Analyzing Scientific Citation Context: Classification and Temporal analysis

Sikhar Patranabis [1] Samprit Biswas [1]
Sourav Sarkar Subham De [1]
Subham Ghosh [1] Dipayan Mukherjee [1]
De Rhitottam R [1] Kumar Krishna Agarwal [2]

[1] Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur
[2] Department of Mathematics,
Indian Institute of Technology Kharagpur

Contents

Abstract

Automatic recognition of rhetorical functions of citations in scientific text has wide ranging applications in impact calculation, text summarization and informative citation indexer building. In this term project, we intend to build a system for classification citation contexts and correlate it with a standard sentiment classification model. We also investigate temporal tends of citations for scientific papers and correlation with number of citations of the paper.

Chapter 1

Introduction

The primary purpose of citation context classification is to identify the reason as to why researchers cite a particular document. Over the years a number of annotation schemes for citation motivation have been proposed and the question has been studied in detail, even to the level of in-depth interviews with writers about each individual citation [2]. This sustained interest in contexts can be attributed in part to the fact that scientific impact of an author is often measured using bibliometric methods that often take into account only the number of citations received by the author. Such purely quantitative analysis has come under heavy criticism. [4] has pointed out that many citations are done *out of politeness, policy or piety*. Besides, others have emphasized that citations made for criticism or in passing should not be given the same weightage as central citations in a paper, or as those citations where a researcher's work is used as the starting point of somebody else's work [1].

1.1 Aim of the Project

The purpose of this project is to build upon an existing classification scheme and observe the behaviour of the classification with respect to both the number of citations received by the scientific document, as well as the time elapsed since its publication. The citation context annotation scheme used by us in this project has been built by CNRG group, IIT-Kharagpur (Chakraborty et al 2014) It classifies each citation context into one of five major categories. We also correlate the tag based classification with standard sentiment classification results, computed using the *TextBlob* library of Python. Finally, we investigate temporal and citation based trends of the classification results. Our work is along the lines of [3]

Chapter 2

Citation Context Classification Model

The classification model used by us in this project was developed by the CNRG team at IIT-Kharagpur (Chakraborty et al, 2014). The model classifies a given tag into one of the following 5 classes of tags:

- **Alternative Approach (AAN):** The cited paper is a source of an alternative approach but is not compared to.
- **Method (MD):** The cited paper describes some method used in the citing paper.
- **Background (BG):** The cited paper is a background/ Mother paper to the citing paper (like some old paper which founded that theory).
- **Comparable(CM):** The citing paper compares its method to the cited paper's method.
- **Neutral (NE):** None of the other tags fit.

For a given citation context, the model outputs a vector that contains a probability of classification of the context into each of the 5 tags.

2.1 Features Used for Classification

The classification model uses a number of features :

- Lexical Features:
  - *Cue Words*: For a particular class, a list of cue words (e.g., the word "compare" for CM class) are defined that represents that class
  - *N-Grams*: Here, different levels of n-grams (1-grams, 2-grams and 3-grams) are used to see the effect of different word combination.
- Word-level Linguistic Features:
  - *POS*: Part-of-speech (POS) tags of the words in the reference context are used as features.

- *1-gram+POS*: A concatenation of each unigram in the reference context and its associated POS tag constitutes a feature.
- *Tense*: A linguistic feature such as tense related to the main verb of the reference sentence sometimes becomes very effective in identifying the background references.
- *Modal*: The presence of modal verbs (e.g., "can", "may") in a sentence often indicates the strength of the claims described by the reference sentence.
- *Main-verb*: The main-verb of the reference sentence is used as a feature.
- *Root*: Along with the main verb, the root word given by the dependency parser is also used as a word level feature.
- *has-1stPRP*: [Jochim and Schutze 2012] mention that the occurrence of a selected POS could sometimes be an useful feature, such as whether a first person is present in the reference sentence. It is a Boolean feature (True, if first person is present; otherwise False).
- *has-3rdPRP*: Similarly, this feature accounts for the presence of a third person in the sentence.
- Comp/Sup: Comparatives and superlatives (e.g., "more", "better") can help identify CM and AA. It is a Boolean feature that accounts for the presence of any comparative/superlative POS in the reference sentence.
- *has-but*: Another feature is used for the contrastive conjunction "but", which is another clue for the AA class. A boolean feature is set to check its presence in the reference context.
- *has-cf*: It is checked whether an abbreviation such as "cf." (compare to/refer to) is present in the sentence.

