Are Self-Tuning and Ensemble Methods Necessary for Sentiment Analysis of Scientific Citation?

Huy Tu
North Carolina State University
Raleigh, North Carolina
hqtu@ncsu.edu

ABSTRACT

Sentiment analysis of citations within scientific papers have been studied to more appropriately indicating the quality of published papers in comparison to the outdated quantitative evaluation approach basing on the citations' frequency. Approximating the popularity, the context, and the impact of the published research are commonly referred to as bibliometric measurements. In this study, we shall review state of the art sentiment analysis algorithms that helps incorporated that qualitative aspect to the bibliometrics. Moreover, we will focus on the improvement of result analysis of those state of the art algorithms by proposing empirical software science approaches such as ensemble methods and parameter tuning to.

***Keywords* -** Citation Sentiment Analysis, Classification; Differential Evolution, Parameters Tuning; Clustering, Cluster Ensemble.

1 INTRODUCTION

Bibliographic citations have kept the research community alive. Iorio et al [12] have expressed the importance of bibliographic citations as they are tools for:

- *disseminating research* - with references as directional edge to point to motivational background, related work, data sources, and methods coming from various publishing platforms.
- *exploring research* - the network of citations (papers or authors or journals are presented by the nodes and the edges representing the referencing act) can be utilized as readable data for searching, visualizing, filtering, and aggregating for interesting problems such as knowledge transfers across domains, scientific collaboration, and the scholarly disciplines map.
- *evaluating research* - characterizing and quantifying the importance and impact of the work/research through the nature of the purposes those citations and not only the mere existence of those papers.

In this paper, we will focus only at the third important aspect of bibliographic citations, evaluating research. Evaluating research's impact should be optimized for fairness and objectiveness through both quantitative and qualitative evaluations. Only considering one without the other aspect of the evaluation would be misinformed, narrow-minded, and problematic.

- For quantitative aspect without qualitative aspect, the nature of citation itself would not be considered and all the citations would be taken in as they carry equivalent weights when they should not. For example, the referencing act of crediting should be counted differently for the referencing act of criticizing.
- For qualitative aspect without quantitative aspect, the interest of other researcher in it and the popularity of the work [1] would be missing. Moreover, there would be no statistics for significance and effect insights from the qualitative evaluation.

The old-fashioned quantitative citation bibliometrics include Hirsh-index, the g-index [8], and PageRank [15]. The moden qualitative NLP-based bibliometrics include supervised learning of polarity and purpose of citation classification.

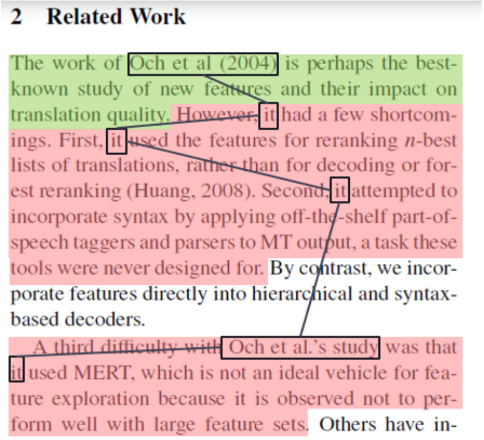
Citation purpose classification is the studying of the motive behind citing that work or research. In 1977, Spiefel-Rosing et al. [16] formerly suggested 13 categories for citation purpose. In 2006, Teufel et al. [17] adopted 12 categories from Spiegel-Rosing's work that can be appropriately grouped to four types: weakness, contrast (4 categories), positive (6 categories), and neutral. In 2013, from those previous work, Abu-Jbara et al. [1] condensed the previous's quantity of categories down to only six categories of: criticizing, comparison, use, substantiating, basis, and neutral(other).

Citation polarity classification through sentiment analysis aims to determine opinions, emotions, and attitudes of the specific granularity of the text region. It can be conducted at three levels of document-level, sentence-level, and context-level. We will focus more on citation polarity classification on context level for this work. The document-level and sentence-level of sentiment analysis have been incomplete and defective due to how citation region can be varied, how multiple citations can be included in one sentence while having different sentiment tags, and how the sentiment can be hidden in the context level. For example:

In the figure 1 above, the citation of Och et al from "Smorgasbord of Features for Statistical Machine Translation" paper firstly included 'best known study' [11], making this citation positive. However, if we read the following sentences after this citation, we find that this citation is being referred in the next 5 sentences anaphorically like the word 'it' and other such phrases which criticizing the original referenced work negatively which makes this citation sentiment difficult to judge.

