Towards Determining Textual Characteristics of High and Low Impact Publications

Yue Chen, Kenneth Steimel, Everett Green, Nils Hjortnaes, Zuoyu Tian, Daniel Dakota, Sandra Kübler
Indiana University
Bloomington, IN, USA
{yc59,ksteimel,evegreen,nhjortn,zuoytian,ddakota,skuebler}@indiana.edu

Abstract

This paper is concerned with the question of whether we can predict the future impact of a paper based on the text of the paper. We create a corpus of papers in computational linguistics, and we create gold standard impact annotations by using their Google Scholar citation counts. We use supervised classification approaches to automatically predict impact of the papers. Our results when using very simple features show some success, but they also show that the classifiers suffer from class imbalance problems.

Keywords: impact detection, data imbalance, corpus

1. Introduction

This paper is concerned with the question whether we can predict the future impact of a paper based on the text of the paper. In other words, are there textual characteristics that increase the impact of a paper? We define the impact of a paper as its citation count. While this question sounds somewhat unrealistic, it does make sense when looked at from the angle that properly advertising one's work should have a positive effect on its reception. A well written paper cannot succeed if there is no academic content. But some papers that have the content, but package it suboptimally may not get as much attention as they deserve. In this vein, our question can be rephrased as: Which textual characteristics do we need to adapt in order to produce a successful paper?

In our work, we investigate papers from the major conferences and journals in computational linguistics. We create a corpus of such papers on the topics of parsing and machine translation, and we create a gold standard of their impact by using their Google Scholar citation counts. We then separate the papers into three classes: low impact, high impact, and highest impact. We use supervised classification approaches to automatically predict impact of the papers. Our results when using very simple features show some success, especially when we use the full papers rather than just the abstracts, but they also show that the classifiers suffer from problems; they have a tendency to group all papers into the low impact class, which is the majority class.

There are two possible reasons for the behavior of the classifiers: One possibility is that the features we use are not predictive enough. The second possibility concerns the problem of class imbalance since the highest impact setting has very few examples. Depending on which of the reasons holds, we need to address the problem by either feature engineering or data sampling. To test the two hypotheses, We removed stopwords from the content, both abstracts and whole texts. We also experimented with both down-sampling and up-sampling. Random down-sampling of the low and high citation classes yields more balanced performance across the classes but results in a reduced overall accuracy due to the small amount of data used. This discrepancy is even more pronounced when only abstracts are used. Synthetic minority up-sampling techniques produced results very similar to the previous experiments.

The paper is structured as follows: We present related work in section 2., followed by a description of the corpus in section 3.. Section 4. presents the experiments and results with section 4.5. presenting an analysis of the features. We conclude with areas of future research in section 5..

2. Related Work

Traditional methods to determine the impact of a publication have heavily focused on citation counts. However, there are many methodological issues to consider as well as many caveats in these results. Furthermore, such metrics are often only retroactively obtainable and cannot indicate future impact. This has led to more focused work examining whether different sorts of features can be utilized to gauge the future impact of a publication. We are aware that this limits the objectivity of our gold standard (see section 3.), but since we are interested in automatic approaches to predicting future impact based on text, we assume that a switch in determining the gold standard will not have impact the usability of our methodology.

2.1. Citation Count Impact

Rankings based on citation counts are often used to demonstrate the "importance" of a publication. This is often performed by simply counting the number of times a publication (or a group of citations) has been cited by a different set of publications. More complex measures aim to account for types of variation and instead focus on the average number of citations on a set of papers and compensate for the length of time publication has been in existence (i.e. weighting publications having existed for three years against those for fifty years). Such methodology does yield a plethora of information. Adams et al. (2005) use citation probability metrics on the the Institute for Scientific Information to discover certain trends including: Higher ranked universities' citation sharing, mutual cross-over between scientific fields

in citations, and that there is a lag of about three years for the diffusion of information.

However, although informative and easy to access in terms of information, relying strictly on citation counts and probability metrics is often misleading and prone to inherent bias based on the given criteria. Meho and Yang (2007) compile a corpus created by fusing different citation metric systems, such as WoS and Scopus, to demonstrate that a selected metric significantly impacts how a publication can be ranked based on citation counts as different metrics exclude different fields, languages, or publication types.