- Linguistic Structure Features:
  - *dep-rel*: Dependency Relations (*dep-rel*)were used.
  - *POS Tags*: Seven regular expression patterns of POS tags were used to capture syntactic information.

- Location Features:
  - *Section*: The section of the paper in which the citation is located was used as a feature.
  - *Paper-loc*: This feature is used to check the location of the citation within a paper.
  - *Paragraph-loc*: This feature checks the position of the citation within a paragraph by dividing each paragraph into three portions.
  - *Sentence-loc*: Further incorporation of more fine-grained position information of a citation within a sentence (located in the first quarter, middle half, and last quarter).

- Citation features:

– *Relevance Degree*: Number of citations in the same sentence can indicate the relevance degree of a citation.
– *Density*: Apart from the reference sentence, the number of citations in the reference context, i.e., the preceding and succeeding sentences apart from the reference sentence itself is also considered.
– *avgDensity*: Count of the average number of words per citation in the reference sentence which in turn indicates the average number of words that are devoted to explain each cited paper in the sentence.
– *self-cite*: This feature indicates if one of the citing authors also (co-)authored the cited work.

Chapter 3

Citation Context Sentiment Analysis Model

For sentiment classification, we used *TextBlob* - a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more. The *textblob.sentiments* module contains two sentiment analysis implementations:

- PatternAnalyzer which is based on the pattern library
- NaiveBayesAnalyzer which is an NLTK classifier trained on a movie reviews corpus.

The default implementation is PatternAnalyzer, but one can override the analyzer by passing another implementation into a TextBlob's constructor. For our experiment, we used the PatternAnalyzer.

The PatternAnalyzer takes the citation context string as input and returns a score tuple of the form (*subjectivity*, *polarity*).

- The subjectivity score (S) is a float in the range $[0.0, 1.0]$ with 0.0 indicating highly objective)and 1.0 highy subjective. High subjectivity indicates neutral opinion, such as quoting people etc.
- The polarity score (P) is a float in the range $[-1.0, 1.0]$ with -1.0 indicating highly negative sentiment and 1.0 indicating highly positive sentiment.

For our classification, we use the following sentiment analysis score κ:

$$\kappa = (1 - S)P \tag{3.1}$$

Finally, we use Table 3.1 for sentiment classification:

However, as evident form the results in 4.2 standard sentiment classification is not very suitable for application to scientific citation contexts and does not give very good results. This is partly because in scientific contexts, negativity or disapproval of a technique is not expressed in the same way as in normal text.

κ	Sentiment Tag
0.2 to 1.0	Positive
-0.2 to 0.2	Neutral
-1.0 to -0.2	Negative

Table 3.1: Sentiment Classification Model

Chapter 4

Description of the Data Set used

We collect the AAN dataset [Radev et al. 2013; Radev et al. 2009b; Radev et al. 2009a] which is a collection of all of the papers included in ACL (Association for Computational Linguistics) publication venues. The entire content of each paper is present in XML format where the sentences, paragraphs and the sections are properly separated using different tags. Apart from this, there are two more files – (i) a metadata file: it stores the matadata information of each paper, namely the title of the paper, unique ID of the paper, name of the author(s) of the paper, the venue of publication, the year of publication; (ii) paper-to-reference file: it stores the references of each paper where for each reference, the unique IDs of the citing and cited paper are separated by a tab. The data set used for this project has the citation contexts stored separately in an XML file. The training data is manually tagged by the CNRG group, IIT-Kharagpur.

Table 4.1 describes the manually tagged dataset used for training the classifier.

Number of Cited Papers	348
Number of Citing Papers	28
Number of Citation Contexts	592
Fraction of AAN	0.4628
Fraction of MD	0.2432
Fraction of BG	0.1604
Fraction of CM	0.0253
Fraction of NE	0.1081

Table 4.1: Training Data: Statistics

4.1 Preprocessing of data

We took the citation contexts and performed some basic pre-processing steps on them before feeding them to the classifier. These include:

- Removal of stopwords from citation context text using nltk.
- Removal of proper nouns from the citation context. In particular we have removed all words not belonging to any synset of WordNet.

We also did some basic analysis of the citation context data from the hand-tagged training data set including the following:

- Computing the conditional probability that a word belongs to a tag given any random word
- Computing the top words most closely associated with each tag.

