, purpose, method, result, and conclusion [5]. They used three datasets: Structured-1, Structured-2, and Unstructured. Structured-2 is randomly selected from MEDLINE structured abstracts. Structured-1 is a randomly selected subset (~2%) of Structured-2. Having two training sets was for evaluating the impact of training set size. The unstructured dataset was from unstructured abstracts of MEDLINE and annotated by human experts. Classifiers were trained on Structure-1 and Structured-2, using section headings as labels, and then tested on the unstructured dataset. The authors tested three feature combinations: Original, SPTV and Combined. The "Original" set of features includes *sentence positions, tense, tf-idf value, presence of auxiliary verb, χ^2 value of terms, χ^2 value of collocation, and χ^2 value of subject-verb pairs.* SPTV set includes only sentence position and term vector, which were used by a previous study. The "Combined" set includes both the "Original" and the "SPTV" set. The three models were thus named "SeCBLiS original," "SPTV" and "SeCBLiS combined." Unsurprisingly, SeCBLiS combined had the best F-scores, and SeCBLiS original is better than SPTV. One shortcoming of SeCBLiS original is its susceptibility of tense because authors' preference of tense varies from one to another. Training set size effect between Structured-1 and Structured-2 (91360 sentences vs 1673 sentences) was observed in method, result and conclusion classifiers, but not in background and purpose classifiers.

Guo et al [6] tested three annotation schemes for biomedical abstracts: those based on section names (S1), argumentative zones (S2), and core scientific concepts (S3). S1 is the closest to my target classes, except that S1 does not have the "background" class. The authors annotated a dataset taken from the cancer risk assessment (CRA) abstracts and found a certain degree of association between the three schemes. The resulted dataset contains 7985 sentences, comparable to my dataset. They also built classifiers using Naive Bayes (NB) and Support Vector Machines (SVM). Because SVM performed considerably better than NB, the paper only reported results from SVM classifiers. Features selected by the authors include *history (i.e., category assigned to the previous sentence), location, word, bi-gram, verb, verb-class (based on an earlier study), POS, grammatical relationships (GR), subject, object and voice*. To understand each feature's impact on model performance, the authors also tested single-feature models and all-but-one-feature models. For single-feature models, word, bi-gram, and verb performed the best, whereas *history* and *voice* performed the worst. For all-but-one-feature models, S1 and S2's performance declined the most from missing the *location* feature, and S3 deteriorated the most from missing the *word* or the *POS* feature. And the best combination of features is all-but-verb.

Nam et al. [3] developed classifiers to reformatting of unstructured abstracts into the IMRAD format. They constructed three datasets from PubMed central open-access subset: one from structured abstract and used for training (SA, 23881 sentences), one from unstructured RCT abstracts (UA1, 1786 sentences), and one from unstructured general abstracts (UA2, 2429 sentences). For SA, the labels were derived from the NLM mapping list to one of the five standard headings (i.e., Objective, Background, Methods, Results, and Conclusion). The "Objective" and "Background" class were then combined to an "Introduction" class and "Conclusion" was renamed to "Discussion." The authors grouped the features selected into four categories: *bag-of-words (B), linguistic features (L), grammatical features (G), and structural features (S)*. BOW (B), grammatical (G) and structural (S) features have been used before in other studies. The innovation of this paper is the construction of linguistic features. For each linguistic feature (e.g., n-gram, verb phrase, noun phrase), the authors first calculated the TF-IDF weight in each class (4 classes in total) of this linguistic feature to create a 4-D weight vector. Then, they calculated the weighted vector for each linguistic feature of a sentence, added all the vectors together, and converted this sum vector to a rank order vector. The rank order vector ranks the likelihood of one sentence belonging to a class from high to low. Their results showed that introducing linguistic features could achieve good accuracy without substantial computational cost. For example, the BGSL combination (1550 dimensions) still had the best accuracy (91.70%) in structured abstracts. However, the GSL combination also achieved 91.00% accuracy with only 75 dimensions, and SL achieved 90.3% accuracy with only 14 dimensions. And F-score of SL were only marginally below those of GSL and BGSL in all IMRAD sections.

De Waard & Pander Maat [7] studied how tense influenced human perception of a sentence's discourse role. They recruited 12 participants to classify sentences into one of the following categories: Fact, Results, Hypotheses, Method, Problem, and Implication. The sentence's content was unchanged when presented to the participants, but its main verb took one of the following tenses: present, past, or modal present. They found evidence that people did associate tense with the discourse roles. Present tense increased the chance of Fact prediction, past tense Result prediction, and the modal verbs Hypothesis prediction. However, the classification of Result sentences was less susceptible to tense variation than others, potentially because they contained words that have "experimental" characteristics. The authors also pointed out the challenge for nonnative authors to adhere to the rules. Yamamoto and Takagi's study also found that using tense as a feature could throw the model off, and the reason may be explained by inconsistent use of tense by a wide range of authors. This is also a baseline study to see whether humans can tell the rhetorical categories apart at single-sentence level. Methods sentences were the most consistently classified (96%), followed by Results (76%), Implication (65%), Fact (64%), and Problem (55%). And Hypothesis sentences were proven to be the hardest to humans, with only 46% of the sentences correctly classified.

Shahriari [8] studied lexical bundles in three main sections (i.e., Introduction, Method, and Results) from 200 papers in applied linguistics. Lexical bundles are phrases made by three or more words and repeatedly appearing in a language, and the study of them is relevant to the n-gram feature in text mining. Setting cut-off frequency at 30 occurrences per million words, the author obtained three sets of lexical bundles, 60 from the Introduction section, 44 from the Method section, and 179 from the Results section. He also studied shared lexical bundles and section-unique bundles. Introduction and Results had the greatest number of bundles in common (27), and Method shared more bundles with Results (14) than with Introduction (1). Also, Results had the most unique bundles, while Introduction had the least.