Another approach taken is correlating the number of citations with the impact factor of the journal of publication to examine the interaction of the two. Levitt and Thelwall (2011) noted that standard citation metrics are not necessarily the best indicator of impact for the subject of economics, as there is also a strong correlation with the journal of publication. This suggests that the forum of publication is also relevant to impact, not just the number of citations and substance of the article.

2.2. Predictive Impact

Traditional methods of impact assessment can only be performed after a reasonable amount of time has passed to allow for the dissemination of the publication into a research community. Much of this work focuses on the use of citation counts to determine impact; however, this is rather limited in terms of future predictability. Thus, approaches utilizing more content for impact prediction have been an area of more recent research.

Ibáñez et al. (2009) examined which types of classifiers and features can be used to predict future citation frequency. They found that certain classifiers, such as Naive Bayes, performed better but also that certain tokens can actually be indicative of a publications of future citation frequency. Dietz et al. (2007) use an LDA-based approach that attempts to detect topical influence of cited documents on the citing document by linking individual references and word distributions on citing papers.

Other recent work has looked at how citation impact can be predicted at a publication's release. This has become relevant due to the electronic publication of many articles upon release. Brody et al. (2006) found a correlation between downloads of arXiv articles in certain scientific fields and their citation and impact. They further argue that downloads can also show a usage impact that is not correlated to citations and that as more databases become available, such impact may only increase.

With the advent of social media, the announcement of the existence of new publications is disseminated through these mediums. This was explored by Eysenbach (2011) who noted that Twitter can help predict high impact publications by the frequency a publication is tweeted within the first few days of publication, suggesting that non-traditional metrics can be used to immediately identify impact.

3. Impact Corpus

We are interested in whether the content of a paper can give us information on whether this paper will have an impact on the field. Since we did not find any corpus that would allow us to investigate this question, we created our corpus. The corpus was sampled from leading publications in the field of Computational Linguistics, and more specifically from major conferences and journals that are incorporated into the ACL Anthology[1]. Specifically, we only took papers from the Computational Linguistics journal, ACL, NAACL, EACL and EMNLP due to their content and stylistic similarities. Since we need to access the text, using the PDFs from the anthology directly is of limited use. Thus, we used the texts available from the ACL Anthology Network[2] for the textual basis. This corpus was created by using OCR to convert the PDFs into text, with additional post-processing using both scripts and manual labor (Radev et al., 2009; Radev et al., 2013). We decided to concentrate on two major topics of computational linguistics, parsing and machine translation. To extract papers on those topics, all texts that use the words "parse" or "parsing" (case invariant) in their title were extracted for the parsing category, and all papers using the words "translate", or "translation" in the title were extracted for the machine translation category[3].

Year	Total Papers	Parsing	Machine Translation
2007	187	83	104
2008	279	108	171
2009	270	135	135
2010	306	130	176
2011	191	67	124
2012	225	81	144

Table 1: Distribution of papers across years

Since we define the impact of paper in terms of the number of citations a paper has received, we need to allow sufficient time between publication of the original paper and of the papers citing it. Thus, we chose a window of 5 to 10 years ago, i.e., we consider papers published between 2007 and 2012. Table 1 displays the distribution of papers across the years for which we collected data.

Citation counts were then collected for each of these papers using Google Scholar[4]. We extracted the citation counts manually, and we list the sum of all citations if a paper is listed more than once on Google Scholar.

Figure 1 shows the distribution of citation counts in the two topics. Based on this distribution, we established three categories: a low citation count (0-29 citations), a high citation count (30-119), and an extremely high citation count (>120). The graphs in Figure 1 show that this split results in a severely imbalanced data set, which will make the automatic prediction of impact very challenging. Citation counts follow a rough Zipfian distribution: 948 papers fall into the low-count bin, 424 papers fall into the high-count

[1] http://aclweb.org/anthology/

[2] http://tangra.cs.yale.edu/newaan/

[3] A third category corresponding to stance detection/sentiment analysis was also collected. However, the resulting collection of papers was too small to be of use.