We estimated the conditional probability of tag given word as follows from the training set:

$$P(\text{Tag}|\text{Word}) = \frac{\text{Number of occurrences of Word with Tag}}{\text{Total number of occurrences of Word}} \tag{4.1}$$

We found a set of top words for each of the 5 tags, the results of which are shown in Fig 4.2

Category	Top Words
AAN	researchers, compilation, prior, deficiencies, propose, explored, both, alternative, replicates, address
MD	tagger, extracted, implemented, align, quick, evaluated, estimate, finalized, optimal, paired, paradigm, follows, versions, template
CM	comparing, numbers, main, favorably, finding, interested, participating, chosen, contradict, qualitative, performing, compares, vaguely, evaluating
BG	community, lack, hindered, benefits, involvement, pivot, poor, cost, constructing, shorter, gaps
NE	biomedical, novel, detecting, adding, exploited, substitute, employed, characterized, disentanglement, quality, gale, here, modifiers, represented

Table 4.2: Top Words for Different Categories

4.2 Statistics of citation context classification

Table 4.3 shows some statistics of the test data after classification of the citation contexts.

Number of Cited Papers	10176
Number of Citing Papers	12518
Number of Citation Contexts	77226
Fraction of AAN	0.7159
Fraction of MD	0.1908
Fraction of BG	0.0832
Fraction of CM	0.0002
Fraction of NE	0.0098

Table 4.3: Test Data: Statistics after Classification

Chapter 5

Citation Based Analysis

We analyzed the variation in citation tag trends with the number of citations of scientific papers. We clubbed together papers in similar citation ranges and observed their average citation tag trends as obtained from the classifier. We also performed a basic sentiment analysis of the citation contexts into 3 categories - positive, negative and neutral, and observed the corresponding variation in trends against the number of citations. We break up the citation range into two broad zones- $0 - 250$ and $400 - 600$, since there were the ranges where we have sufficient data for analysis.

Figures 5.1 and 5.2 illustrate the variation in citation tags as output by the classifier for each citation range. The X-Axis shows the number of citations while the Y-Axis depicts the fraction of citations corresponding to each tag. We note that for papers in the citation range $0 - 250$, the majority of the citations received are of category AAN, with MD being the second most frequent. CM is the least common of all tags. The citation range $400 - 600$ shows an interesting trend. In the initial phases, AAN dominates, but with increase in number of citations, MD gains over AAN in terms of dominance. CM is the least common once again.

Figures 5.3 and 5.4 illustrate the variation in citation tags as output by the sentiment classifier for each citation range. The X-Axis shows the number of citations while the Y-Axis depicts the fraction of citations corresponding to each sentiment. Unfortunately, we do not observe any interesting trend, indicating that normal sentiment classification is probably not effective for scientific citation contexts.

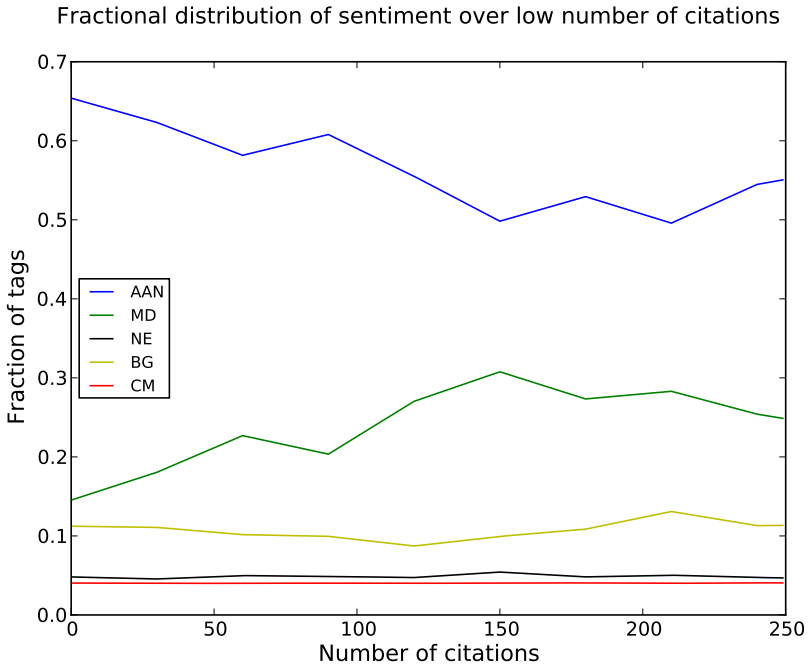

Figure 5.1: Citation Tag Fractions vs Number of Citations: Citation range : $0 - 250$