Agibetov et al. [9] used fastText, a library that implemented shallow-and-wide neural models to classify biomedical text into Objective, Background, Methods, Results, and Conclusions. FastText has a low barrier to use. It does not require a GPU and requires little data-preprocessing and hyper-parameter tuning. The authors used PubMed 20k RCT, PubMed 200k RCT, and an extended corpus as their datasets, with PubMed 200k RCT as the benchmark. They tested two approaches, fully supervised learning and a mix of unsupervised learning followed by supervised learning. The supervised learning approach relied on n-gram representation for training in fastText. The mixed approached used sent2vec first to pre-training N-grams embeddings on the full PubMed 200k RCT training set and then switched to fastText for fully supervised training. Given sufficient data (>50000), the fully supervised approach and the mixed approach generated comparable performance, and the mixed approach had a significant advantage with smaller datasets (<1000). In many of the supervised learning examples reviewed previously, structural information such as sentence location was indispensable. Unfortunately, the fastText models also encounter performance deterioration when sentence context and sentence position were removed. Therefore, single sentence model remains a challenge even for this state of art neural network technique.

3. Dataset and Experiment Design

The dataset includes 5517 sentences from 500 abstracts randomly selected from PubMed. Each sentence was manually annotated to one of the following rhetorical categories: Background, objective, methods, results, and conclusions. As shown in Figure 1, instances are not evenly distributed among the five categories. “Results” has the most instances, followed by “methods.” And “objective” has the least number of instances.

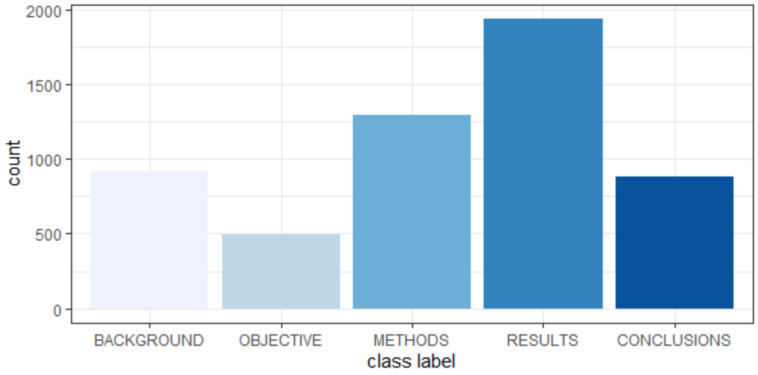

Figure 1. Distribution of instances among five rhetorical categories

Two experiments were conducted. For the first experiment, I intended to compare three different classification algorithms, support vector machine (SVM), naïve bayes (NB), and decision tree (DT). I used the bag-of-words representation. Preprocessing included removal of stop words but no further. Features were selected based on information gain. For experiment 2, the goal is to experiment with new representation strategies, especially to take advantage of formulaic expressions in scientific writings. Three different representation strategies were tested: Bag-of-words with and without punctuation removal and stemming, and trigram representation. Support vector machine was selected as the classification algorithm, and features are still selected based on information gain. The number of features was varied from 100 to 1000 for bag-of-words representation and from 100 to 800 for trigram representation. Four metrics were recorded to evaluate the performance, including accuracy, precision, recall, and F1 scores.

A third experiment evaluated whether models trained on manual label outperform the model trained on automatically generated labels. I continued to use bag-of-words representation with 100 to 1000 features selected by information gain. Two sets of models were created with the same training dataset, one based on manual labels and one based on automatically generated labels. Those two sets of models were subsequently evaluated on a test set with manual labels.

Lastly, a set of classifiers built using bag-of-words representation (1000 terms) without stemming and punctuation removal and support vector machine as classification algorithm was applied to ten citation context sentences (i.e., sentences surrounding citations) extracted from biomedical and chemistry research articles. All except one (sentence 4) are “keystone citation context” [1].

4. Results

4.1 Experiment 1

Results from Experiment 1 are shown in Figure 2 to Figure 6. SVM consistently yields good precision (0.8 to 0.9). Recalls are poor (0.2 to 0.3) for three categories: Background, objective, and conclusions, and decent (around 0.5) for two categories: Results and methods. Background, objective, and conclusions also happen to be the three classes have fewer number of instances. Numbers of features have mild to negligible influence on classifier's performance.

Naïve Bayes achieved better recalls than SVM. However, precision is rather poor (often less than 0.5), which renders good recalls almost useless. Also, in most cases, adding features hurts the recall, but the impact to precision is mild to negligible.

Precision and recall from decision tree classifiers are not so wildly apart as in Naïve Bayes and SVM, and they tend to converge as number of features increases. Objective, methods, and results have decent precision and recall (0.5 to 0.6). Background and conclusions have poor precision and recall (0.2 to 0.4).

Given the results from Experiment 1, SVM is the best choice among the three classification algorithms, and it will become the algorithm for Experiment 2.