[4] https://scholar.google.com/

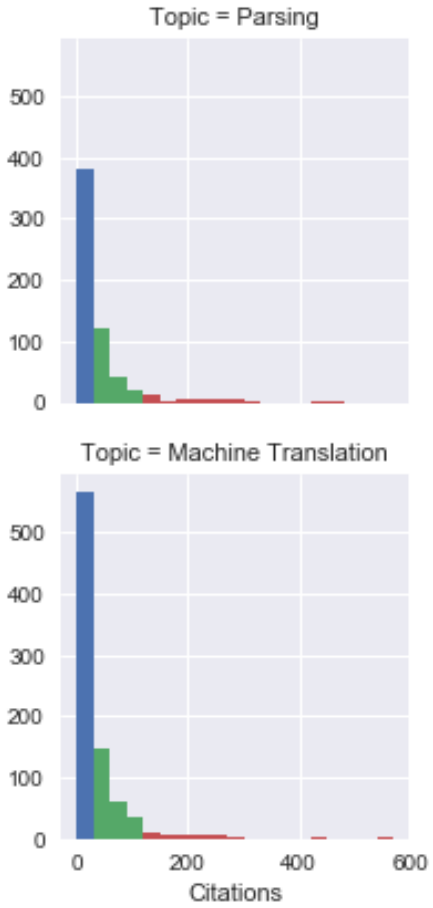

Figure 1: Citation class distribution with 3 classes

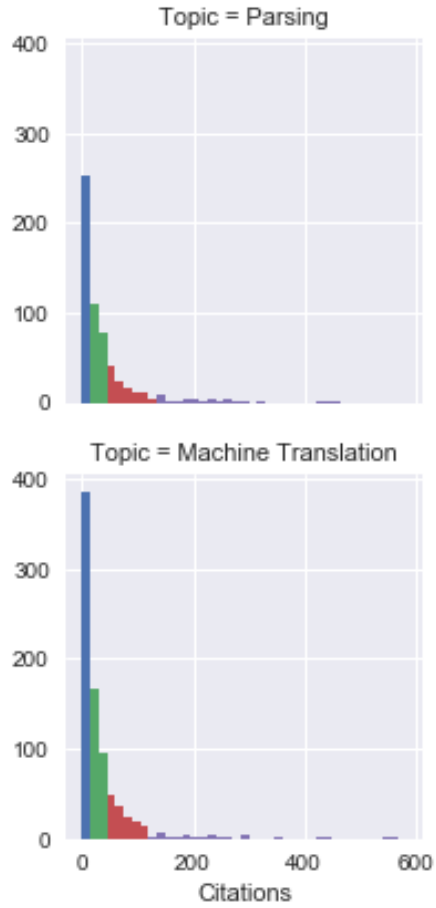

Figure 2: The 5 class split

Topic	Low	High	Highest
Parsing	381	181	42
MT	567	243	44

Table 2: Class distribution by topic

bin and 86 papers fall into the final highest-count bin. Table 2 shows the distribution across the two topics.

We are looking into an alternative classification using five citation classes as a way to mitigate the imbalance in the data. The classes consist of a no impact class for papers receiving 0 citations, a low class for papers with 1-15 citations, a moderate class for papers with 16-45, a high class for papers with 46-125 and a very high class for papers with more than 125 citations. The number of papers for this 5-class system are shown in figure 2. This graph shows that we obtain a less skewed data set.

4. Experiments

Our interest is whether we can predict a paper's future impact based on characteristics in the paper. We conducted a series of experiments to investigate how well the impact class can be predicted based on characteristics of the text in papers. For these experiments, we use a simple bag of words approach. All of the experiments presented here are based on the skewed 3-class data split.

We experiment with two types of texts: paper abstracts and full texts. This will ultimately answer the questions whether we can determine the impact of a paper based solely on the abstract and whether the abstract is as informative as the full paper. We also experiment with an additional pre-processing step: removing the stopwords. The list of stopwords is obtained from NLTK (Bird et al., 2009).

4.1. Extracting Abstracts

We extract abstracts automatically from the corpus using regular expressions. The regular expression will take the texts between the word "Abstract" and the word "Introduction". As some of the papers do not follow this pattern, their abstracts were not extracted. For these 70 papers, the abstracts could not be identified successfully, therefore we extracted those abstracts manually.