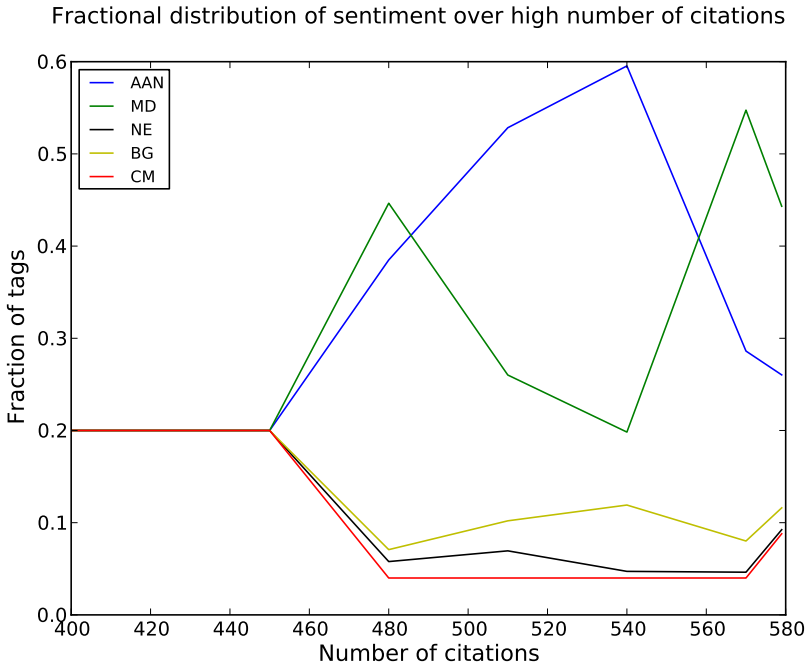

Figure 5.2: Citation Tag Fractions vs Number of Citations: Citation range : $400 - 600$

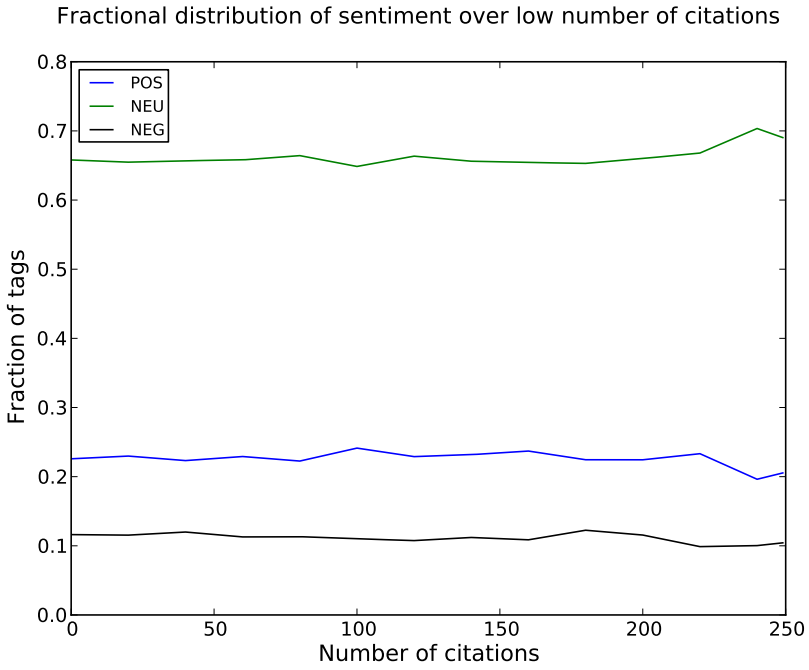

Figure 5.3: Sentiment Tag Fractions vs Number of Citations: Citation range : $0 - 250$