Specifically, there are still many difficulties and open research opportunities within the existing models for citation sentiment analysis [2]:

- Sentiment is often subtle or even hidden in citation within research community. Negative polarity is often embedded contrastively.
- Sentiment of the citation within sentences are often neutral.
- Diverse variation of sentiment lexicon or sentiment carrying scientific and/or technical terms/phrases (and specifically from the authorâĂŹs research area). Some are longer in



2 Related Work

The work of Och et al (2004) is perhaps the best-known study of new features and their impact on translation quality. However, it had a few shortcomings. First, it used the features for reranking *n*-best lists of translations, rather than for decoding or forest reranking (Huang, 2008). Second, it attempted to incorporate syntax by applying off-the-shelf part-of-speech taggers and parsers to MT output, a task these tools were never designed for. By contrast, we incorporate features directly into hierarchical and syntax-based decoders.

A third difficulty with Och et al.’s study was that it used MERT, which is not an ideal vehicle for feature exploration because it is observed not to perform well with large feature sets. Others have in-

Figure 1: Difficulties for Sentiment Analysis for Contextual Citations [3]

length (such as âĂIJstate of the artâĂİ) which suggests that considering higher n-grams would be useful.
- From a single clause level to multiple paragraphs level, the region of influence of citations differ.

Sentiment polarity varied within context which inherits the cognitive and social values from the research communityâĂŹs culture. It is definitely an interesting problem that has been studied as a subject of scholarly analysis in the research community. From the previous work, they can be grouped into two main type of sentiment analysis research. The old-fashioned one focused on rule-based schema based on a reconstructed decision tree classification having pre-defined cue words and phrases set to classify extracted citation scope (Garzone, 1997; Nanba et al., 2000, Pham et al., 2003) [7]. The more advanced one incorporated state of the art machine learning models such expert knowledge of lexicon (scientific and technical terms) or phrases (cues) (Angrosh et al., 2010; Teufel et al., 2006; Athar, 2011 & 2012) [7]. Those machine learning based classifiers included IBk k-NN (k- Nearest Neighbors), support vector machine (SVM), and CRF. The summary of the previous efforts is recorded in the Table 1 based on their features and classifier.

Those second type of sentiment analysis of scientific citation research applied popular state of the art machine learning models stand alone with their based metrics/parameters. The author(s) prepared and cleaned the data as the citation text for tagging, parsing, and transforming before applying the classifier. However, fine tuning of parameters have been studied to improve the result analysis by optimizing the textual data which in this case are citation context using differential evolutionary algorithm. Moreover, Colleta et al. [6] has employ an ensemble method including SVM classifier with cluster ensembles combination which resulted in better classification accuracy. According to Menzies [10], tuning is under-explored for optimization problems. Consequently, the execution of this essay/plan for the project will strive for employing those ideas into the sentiment analysis of scientific citation problem and possibly combining both the differential evolutionary methods with ensemble methods.

Table 1: Comparison of Citation Purpose & Polarity Analysis Schemas [7]

Work	Features	Classifier
Teufel et al. 2006	Cue phrases Verb tense/voice Modality Location (paper/paragraph)	IBk (k-NN)
Angrosh et al. 2010	Generalization terms (Lexicon) (Prev.) Sentence has citations	CRF
Dong&Schafer 2012	Cue words Boolean and weight POS-tag	SMO BayesNet NaiveBayes
Athar 2012	four-class anotation 1-3 grams Scientific lexicon POS-tag Contextual Polarity Dependency Structure Sentence Splitting (removing) Negation	SVM
Aju-Jbara et al. 2013	Reference Count Reference Separation Cue words Self-Citation Contains 1st/3rd Person Pronouns Contrary Expression Dependency Relations Negation	SVM

The rest of the paper is organized as follow: [2] discusses the diverse criteria of their usefulness and feasibility of incorporating that criteria into state of the art models for sentiment analysis of scientific citation, [3] expands in details on the self-tuning criteria that this project will focus on, [4] reviews the state of the art tools and research and how those state of the art work fail the self-tuning criteria , [5] proposes the simple engineering approach to enable self-tuning criteria, and [6] maps out the plan for building the self-tuned model and comparing it with the state of the art models.