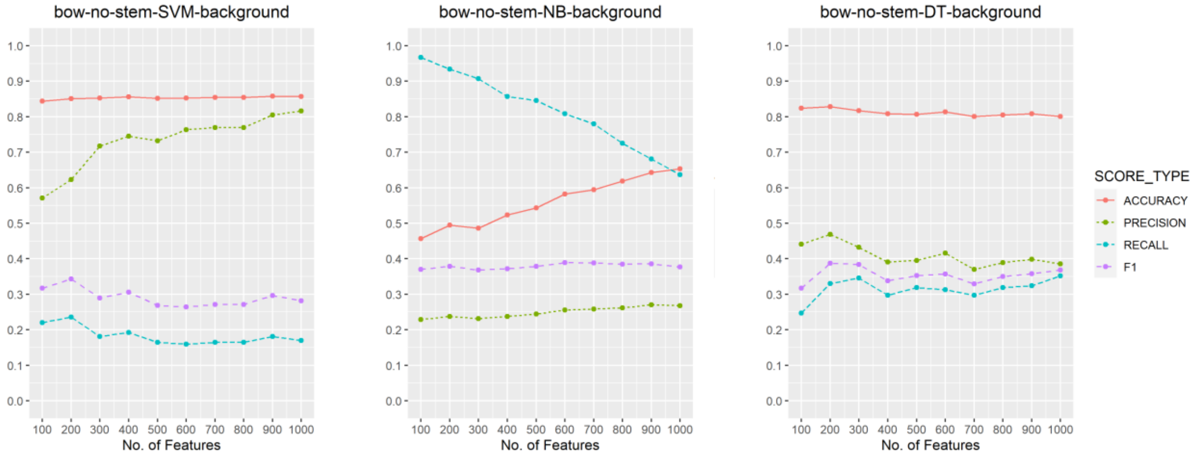

Figure 2. Experiment 1 results for the "background" class

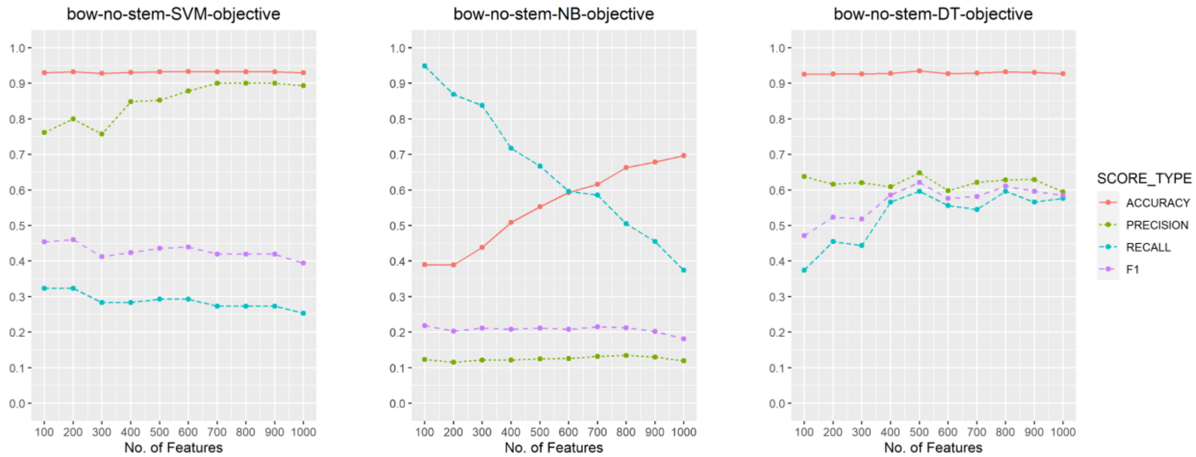

Figure 3. Experiment 1 results for the "objective" class

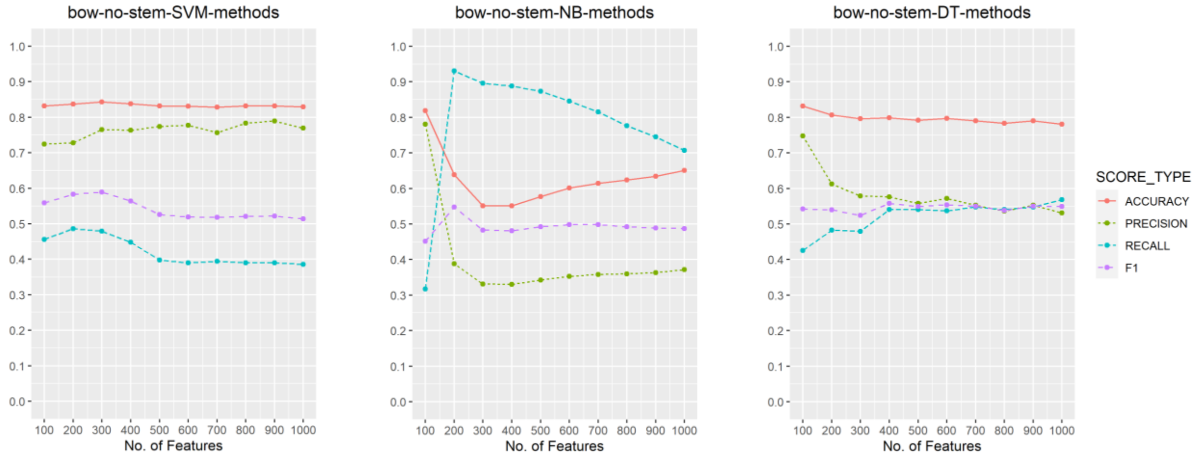

Figure 4. Experiment 1 results for the "methods" class

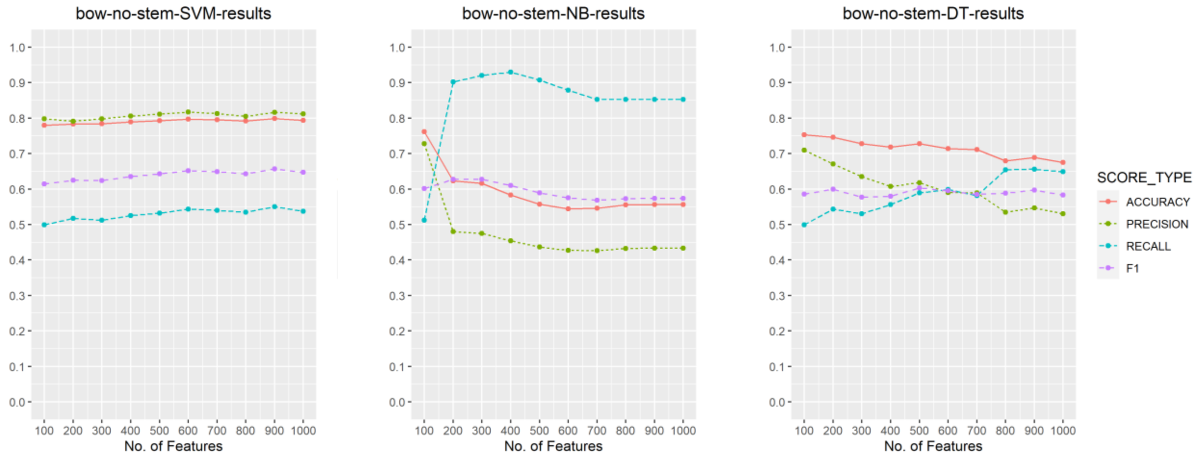

Figure 5. Experiment 1 results for the "results" class

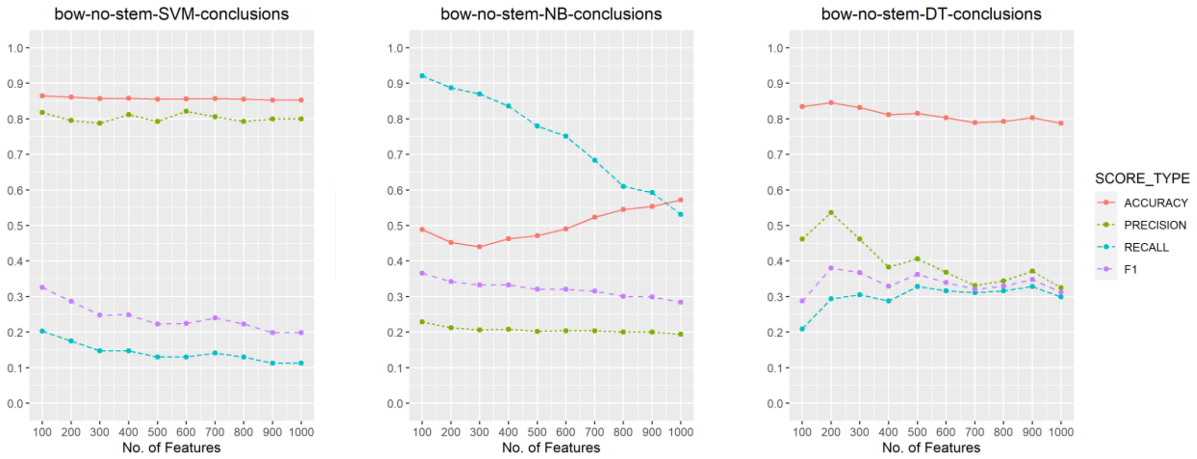

Figure 6. Experiment 1 results for the "conclusions" class

4.2 Experiment 2

Three different representation strategies were tested: Bag-of-words with and without punctuation removal and stemming, and trigram representation. I chose trigram, the three-word cluster, to represent formulaic expressions in scientific writings that cannot captured at the single-word level. Examples of high-ranking trigrams are listed in Table 1.

Table 1. Representative examples of trigrams

Background	Objective	Methods	Results	Conclusions
littl is known	studi wa to	wa used to	the effect of	suggest that the
ha been report	the aim of	wa assess use	p 0001 and	find suggest that
ha not been	studi aim to	use to assess	there were no	are need to
is known about	to evalu the	use to evalu	were use to	result suggest that
been shown to	to investig the	wa carri out	the present studi	suggest that the
is one of	the purpos of	were obtain from	confid interv ci	may be a

Results from Experiment 2 are shown in Figure 7 to Figure 11. Unfortunately, removing punctuations and stemming do not improve classification performance. Poor recall remains a problem.

Trigram representation alone does not perform better than the "bag-of-words" representation. However, there is one exception, the "objective" class. For objection, using 100 trigrams, SVM can achieve 0.7 precision and 0.45 recall, which is better than either of other two representations could achieve. It is interesting because objective is also the class that is supposed to be very formulaic.