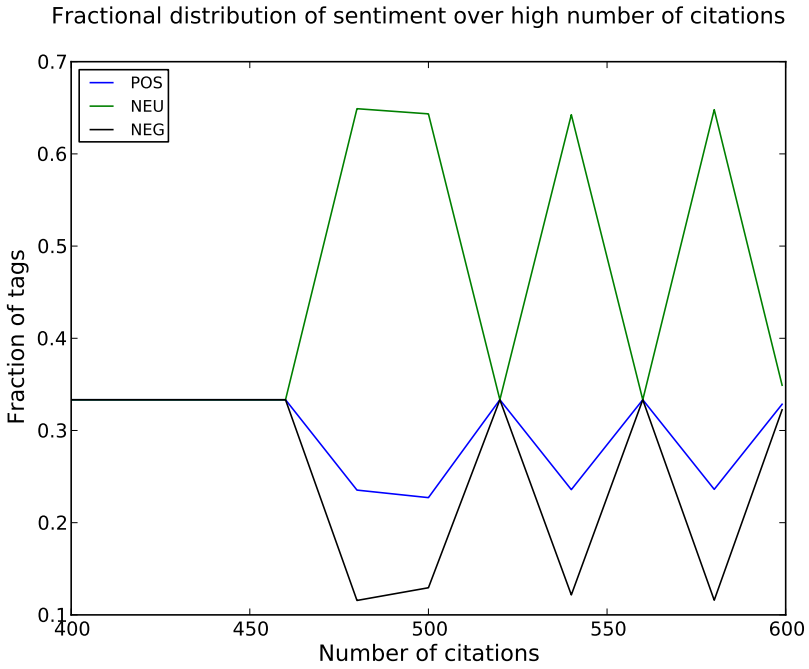

Figure 5.4: Sentiment Tag Fractions vs Number of Citations: Citation range : $400 - 600$

Chapter 6

Temporal Analysis of Citation Contexts

We also analyzed the temporal variation of the citation trends for scientific papers in various citation ranges. We classified papers into 3 major citation zones as follows:

- Low Citations : Number of citations in the range $0 - 100$
- Medium Citations : Number of citations in the range $100 - 300$
- High Citations : Number of citations in the range $400 - 600$

For each citation zone, we obtained the variation in fraction of tags assigned to citation contexts over a period of 30 years from the date of publication of the paper, averaged over all papers in the citation zone. Figure 6.1 shows the average temporal variation in citation tags as output by the classifier over a period of 30 years from the date of publication of a paper. Figures 6.2, 6.3 and 6.4 illustrate the temporal variation in citation tags as output by the classifier for each citation range. The X-Axis shows the number of years following the publication of a paper while the Y-Axis depicts the fraction of citations corresponding to each tag in each year. For the low and medium citation ranges, AAN dominates other tags. In the low citation range, towards 30 years, the BG tag is almost as frequent as AAN, but it again drops steeply. For the high citation range however, the MD tag dominates earlier on, indicating highly cited papers are cited in their earlier years for their methods and only later on are they depicted as an alternative approach.

Figures 6.5, 6.6 and 6.7 illustrate the temporal variation in citation tags as output by the classifier for each citation range. The X-Axis shows the number of years following the publication of a paper while the Y-Axis depicts the fraction of citations corresponding to each tag in each year. In all three citation ranges, the positive sentiment dominates.
’