2 EVALUATION CRITERIA

2.1 Model Readability

Human readable and interpretable models are essential for transferable knowledge that are often overlooked. In development process, the collaboration, experiment, and management of project can be optimized within a restricted time frame. For business usage, managers and non-technical audiences can have a better understanding while brainstorming for potentially new ideas. However, achievability of model readability is hard since the performance of the state of the art models will be compromised due to lower complexity (No Free Lunch Theory).

2.2 Actionable Conclusions

Similar to the first criteria, actionable conclusion thrives from the interpretable aspect of the results to take the next appropriate step to move forward with the model. For actionable conclusions, it would be more essential to understand the differences and impacts of the features/variables in the model that lead to the results instead of the architecture of the model and which algorithms that built the model.

2.3 Learnability and Repeatability of the Results

The availability and weight of data have grown exponentially over time in this 21st century which is difficult when you need to develop a model that can be applied to most data appropriately. Therefore, the learnability and repeatability of the results is important to improve and reproduce results to confirm performance and adjusting the developed model is an continuous act. The small memory footprint (RAM, disk, CPU, and GPU) models/algorithms are appealing such as Mini Batch K-Means and Naive Bayes.

2.4 Multi-Goal Reasoning

Most problems out there in real world situations have multiple goals to solve. In computer science, goal models have represented software requirements, business objectives, and design qualities. However, it is a complex problem, there have been limitations of expressiveness and/or tractability in coping with the real world situation from existing techniques. One of the methods of obtaining solutions for multiple goal reasoning is by converting multiple goal measures into single goal (e.g. domination score) and then use the existing single goal algorithms to obtain solutions.

2.5 Anomaly Detection

Anomaly is the entry item, event, or observation which do not approximately fit in with the expected pattern or the homogeneous school of existing items in the dataset. Anomaly detection is an act of identifying those anomalies or outliers. The integration of anomaly detection with the existing models would have low complexity and computational power. However, there are difficulties with implementing of anomaly detection that is aware of the context and one-class anomaly classification would be more difficult to achieve.

2.6 Incremental Learning

Increment is an act of addition or increase for one on a scale according to a problem. Incremental learning algorithm describe the machine learning algorithm where the learning process takes place whenever new entries emerge and adjusts the knowledge bank according to the new entries without started the learning from scratch. This machine learning paradigm is essential when the new example(s)/data need to be integrated to the existing model or they are continuously emerging (streaming data like weather, social media, and networking). Decision trees/rules, artificial neural networks, incremental SVM are a few among many state of the art machine learning algorithms that support incremental learning.

2.7 Sharable

With the exponential growth in the machine learning fields along with the hype of deep learning area, accessibility to model and especially the data to train the model is imperative than ever. Data privacy becomes an issue when the demands for data drastically increase. Moreover, the feasibility of shared learning models to run smoothly across hardware and software platforms is notable. Two approaches to this criteria can be summarized below:

- the training data (somehow) can be succinctly described
- and/or that dataset can be mutated to lower the odds of detecting single within the data

Noting that, that summarized/mutated dataset still applicable for a model to learn from.

2.8 Context Aware

Context-aware is another aspect that most state of the art work have also overlook. The data is analyzed as an homogenous body of information without disaggregating them to analyze in different situations or groups (e.g. context). Deriving analysis and decision making based on that "whole" data will result in missing out on important trends appearing in groups of data but disappearing or reversed then the data is aggregated. The analysis act will suffer from the SimpsonâĂŹs paradox or the Yule-Simpson effect.

2.9 Self-Tuning

Self-tuning or auto-tuning refers to the aspect of able to optimize itself (by adjusting it's own parameters) which is whatever models and software we are using. Self-tuning aspect helps to satisfy the objective function by maximizing or minimizing the appropriate requirements. Examples of self-tuning include increase of analysis results, maximization of efficiency, or error minimization. The self-tuning method is based on finding the optimal set of gains from a pre-generated training set for a further selection of the best seeds through a membership function.

3 KEY CRITERIA

No machine learning model work for all dataset and situations, but most machine learning models are predefined with sub-optimal parameters and the built model depends on those parameters and the data to learn. The act of optimization for parameters within a learner can impact the performance in term of the efficiency and result analysis of the learner. Some of the parameters of state of the art machine learning algorithms are below:

- Parameter C to set the amount of regularization of SVM (Support Vector Machine) module in *Scikit-learn*
- Parameter *number-of-trees* to set the quantity of decision trees in a random forest
- Parameter K to set the number of nearest neighbor in kNN (K Nearest Neighbors)

These parameters have similar function within their machine learning models. A small value for parameter C, number of decision trees, and K will generate a simple model with potentially more training errors and a larger value setting for those parameters would result in a more refined model with less training error. However, the improvement of result analysis decreases as the quantitative settings for those parameters increases, i.e. at a certain point the benefit in prediction performance from operations (applying more regularization, learning more trees, and checking more members) will be lower than the cost in computation time for these operations. Moreover, there are abundant considerations or settings for the control parameters (1000+ combinations for several models) that leads to intensive computing power. Efficiency and caution are important aspects in the search for the right settings of parameters to obtain the best performance for the model.