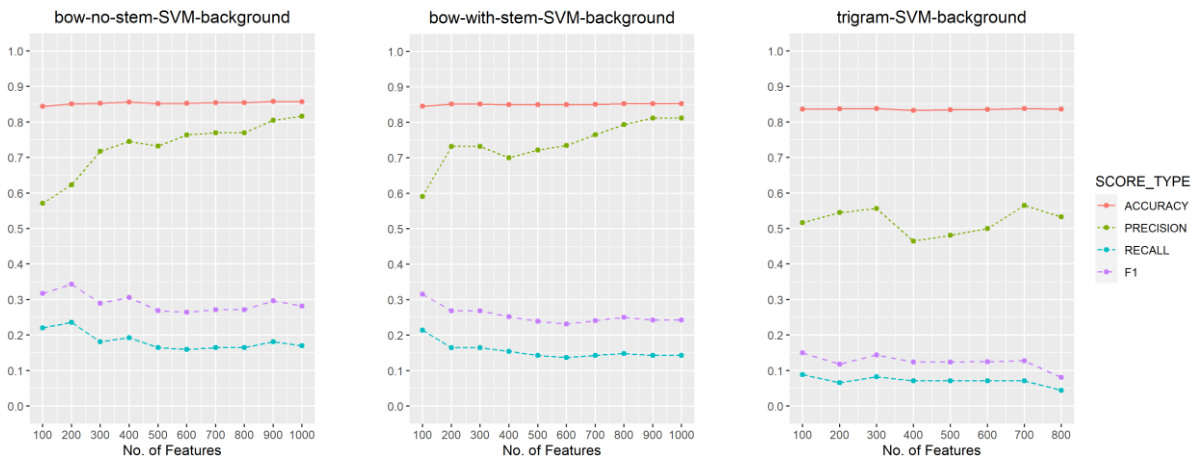

Figure 7. Experiment 2 results for the “background” class

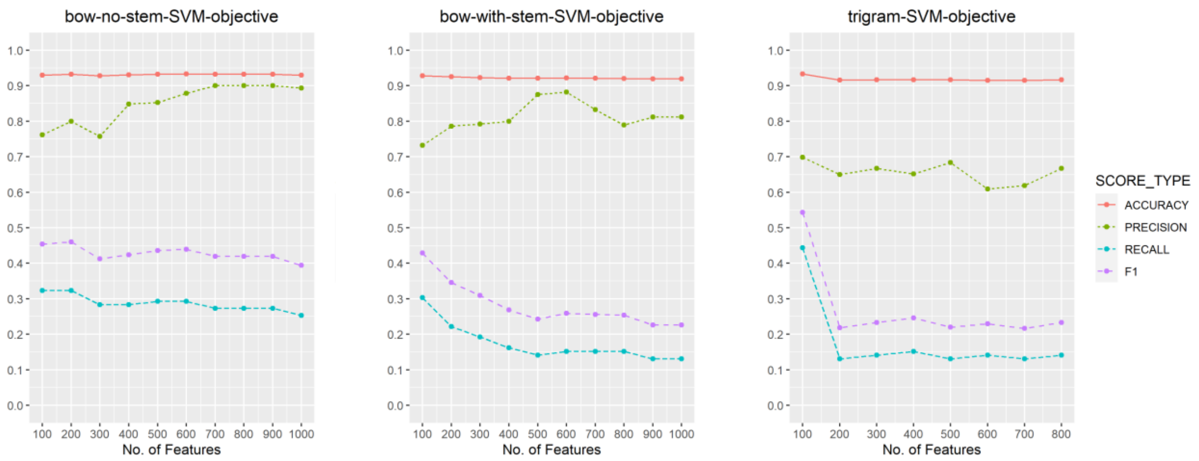

Figure 8. Experiment 2 results for the “objective” class

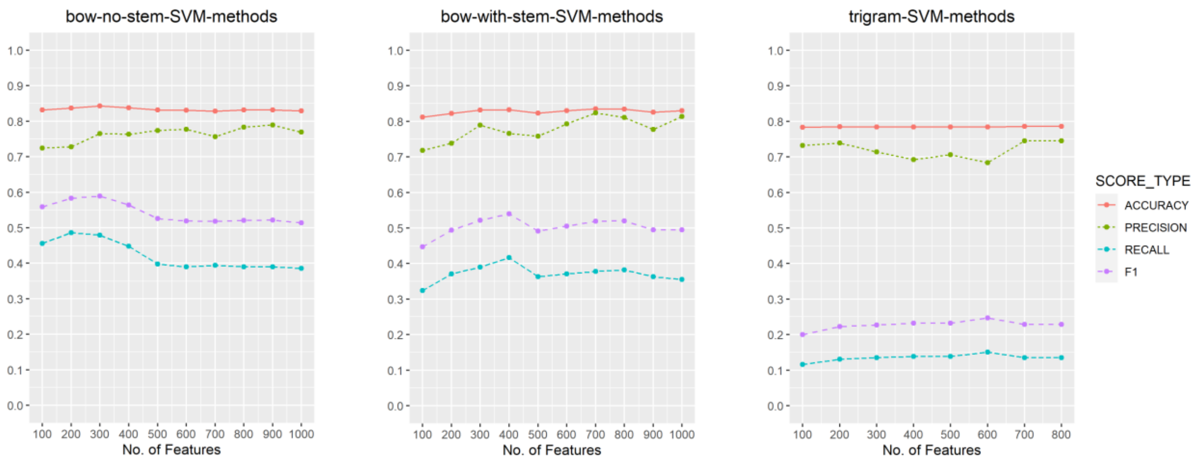

Figure 9. Experiment 2 results for the “methods” class

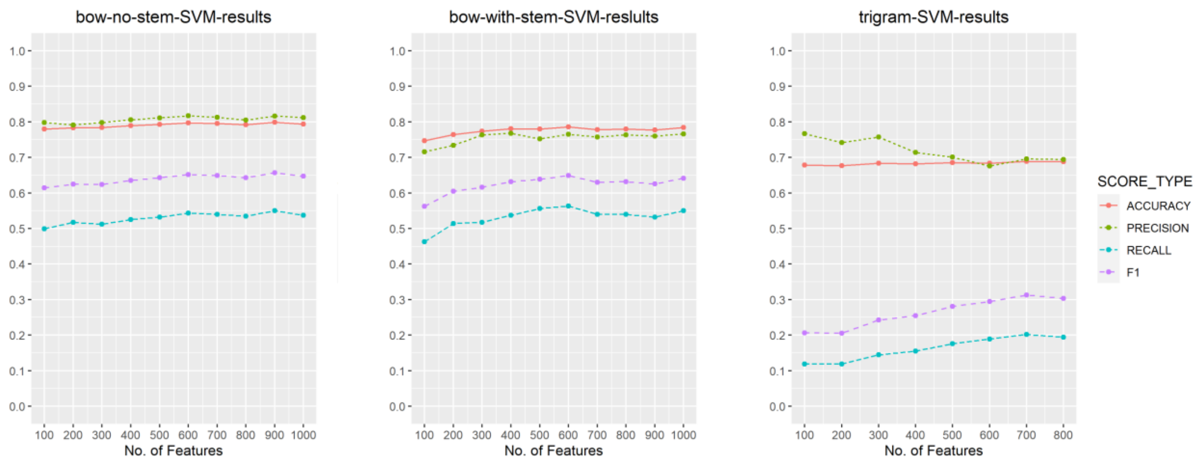

Figure 10. Experiment 2 results for the "results" class

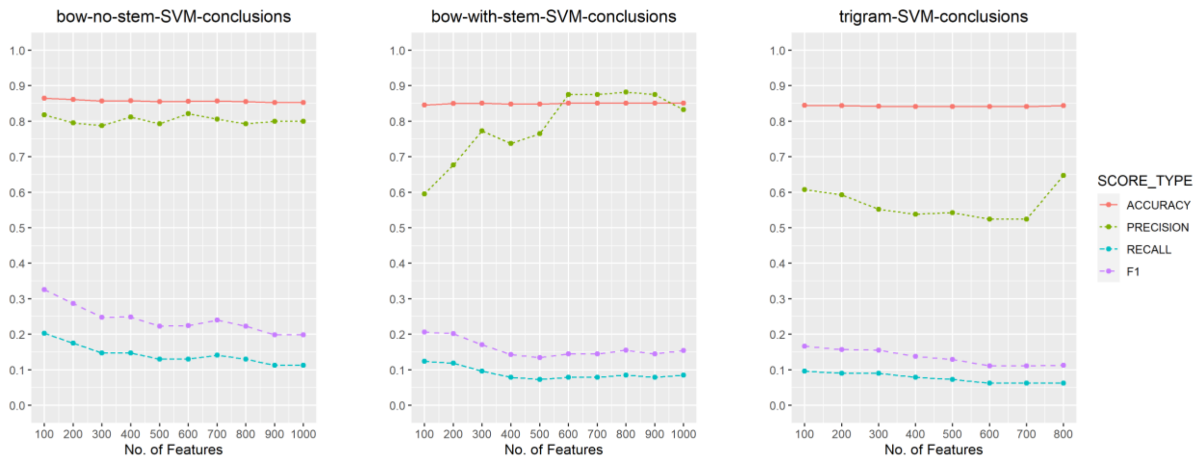

Figure 11. Experiment 2 results for the "conclusions" class

4.3 Error analysis

False negative and false positives were analyzed to identify ways to improve the models. Twenty cases were analyzed, and three categories emerged. They are listed in Table 2. Category 2, blurred boundaries, poses fundamental limits to how well our rhetorical category classifiers can perform, while adding syntax-based features may help us do better with Category 1 and 3.

Table 2. Identified error categories

Category	Example	Causing FN, FP, or both?
Terminology overloading	"A P(II) signal transduction protein, GlnK, is the final regulator of transmembrane ammonia conductance by the ammonia channel AmtB in Escherichia coli."	FN. Because the sentence is full of technical terms, there is not enough signal to help it get classified.