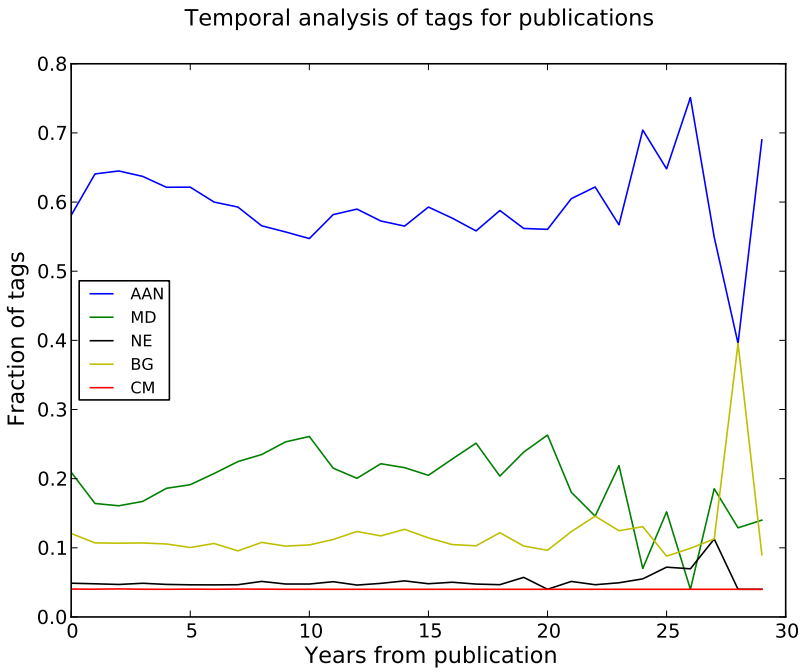

Figure 6.1: Citation Tag Fractions vs Number of years from Publication: Overall

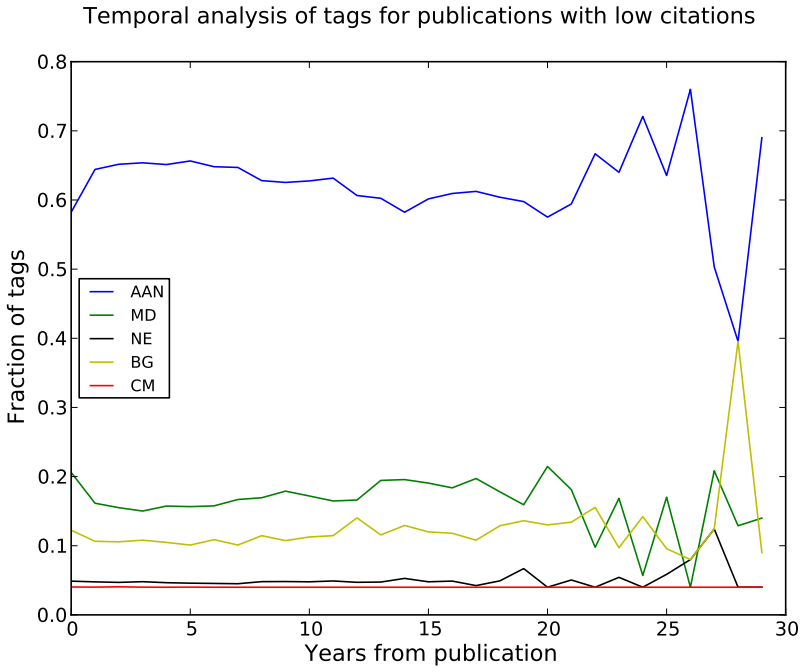

Figure 6.2: Citation Tag Fractions vs Number of years from Publication: Low Citation Range

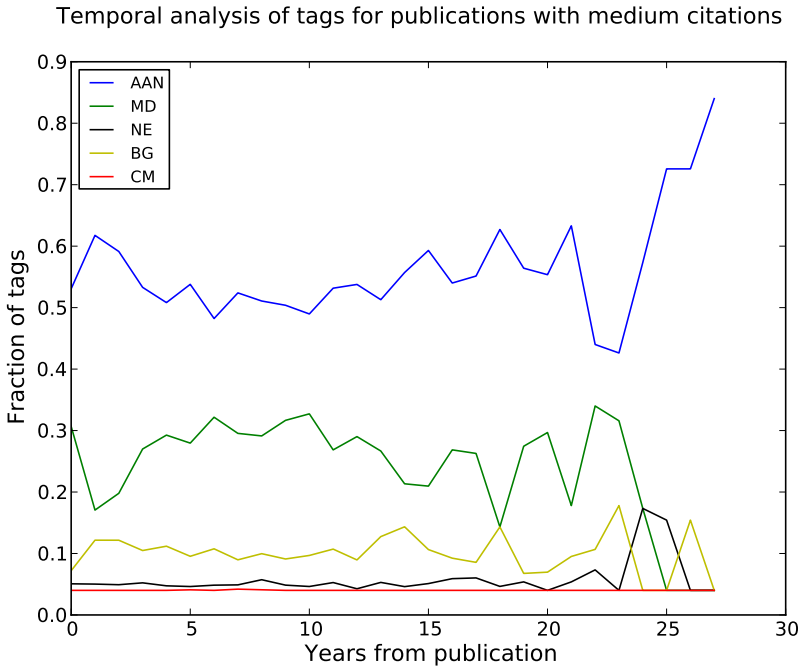

Figure 6.3: Citation Tag Fractions vs Number of years from Publication: Medium Citation Range