4 CRITIQUE

Most state of the art machine learning algorithms for sentiment analysis are stand-alone and stand-alone tools in scientific citation analysis. There have been a lot of efforts putting into detecting citation region and defining/analyzing useful structural features (in word-level, sentence-level, and context-level). Those machine learning models include IBk k-NN, SVM, CRF, and NB. Those work are valuable as discussed above, yet there are several concerns with that approach:

- Off-the-shelf algorithms have parameters that are not tuned to maximize the performance.
- The instability of the prediction method that results in very different outputs with slightly different inputs.

As a result, they potentially miss out the reliability and optimization for the result of the model. Therefore, all of the approaches and efforts on scientific citation sentiment analysis have failed to attempt to self-tune their models.

5 REVIEW

In the field of software engineering, features optimization have not been common till recent [9]. Regarding those previous work on sentiment analysis for scientific citations, with no parameters tuning methods applied, any parameters tuning method would be sufficient to start with. The traditional and popular Grid Search do search for the entire space but they are very slow and may not be the most effective for rerun the algorithm on the dataset repeatedly. Random search algorithms (e.g. differential evolution algorithm) are arguably can outperform Grid search algorithms with efficiency and performance according to Bergstra et al. [4]. Differential Evolution(DE) algorithms are simple to code and have been shown to tune parameters effectively. DE iteratively tries to improve a candidate solution with regard to a given measure of quality or parameter. Moreover, according to Menzies work [10], Grid Search can take to thousands of iterations to converge while the Evolutionary algorithms converge drastically faster in most cases, around 100 iterations.

Ensemble Methods uses multiple learning algorithms instead of any stand-alone learning algorithms in order to obtain better (and stable) result analysis. For example, the random forest algorithm has multiple CART models which perform better than an individual CART model by counting the votes of class by each tree and picking the one with the most votes. Two common types of ensemble methods include:

- Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The instability of the prediction method for state of the art algorithms can be utilized by running multiple instances, it can be shown that the reduced instability leads to lower error.
- Boosting predictors is a method where the first predictor is built on the whole dataset while the following is built on the training set based on the performance of the previous one. The pre-defined equal weights will be justified when the instance being classified incorrectly which leads to being more cautious with the misclassification that leads to generally improved accuracy from bagging.

6 PLANNING

The collective goal of the project is to develop the self-tuning criteria of the 'tuned' machine learning algorithms and then compare it with the state of the art algorithms for sentiment analysis of scientific citations proposed in these previous work.

Parameter Tuning method such as Differential Evolution shall be used to maintain and optimize the right parameters while the Ensemble methods would help optimize and stabilize the analysis results after running the aggregated and/or stacked algorithms on the pre-processed data.

For the dataset that we will use for this project, the annotated 8736 citations from 310 research papers taken from the ACL Anthology from Bird et al. [5] will be applied. Moreover, the project ideally can take a sample from the dataset includes 35,391 Software Engineering papers from the last 25 years published in 34 top-ranked conferences and journals that have been used in [13]. There would be three classes for sentiment classification of scientific citation: negative, positive, and objective (or neutral) as Athar proposed in [2].

With the collective goal in mind, the project shall ask these research questions:

RQ 1: Can we reproduce Athar's baseline results with the ACL Anthology dataset? Using such a baseline, we can compare our methods to those of Athar.

RQ 2: Is the results from tuning the baseline algorithms with parameters tuning and ensemble methods would outperform the results achieved by the standalone baseline algorithms from previous work?

RQ 3: Is the cost of tuning (space and time complexities) acceptable?

RQ 4: What would be the results if we applied those models on the SE papers dataset?