Blurred boundaries between rhetorical categories	“Fluorescein angiography showed multiple nummular hyperfluorescent lesions surrounded by zones of hypofluorescence.”	Both. Some sentences, when taken out of context, cannot easily be determined as one of the rhetorical categories.
Triggered by key terms	“Further study is required to investigate the correlations between the parameters investigated and efficacy of the bleaching process.”	FP. Sometimes, high-ranking terms can trigger a sentence being classified into a wrong category. The example here is clearly a conclusion sentence. However, the term “investigates” probably trigger the misclassification into the “objective” class.

4.4 Experiment 3

Two categories, “background” and “objective” benefited the most by having manually generated labels. For the “background” class, models trained on manual labels yield both better precision and better recall (Figure 12). For the “objective” class, recall is extremely low for models trained on automatically generated labels. Models with 400 to 700 features picked out no true positives, resulting in a recall of 0 (Figure 13). The 1.0 precision scores are also deceptive, because the classifiers only picked out 1 true positive in those cases. For the “methods” class, manual labels improved recall, and the precision remained similar between the two sets of models (Figure 14). For “results” and “conclusions,” performance of the two are comparable (Figure 15 and Figure 16).

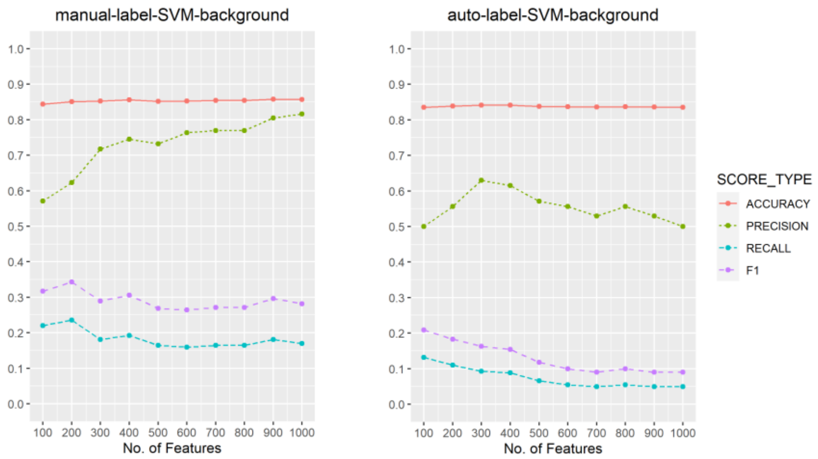

Figure 12. Experiment 3 results for the “background” class

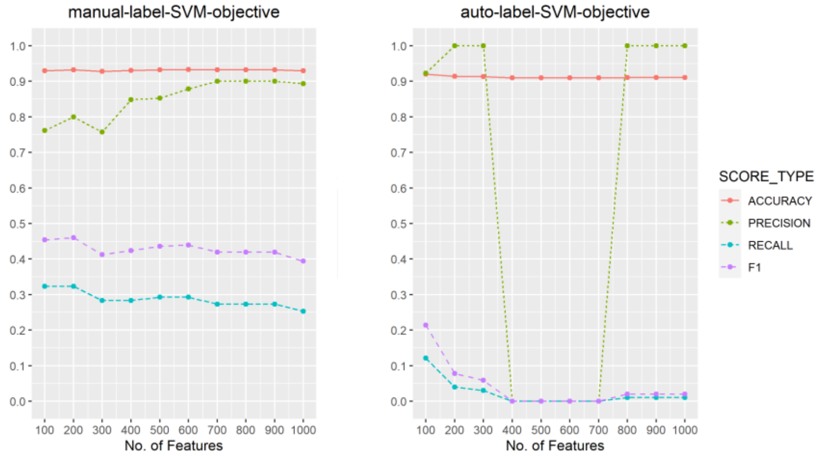

Figure 13. Experiment 3 results for the "objective" class

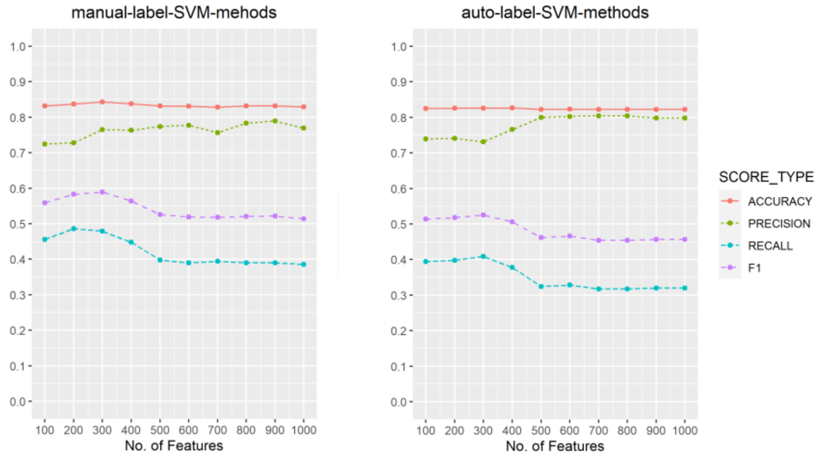

Figure 14. Experiment 3 results for the "methods" class

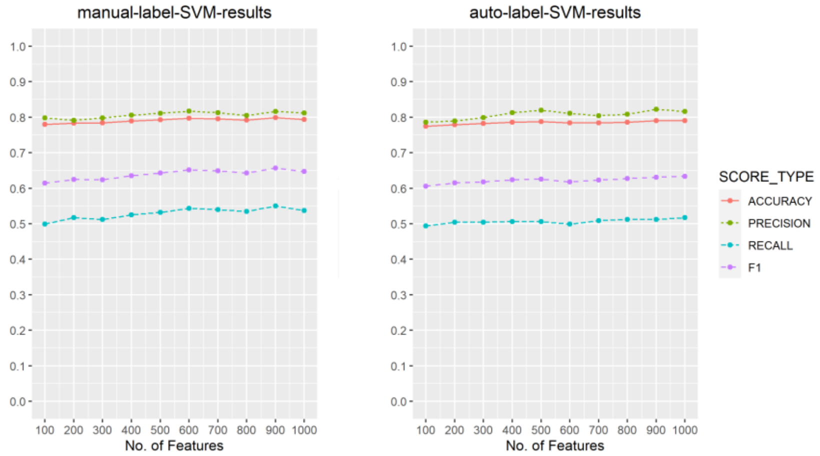

Figure 15. Experiment 3 results for the "results" class

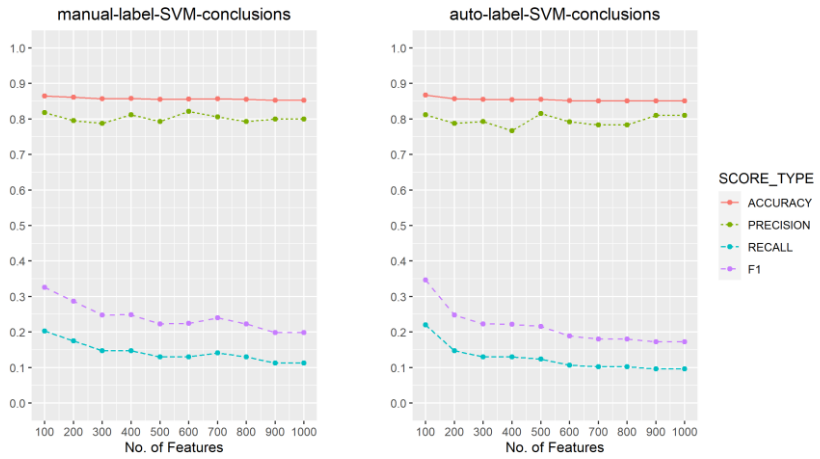

Figure 16. Experiment 3 results for the "conclusions" class

Table 3 lists numbers and percentages of instances whose manual labels and automatically generated labels do not match. "Objective" has the most instances scattered in other sections (33.7%), which may explain why switching to manual labeling brought so much benefit. However, "methods" class has a comparable percentage of mismatched labels to "results" and "conclusions" but received more benefit in terms of recall that the other two.

Table 3 Mismatched labels

Manual Label	Automatically generate label	No. of mismatched instances	Percentage
Background	Objective, methods, results, conclusions	162	17.8%
Objective	Background, methods, results, conclusions	166	33.7%
Methods	Background, objective, results, conclusions	135	10.4%
Results	Background, objective, methods, conclusions	193	10.0%
Conclusions	Background, objective, results	80	9.05%

4.5 Apply classifiers to citation context sentences

Despite the poor performance of classifiers and the potential difference in language used in abstracts and the full text, when applied to citation context sentences, the classifiers produced surprisingly good results (Table 4). The "methods" classifiers picked out three "methods" sentences, sentence 6, 7, and 8. And the "results" classifier correctly classified sentence 5. Interestingly, the "results" classifier also classified sentence 3, which I annotated as "methods." Close examination revealed that it is a hybrid type: It described a method, the use of a monoclonal antibody to confirm the expression of tau protein, and a result, the confirmation of the strong expression of tau protein. In other words, the classifier is correct.

Table 4. Classification results of citation context sentences

Sentence No.	Citance to be classifed	Machine classification results	My annotation
1	We took advantage of a mouse line in which expression of a tet transactivator transgene is under control of the neuropsin gene promoter.	No hit	Methods
2	This line was crossed with the Tg(tetO-tauP301L)4510 line that only expresses human tau carrying the P301L frontotemporal dementia mutation in the presence of a tet transactivator.	No hit	Methods