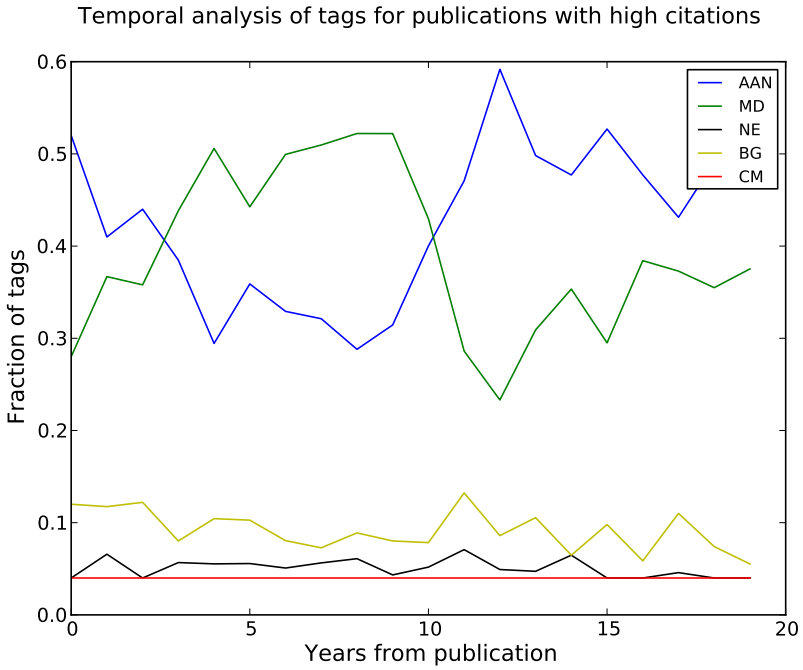

Figure 6.4: Citation Tag Fractions vs Number of years from Publication: High Citation Range

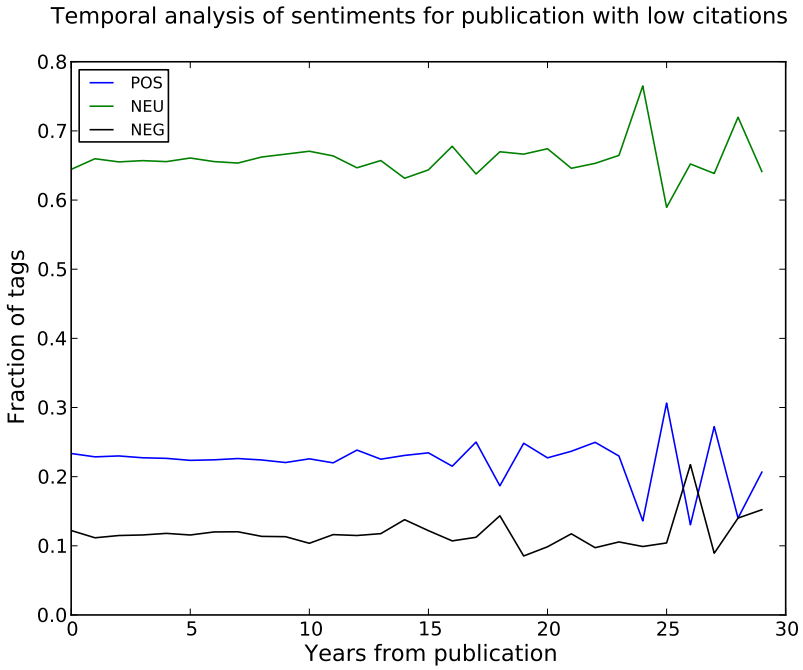

Figure 6.5: Sentiment Tag Fractions vs Number of years from Publication: Low Citation Range

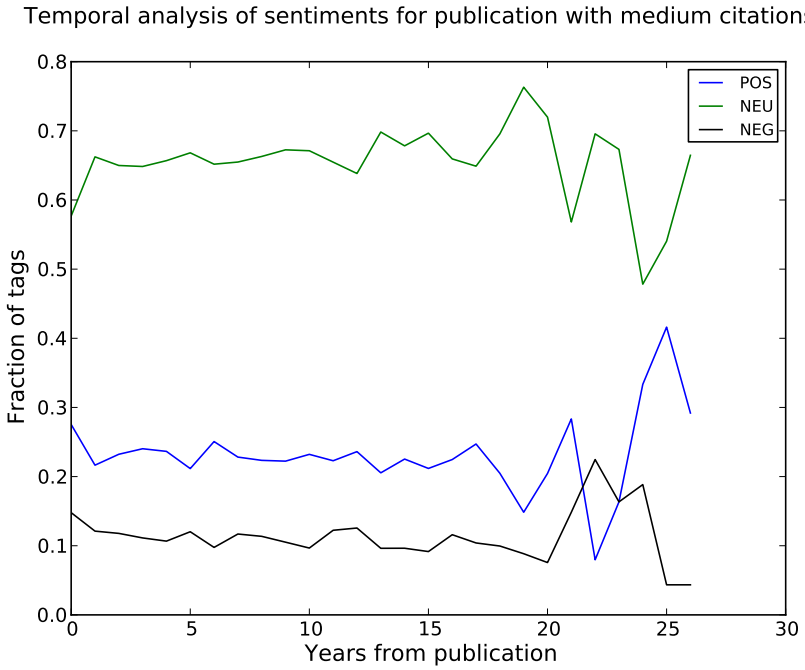

Figure 6.6: Sentiment Tag Fractions vs Number of years from Publication: Medium Citation Range