4.2. Experimental Setup

To create the training, development and test datasets used, we split the corpus per topic, i.e., we created separate training, development, and test sets for the parsing and the MT

	Parsing			Machine Translation		
Classifier	# features	Accuracy	F-score	# features	Accuracy	F-score
Random Forest	All	**60.61**	**45.74**	All	67.82	56.87
	10 000	**60.61**	**46.18**	3 000	68.97	59.15
Gradient Boost Trees	All	**60.60**	**46.18**	All	65.52	56.87
	5 000	60.60	48.91	10 000	67.82	58.89
Adaptive Boosting	All	**60.60**	**45.74**	All	66.67	58.89
	4 000	**60.60**	**45.74**	2 000	67.82	61.79
SVM	All	62.12	53.67	All	68.97	65.66
	10 000	62.12	57.22	10 000	68.97	65.66

Table 3: Results for both topics using only the papers' abstracts (boldface: majority classification)

domain. Out of every 10 papers, we randomly selected 1 paper for the development dataset, 1 paper for the test data, and the remaining 8 papers for the training set.

For the features, we extracted word unigram, bigram, and trigram counts from the texts of the training set. In the experiments shown in section 4.3., only the abstract is used for feature extraction while the experiments in section 4.4. use the entire text including the abstract. Then, we performed feature selection via a filter method, using both χ^2-goodness of fit and Mutual Information. The features with the highest scores below a specified count threshold are kept while all others are removed. We only report results using Mutual Information. χ^2 tends to result in similar, occasionally somewhat lower performance.

To gauge how sensitive performance is to specific machine learning approaches, we experiment with a variety of classification algorithms: Random Forest, Support Vector Machines (SVM), Adaptive Boosting, and Gradient Boost using shallow decision trees. We use the implementation in `scikit-learn` (Pedregosa et al., 2011). Each of the classifiers is trained using an exhaustive search over hyperparameter values.

4.3. Classifying Abstracts

The results for both topics are shown in table 3. The table shows several interesting results: First, it is clear from looking at accuracy that word n-grams do not provide enough information to determine impact of papers reliably. Additionally, the F-scores are considerably lower than the accuracies. This difference gives us an indication one problem: Many of the results are based on majority classification, i.e., the machine learner exclusively chooses the class that constitutes the majority class in the training data. Such cases are indicated in bold in the table. This shows that most of the classifiers prefer majority classification. Feature selection, which has been shown to have the potential of being useful in problems with class imbalance (Kübler et al., 2017), does not have any effect on accuracy in parsing. For the machine translation topic, it has a positive effect on all ensemble methods but does not improve the accuracy of SVMs. We will return to the question of majority classification below and have a closer look at performance per class.

As we described above, we repeated the experiments after having removed the stopwords. For abstract only features, this pre-processing step did not help with either accuracy or F-score.

A second trend that is obvious from table 3 is that predicting impact for the machine translation topic is more successful than for parsing: The highest accuracy reaches almost 70% while for parsing, the highest accuracy is around 62%. This cannot be explained by the imbalance in the data since machine translation has a higher skewing factor (the majority class is 1.98 times more likely than the other two classes combined) than parsing (1.71 times). Especially for SVMs, the F-scores are close to the accuracies, which means that the classifier goes beyond majority classification.

Returning to the issue of majority classification, table 4 shows the results in terms of precision and recall for selected experiments. These results show how serious the issue is: for parsing, SVM and Gradient Boosted Trees are the only classifiers that can identify at least some papers in the High class. For machine translation, all classifiers successfully identify some of the High class. However, none of the settings identifies any of the papers in the Highest class. At this point, it is unclear whether this is a consequence of the class imbalance in the data set or whether our feature set is not expressive enough to distinguish the classes. Further experiments using methods to address class imbalance are needed.

4.4. Classifying Full Papers

We now turn to the experiments where we use the full text instead of abstracts. The results of those experiments are shown in table 5. These results show that predicting impact based on the full text is more successful than predictions using only the abstract: For parsing, Adaptive Boosting reaches an accuracy of 77.27%, which is about 17% absolute higher than for abstracts. For machine translation, the same classifier reaches 72.41%, which is 5% absolute higher than its results on abstracts. Interestingly, both of these results are based on a small number of features chosen by feature selection. The corresponding F-scores show similar trends.