3	Immunohistochemistry using the 5A6 antibody (courtesy of Dr.G.V. Johnson, Universityof Rochester), a monoclonal antibody raised against the longest form of recombinant human tau which recognizes an epitope between amino acids 19 and 46, confirmed strong expression of tau protein in superficial layers of the MEC and parasubiculum in rTgTauEC mice at 3 months of age compared to a control brain (Figure 1D).	Results	Methods
4	The major output of the entorhinal cortex is a large axonal projection called the perforant pathway that carries input from layers II and III of EC to the hippocampus, terminating in the middle molecular layer of the DG.	No hit	Background
5	Alz50-positive aggregates were also found in large numbers of neurons without detectable transgene expression in the DG, anterior cingulate cortex, CA1, and CA3, all major targets of the EC. Importantly, unlike the anterior cingulate cortex, cortical areas that showed limited transgene expression outside of the EC, but do not receive direct input from the EC, did not show any tau aggregation. Moreover, the cerebellum, which expresses human P301L tau mRNA, did not develop any fibrillar accumulation of htau in the soma.	Results	Results
6	Therefore, we assessed two synaptic markers in the perforant pathway terminal zone of rTgTauEC mice: synapsin-I, a marker of synaptic vesicles, and PSD-95, a postsynaptic marker that has been reported to decrease early in neurodegeneration.	Methods	Methods
7	Therefore, we turned to a protocol that relies on density functional theory-based computations of 1H and 13C NMR chemical shifts and the use of statistical tools to assign the experimental data to the correct isomer of a compound.	Methods	Methods
8	Conformational analysis of 4 was performed with Schrödinger MacroModel 2016 by following the method of Willoughby et al.	Methods	Methods
9	Scaling factors (slope = -1.0522, intercept ¼ 181.2412) are applied to the 13C NMR shielding tensors (B3LYP/6-311 þ G (2d,p)//M06-2X/6-31 þ G (d,p) to calculate the 13C NMR chemical shifts.	Results	Methods
10	In AD, early hallmarks include the loss of synapses, and comparison of AD patients to age-matched control individuals showed that the density of synapses correlated strongly with cognitive impairment, suggesting that loss of connections is associated with the progression of the disease.	No hit	Background

To understand the results, I must explain my annotation procedure, which is quite different from the procedure that created the dataset. I first create an argument diagram for each main claim stated in the article using an ontology called micropublication model. Then, I align citation context sentences to components in the diagram. Although I will annotate a citation context sentence as "methods", it may not be a "method" sentence. It only contains the information about the "methods." One of the valuable take-away from the project is the tiny difference between the two definitions. Moreover, I labeled sentence 4 and 10 as "background," in the sense that they provided “background” information for reasoning. They are not technically “background” sentences, and they were not picked up by any of the classifiers.

5. Discussions

5.1 Comparing my results to published work

My models are quite primitive compared to published work [3], [5], [6]. However, after conducting my experiment, I am in a better position to understand their feature engineering methods than before. First, my precision and accuracy are comparable to what they have achieved. The main difference lies in recalls, meaning there is room to improve by adding new features. Second, all studies have used “condensed” features to deal with data sparsity, such as replacing numbers with the string “NUM” to mask the diversity or using “T/F” to replace a chunk of the term vector. The “linguistic features” from [3] also belong to this category, although I still cannot grasp how the conversion was done. My understanding is that such operations eased the burdens for classifiers. In the future, this is a technique that I need to explore further.