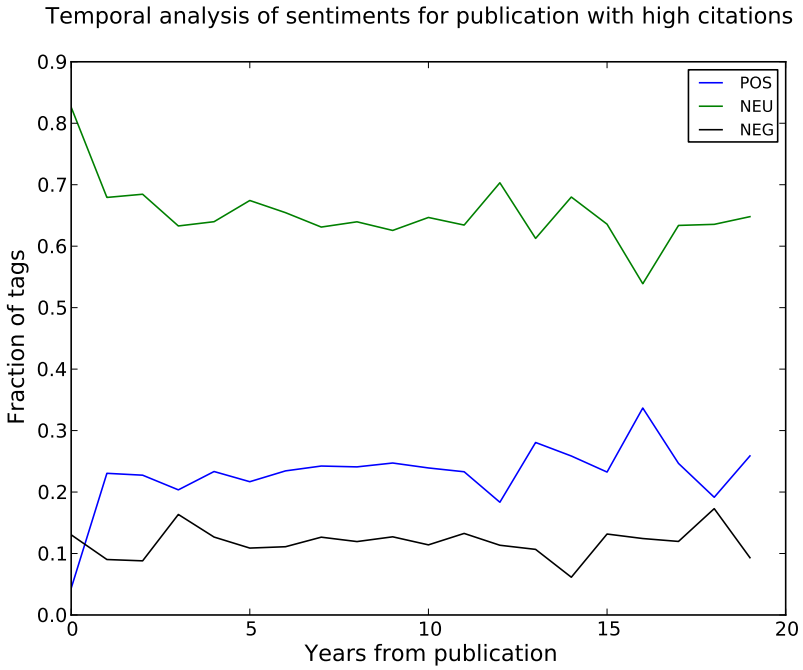

Figure 6.7: Sentiment Tag Fractions vs Number of years from Publication: High Citation Range

Chapter 7

Web Module

The web module consists of a search engine where the user is allowed to enter a query related to a paper and the module displays the corresponding paper(s) matching the entered query. The user can search papers either by keywords in the title or the year of publication. On selecting a particular paper, the module shows the citation trends of the paper starting from its year of publication, along with its number of citations. The time span is decided according to the citation information available for that paper.

Chapter 8

Conclusions

In this project, we have analyzed the citation trends of scientific papers with respect to their number of citations as well as with time. We have also demonstrated that standard sentiment analysis techniques fail to capture these trends when applied to scientific contexts. Our experiments have demonstrated that highly cited papers tend to be cited more as a standard method than as an alternative approach, especially within the first 10 years after publication. Also, highly cited papers tend to be cited as a method more than an alternative approach in general. With a larger data set, we might be able to find some more interesting trends. But for the current data set, these are the most important observations.

Bibliography

[1] Susan Bonzi. Characteristics of a literature as predictors of relatedness between cited and citing works. *JASIS*, 33(4):208–216, 1982.

[2] T.L. Hodges. Citation Indexing: Its Potential for Bibliographical Control., 1972.

[3] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. *Conference on Empirical Methods in natural Language Processing*, 2006.

[4] John M. Ziman. Public Knowledge: An Essay Concerning the Social Dimensions of Science. 1968.

Appendices

Appendix A

Work Division

The following is the work load division between members of the project group:

Member Name	Roll Number	Module
Samprit Biswas	11CS10038	Statistics Extraction
Sourav Sarkar	11CS30037	Statistics Extraction
Subham Ghosh	11CS10046	Statistics Extraction
Sikhar patranabis	11CS10044	Citation and Temporal Based Analysis of Citation Contexts
Dipayan Mukherjee	11CS30045	Citation and Temporal Based Analysis of Citation Contexts
Kumar Krishna Agarwal	11MA20052	Sentiment Analysis of Citation Contexts
Subham De	11CS10046	Design of Web Module
De Rhitottam R	11CS30010	Design of Web Module

Table A.1: Work Division

Appendix B

Feedback

The project helped us gain an insight into how the citation trends vary for papers in different citation ranges. It also gave us an idea about standard natural language processing techniques and their use in real applications. However, our results could have been better with a larger data set.