The results for the experiments in which we removed the stopwords are shown in table 6. We focus on the same settings as in table 5 to allow for a direct comparison between the two settings. These results show several interesting trends: For parsing, removing stopwords results in a massive deterioration across all classifiers. For machine translation, in contrast, Adaptive Boosting shows a minimal gain of 0.8% absolute in terms of F-score, and Random Forest gains close to 8% absolute. The reason for these gains

Topic	Classifier	# features	Class	Precision	Recall
Parsing	Random Forest	10 000	Low	61.54	100.00
			High	0.00	0.00
			Highest	0.00	0.00
	SVM	10 000	Low	67.31	87.50
			High	42.86	27.27
			Highest	0.00	0.00
	Gradient Boost Trees	5 000	Low	62.90	97.50
			High	25.00	4.55
			Highest	0.00	0.00
	Adaptive Boosting	4 000	Low	60.61	100
			High	0.00	0.00
			Highest	0.00	0.00
Machine translation	Random Forest	3 000	Low	67.86	100.00
			High	100.00	12.00
			Highest	0.00	0.00
	SVM	All	Low	71.83	89.47
			High	64.29	34.62
			Highest	0.00	0.00
	Gradient Boost Trees	10 000	Low	72.05	85.96
			High	62.50	38.46
			Highest	0.00	0.00
	Adaptive Boosting	2 000	Low	68.83	92.98
			High	60.00	23.08
			Highest	0.00	0.00

Table 4: Per class precision and recall for abstracts

	Parsing			Machine Translation		
Classifier	# features	Accuracy	F-score	# features	Accuracy	F-score
Adaptive Boosting	1 000	77.27	74.56	3 000	72.41	65.38
Support Vector Machines	50 000	68.18	60.35	50 000	71.26	64.43
Random Forest	2 000	71.21	65.56	1 000	71.26	64.50

Table 5: Results for both topics using the whole text (including stopwords)

	Parsing			Machine Translation		
Classifier	# features	Accuracy	F-score	# features	Accuracy	F-score
Adaptive Boosting	1 000	57.58	56.61	3 000	71.26	66.19
Support Vector Machines	50 000	54.55	56.36	50 000	59.77	58.73
Random Forest	2 000	66.67	60.00	1 000	74.71	72.44

Table 6: Results for both topics using the whole text (no stopwords)

require further investigation.

Table 7 shows the results in terms of precision and recall per class. These results corroborate our findings from table 5: The classifiers are all more successful in identifying High Impact papers than when they only have access to abstracts. This means that full papers contain more information about whether a paper has future impact on the field. When we allow stopwords in the features set, we do not find any of the Highest Impact papers. When we remove stopwords, however, SVM is able to predict the highest class with a precision of 12.50% and a recall of 25.00%. Even though these results are not stellar, we find this very encouraging in that feature engineering shows some impact on finding these highly cited papers.

4.5. Feature Analysis

Here we examine what types of features are selected by the feature selection model on abstracts. We focus on general trends within the features and potential correlations with known real world events during the selected time frame.
We first have a look at the experiments using abstracts only. For the top features for parsing, it is easier to identify common patterns and trends than for their machine translation counterparts. For example, the CoNLL 2007 shared task (Nivre et al., 2007) played an influential role in the direction of the field and is aligned with our time interval. This is noted in the returned features return for parsing as

				With Stopwords		Without Stopwords	
Topic	Classifier	# features	Class	Precision	Recall	Precision	Recall
Parsing	Adaptive Boosting	1 000	Low	79.17	95.00	70.00	70.00
			High	76.47	59.09	40.00	45.45
			Highest	0	0	0	0
	Support Vector Machines	50 000	Low	65.57	100.00	74.29	65.00
			High	100.00	22.73	39.13	40.90
			Highest	0	0	12.50	25.00
	Random Forest	2 000	Low	71.43	100.00	69.64	97.50
			High	77.78	31.82	33.33	9.09
			Highest	0	0	0	0
Machine Translation	Adaptive Boosting	3 000	Low	71.25	100.00	75.81	82.46
			High	85.71	23.08	52.00	50.00
			Highest	0	0	0	0
	Support Vector Machines	50 000	Low	70.00	100.00	70.49	75.44
			High	85.71	23.08	39.13	34.62
			Highest	0	0	0	0
	Random Forest	1 000	Low	70.89	98.25	78.13	87.72
			High	75.00	23.08	65.23	57.69
			Highest	0	0	0	0