5.2 Building classifiers to find keystone citations: insight from this experiment

The idea of using rhetorical category classifiers to screen keystone citations is based on the two assumptions. First, methods (including materials) are of predominant importance in experiment-based research, and therefore, citation context sentence classified as “methods” are highly likely to be keystone citations. Second, citations appearing in discussions sometimes supply crucial knowledge to support author’s arguments, and therefore, citation context sentences classified as “conclusions” (the category that “discussions” are normalized to) are likely to be keystone citations too. This project helped me validate and adjust my original ideas.

First, extracting “methods” citation context sentences is possible. Such classification by supervised learning has been done before but with a slightly different approach [10]. The cited papers were first tagged as “methods” or “non-methods” paper. Their citation context sentences are subsequently tagged as “methods” or “non-methods” and used as the training set. Our approach is more precise because the sentences are directly labeled as “methods” or not. However, one revelation came from this experiment that sentence can be of hybrid type (e.g., sentence 3) and end up not being classified as “methods.”

Secondly, my experiment challenged the second assumption. The reasons are two-fold. First, with the absence of a distinct “discussion” category, it seems that this type of citation context sentences was classified to “results” rather than “conclusions” (e.g., sentence 5). Secondly, when taking out of context, some of them read like “background” (e.g., sentence 10). Therefore, the second assumption (and subsequent solution) need to be adjusted. What I want to identify is whether the text surrounding the citations has an argumentative nature. I probably need to extend the span to several sentences rather than one. In terms of detecting “argumentative” nature, there are two possible venues. First, I can classify a block of sentences as argumentative or non-argumentative, which has been done before and is also a core task for argumentation mining [11]. Otherwise, I can also detect hints of argument scheme in a block of text, which also has been attempted before, but it is a much harder task [12] as it goes deeper into detecting the structure of arguments. The associate benefit is that we can finally leave the regime of using proxies and will have the chance to locate role citation plays in scientific argument.

6. Conclusions

I have built a set of models to classify sentences from abstracts to one of the five rhetorical categories: Background, objective, methods, results, and conclusions. I compared three classification algorithms and three representations. Among three classification algorithms, support vector machine so far delivers the best performance, with good precision (0.8 to 0.9) and moderate to poor recall (0.2 to 0.5). Representation wise, stemming and removing punctuations do not improve the performance, and trigram-alone does not perform better than s "bag-of-words" representation, except the class of "objective." Switching to manual labels brings the most benefit to the "background" and "objective" categories, improving precision and recall. It also brings benefit to the "methods" category by improving the recall. Benefit to "results" and "conclusions" is marginal. I also applied classifiers to classify sentences surrounding "keystone" citations ("keystone citation context"). Surprisingly, those classifiers produced decent results. They correctly classified 4 out of 10 sentences. Even the misclassified cases generate new insights into the problem.

7. Future Work

In the future, I will combine trigram and bag-of-words representations. Syntax-based representations can be explored as a strategy to deal with "terminology overloading." I also need to rethink my project. For example, what differentiate those rhetorical categories, if boundaries between them are so blurry? When a citation context sentence is "about" methods, does it mean that it has to be a "methods" sentence?

References

[1] Y. Fu and J. Schneider, "Towards knowledge maintenance in scientific digital libraries with the keystone framework," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, New York, NY, USA, 2020, pp. 217–226, doi: 10.1145/3383583.3398514.

[2] A. Jimeno Yepes, J. Mork, and A. Aronson, "Using the Argumentative Structure of Scientific Literature to Improve Information Access," in *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, Sofia, Bulgaria, Aug. 2013, pp. 102–110, Accessed: Aug. 20, 2020. [Online]. Available: https://www.aclweb.org/anthology/W13-1913.

[3] S. Nam, S. Jeong, S.-K. Kim, H.-G. Kim, V. Ngo, and N. Zong, "Structuralizing biomedical abstracts with discriminative linguistic features," *Computers in Biology and Medicine*, vol. 79, pp. 276–285, Dec. 2016, doi: 10.1016/j.compbiomed.2016.10.026.

[4] P. Ruch *et al.*, "Using argumentation to extract key sentences from biomedical abstracts," *International Journal of Medical Informatics*, vol. 76, no. 2, pp. 195–200, Feb. 2007, doi: 10.1016/j.ijmedinf.2006.05.002.

[5] Y. Yamamoto and T. Takagi, "A Sentence Classification System for Multi Biomedical Literature Summarization," in *21st International Conference on Data Engineering Workshops (ICDEW'05)*, Apr. 2005, pp. 1163–1163, doi: 10.1109/ICDE.2005.170.

[6] Y. Guo, A. Korhonen, M. Liakata, I. S. Karolinska, L. Sun, and U. Stenius, "Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes," in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, USA, 2010, pp. 99–107.

[7] A. de Waard and H. Pander Maat, “Verb form indicates discourse segment type in biological research papers: Experimental evidence,” *Journal of English for Academic Purposes*, vol. 11, no. 4, pp. 357–366, Dec. 2012, doi: 10.1016/j.jeap.2012.06.002.

[8] H. Shahriari, “Comparing lexical bundles across the introduction, method and results sections of the research article,” *Corpora*, vol. 12, no. 1, pp. 1–22, Apr. 2017, doi: 10.3366/cor.2017.0107.

[9] A. Agibetov, K. Blagec, H. Xu, and M. Samwald, “Fast and scalable neural embedding models for biomedical sentence classification,” *BMC Bioinformatics*, vol. 19, no. 1, p. 541, Dec. 2018, doi: 10.1186/s12859-018-2496-4.

[10]H. Small, “Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty,” *Journal of Informetrics*, vol. 12, no. 2, pp. 461–480, May 2018, doi: 10.1016/j.joi.2018.03.007.

[11]M. Lippi and P. Torroni, “Argumentation Mining: State of the Art and Emerging Trends,” *ACM Trans. Internet Technol.*, vol. 16, no. 2, p. 10:1–10:25, Mar. 2016, doi: 10.1145/2850417.

[12]V. W. Feng and G. Hirst, “Classifying Arguments by Scheme,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, USA, 2011, pp. 987–996.