In order to explore those research questions, we shall benchmark the three 'tuned' models with the baseline state of the art methods (IBk k-NN, SVM, CRF, and NB) from the previous work on citation sentiment analysis:

- The baseline method using parameters tuning algorithms
- The baseline method using ensemble methods
- The baseline method using both parameters tuning algorithms and ensemble methods

Table 2 below summarizes the requirements for the project. First column shows the list of datasets that will be used to build the model. Second and third column indicate the ensemble methods and tuning algorithms that will be used to optimize the result analysis from the machine learners in the last column.

Table 2: Dataset, Ensemble methods, Self-tuning algorithms, and Classifiers in this project

Dataset	Ensemble Methods	Tuning Algorithms	Classifier
ACL Anthology	Bagging	Grid Search	SVM
SE Papers	Boosting	DE	NB
		Alternate DE	k-NN
			CRF

The performance assessment can be applied with standard measures of Accuracy, Precision, F-score, Recall, and Area Under the Curve (AUC). To further evaluate results obtained in empirical analysis, two-way analysis of variance (ANOVA) can be applied to observe the statistically meaningful differences the classification accuracies of the algorithms. There is no best way to compare or rank the results of the baseline model with the tuned models, depending on the goals, the data, and the audience [14]. However, in order to explore more of the comparison act, stats.py can be utilized [14].

REFERENCES

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R. Radev. 2013. Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In *HLT-NAACL*.

[2] Awais Athar. 2011. Sentiment Analysis of Citations Using Sentence Structure-based Features. In *Proceedings of the ACL 2011 Student Session (HLT-SS '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 81–87. http://dl.acm.org/citation.cfm?id=2000976.2000991

[3] Awais Athar and Simone Teufel. 2012. Context-enhanced Citation Sentiment Detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 597–601. http://dl.acm.org/citation.cfm?id=2382029.2382125

[4] James Bergstra and Yoshua Bengio. 2012. Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.* 13 (Feb. 2012), 281–305. http://dl.acm.org/citation.cfm?id=2188385.2188395

[5] S. Bird, R. Dale, Bonnie J Dorr, B. Gibson, M.T. Joseph, M.Y. Kan, D. Lee, B. Powley, D.R. Radev, and Y.F. Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC'08)* (2008/// 2008), 1755 – 1759.

[6] Luiz Coletta, Nadia Felix, Eduardo Hruschka, and Estevam Hruschka. 2014. Combining Classification and Clustering for Tweet Sentiment Analysis. (01 2014).

[7] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology* 65, 9 (2014), 1820–1833. https://doi.org/10.1002/asi.23256

[8] Leo Egghe. 2007. Dynamic h-index: The Hirsch index in function of time. *Journal of the American Society for Information Science and Technology* 58, 3 (2007), 452–454. https://doi.org/10.1002/asi.20473

[9] Wei Fu and Tim Menzies. 2017. Easy over Hard: A Case Study on Deep Learning. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017)*. ACM, New York, NY, USA, 49–60. https://doi.org/10.1145/3106237.3106256

[10] Wei Fu, Tim Menzies, and Xipeng Shen. 2016. Tuning for Software Analytics. *Inf. Softw. Technol.* 76, C (Aug. 2016), 135–146. https://doi.org/10.1016/j.infsof.2016.04.017

[11] Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Simon Fraser U, Kenji Yamada, Alex Fraser, Libin Shen, David Smith, Johns Hopkins U, Katherine Eng, Stanford U, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A Smorgasbord of Features for Statistical Machine Translation. (05 2004).

[12] Angelo Di Iorio, Andrea G Nuzzolese, and Silvio Peroni. 2013. Towards the Automatic Identification of the Nature of Citations. *SePublica* (2013), 63âĂŞ74. http://ceur-ws.org/Vol-994/paper-06.pdf

[13] George Mathew, Amritanshu Agarwal, and Tim Menzies. 2016. Trends in Topics at SE Conferences (1993-2013). *CoRR* abs/1608.08100 (2016). arXiv:1608.08100 http://arxiv.org/abs/1608.08100

[14] Timothy Menzies. 2017. Fss17: Evaluation. (2017). https://txt.github.io/fss17/stats

[15] Dragomir R. Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2016. A bibliometric and network analysis of the field of computational linguistics. *Journal of the Association for Information Science and Technology* 67, 3 (2016), 683–706. https://doi.org/10.1002/asi.23394

[16] Ina Spiegel-RÃűsing. 1977. Science Studies: Bibliometric and Content Analysis. *Social Studies of Science* 7, 1 (1977), 97–113. http://www.jstor.org/stable/284635

[17] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic Classification of Citation Function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 103–110. http://dl.acm.org/citation.cfm?id=1610075.1610091