Table 7: Precision and recall using the whole text

not only are references to the shared task returned, but many related terms: multilingual, dependency parsing, track. Not only was the shared task influential, but many of the then state-of-the-art systems participated in the task. This explains why so many of the top features can easily be associated with this knowledge. This leads to an interesting aspect: that by taking a small time interval, the currently most prominent topics will lead to the highest correlation to impact. One way to address this issue may be the use of topic modeling, for modeling this association between current topics of interest in the field and citation counts.This needs to be investigated further.

The features selected for machine translation, however, are not particularly informative. In the experiments using stopwords, many of the high-ranking features for MT are stopwords: the, we, to. While it is possible that certain grammatical constructions may be more clear, and thus papers that use these constructions may be cited more often, it does not seem likely. Comparing the features returned by both Mutual Information and Chi-square do not yield particularly interpretable features. In the experiments disregarding stopwords, many features can easily be associated with the field in general: system, evaluation, domain. Such features should provide any value in distinguishing between different levels of impact. This would help explain why there is little improvement gained without adding large quantities of features.

One exception is "Joshua" which refers to an MT system released in 2009 (Li et al., 2009) and is returned as a high ranking feature. This is interesting given that it was intended to be an alternative to the MT system "Moses" (Koehn et al., 2007) released in 2007 which is also a returned feature but with a much lower ranking. This further supports the notion of using topics as features for the classifier may give us access to current trends in the field as an indication of high impact features. However, one downside to using topics in this manner is that these topics may be too specific to a given time interval, and would not have the same usefulness in terms of determining impact for publication during a different era given that trends change.

5. Future Work

We have only scratched the surface of the problem of identifying the future impact of a paper based on textual features only. More experimentation and examination of the features is still required, particularly with regard to preprocessing decisions and the additions of various types of representations (e.g., lemmatization). We predict that these preprocessing decisions will have a strong impact on our results. Unlike many prediction tasks in which text is often shorter or limited (such as opinion mining of Twitter data), more text is available, thus there is a need to determine the best way of preprocessing such texts to eliminate as much noise as possible while also keeping specific types of non-standard features (e.g., keeping track of the number of figures or tables).

Additionally, while we see some success of classifiers in predicting high impact papers, we need to investigate whether other feature types are useful or whether we can improve results by using methods to address class imbalance in the data. Additional feature types will include character n-grams, which have been used successfully in stance detection tasks with imbalanced data (Mohammad et al., 2016), but also dependency triples and chains. Class imbalance can be addressed by upsampling methods that create artificial examples. In addition, the five way split described in section 3. may balance the classes better.

6. Bibliographical References

Adams, J. D., Clemmons, J. R., and Stephan, P. E. (2005). Standing on academic shoulders: Measuring scientific

influence in universities. *Annales d'Économie et de Statistique*, pages 61–90.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Brody, T., Harnad, S., and Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 233–240, Corvalis, OR.

Eysenbach, G. (2011). Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4):e123, December.

Ibáñez, A., Larrañaga, P., and Bielza, C. (2009). Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Kübler, S., Liu, C., and Sayyed, Z. A. (2017). To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. *Natural Language Engineering*.

Levitt, J. M. and Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing and Management*, 47(2):300–308.

Li, Z., Callison-Burch, C., Dyer, C., Khundapur, S., Schwartz, L., Thornton, W., Weese, J., and Zaidan, O. (2009). Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece.

Meho, L. I. and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, 58(13):2105–2125.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, CA.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL 2007 Shared Task. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radev, D. R., Muthukrishnan, P., and Qazvinian, V. (2009). The ACL Anthology Network Corpus. In *Proceedings of the ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore.

Radev, D., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. (2013). The ACL Anthology Network Corpus. *Language Resources and Evaluation*, pages 1–26.