# CITATION SENTIMENT ANALYSIS

*Identifying sentiments hidden in citations*

# CONTENTS

# “WHAT” AND “WHY” OF CITATION SENTIMENT ANALYSIS

# CITATION ANALYSIS

**WHAT?**

- Citation Analysis is the study of citations in scholarly articles based on their frequency, quality and various other factors.

**WHY?**

- Analyzing citations allow us to detect, analyze and predict
  - Impact of the papers, authors and institutions
  - Trace the interdisciplinary border
  - Access the impact of the research

**HOW?**

- Simplest method of performing citation analysis involves counting the number of times an article has received citations and then measuring its impact.

# DRAWBACKS OF CITATION COUNT

- Only Quantitative
- Measures Popularity; Not Impact
- Unattended factors like criticism not measured
- Misleading and False results
- Lacks Accuracy in the real world

# OVERCOME THE LIMITATIONS

**Study the citations not only *Quantitatively*, but also, *Qualitatively*.**

For studying *Quantitatively*, study citations *syntactically*.

- Learn the “how” and “where” of a citation i.e. See where cited in the paper itself and See how many times cited in one paper

For studying *Qualitatively*, study citations *semantically*.

- Learn the “why” of a citation i.e. Study the citation text to identify the reason/intention/motivation behind the citation; (For Eg. Does is it appraises the cited text, or criticizes it?)

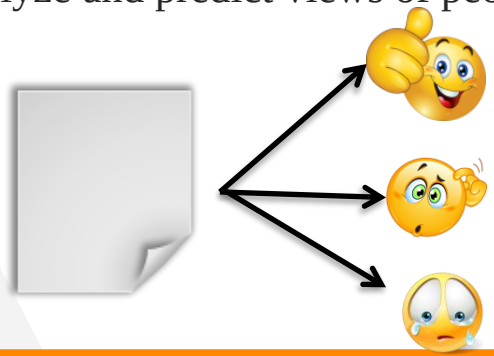
# SENTIMENT ANALYSIS

**WHAT?**

Sentiment Analysis also known as Opinion Mining is a process of computational linguistics that identifies and categories opinions present in a piece of text based on their polarity.

**WHY?**

Allows us to understand analyze and predict views of people towards any entity.



# SENTIMENT ANALYSIS MEETS CITATION ANALYSIS

- Citation Sentiment Analysis the task of analyzing the sentiment behind citations.
- Following are some examples of negative, positive and neutral citations.

*Eg: "While SCL has been successfully applied to POS tagging and Sentiment Analysis (Blitzer et al., 2006), its effectiveness for parsing was rather* ***unexplored****."*

*(A) Negative Citation*

*The window length was motivated by recent research CIT (Qazvinian and Radev, 2010)* ***which shows the best score*** *for a four sentence boundary when detecting non-explicit citation.*

*(B) Positive Citation*

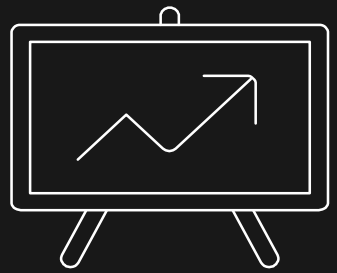
*Piao et al. CIT (2007) proposed a system to attach sentiment information to the citation links between biomedical papers by using existing semantic lexical resources.*

*(C) Neutral Citation*

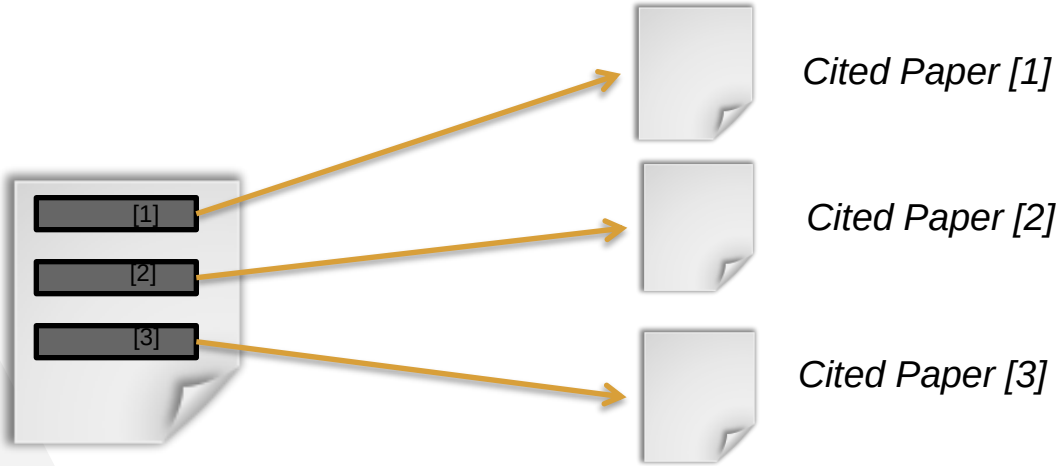
# WHY DO WE NEED CITATION SENTIMENT ANALYSIS?

- Tells us sentiment behind the citation i.e. Determines why the citation was made in the first place.
- Not dependent on what masses say, but what each individual says.
- True indicator of one's work
- Thorough process of analysis
- Let's us know the detail of citation as, it is an *inside out* study of citations.
- Few examples of its benefits are:
  - Pure study especially of poor citations; Since all citations do not mean good news.
  - No misleading conclusions.
  - Rules out the references made out of pity, or courtesy.

# BASIC TERMINOLOGY



# TERMINOLOGIES (a) TYPES OF PAPERS




***Citing Paper***
( Papers that *cite/ refer* others )

***Cited Paper***
( Papers that get *cited/ referenced* )

# TERMINOLOGIES (b) TYPES OF CITATIONS

**Explicit Citations :** The sentences where the citations have been mentioned explicitly i.e. referenced specifically. For eg:

*" While SCL has been successfully applied to POS tagging and Sentiment Analysis (Blitzer et al., 2006), its effectiveness for parsing was rather unexplored."*

**Implicit Citations :** They involve mentions of other research papers and authors with informal or indirect references and no proper explicit reference. For eg:

*"Exploring more lexical features in a later work, Wiebe and Riloff (2005) developed a Naive Bayes classifier using data extracted by a pattern learner. This pattern learner was seeded with known subjective data."*

# APPROACHES

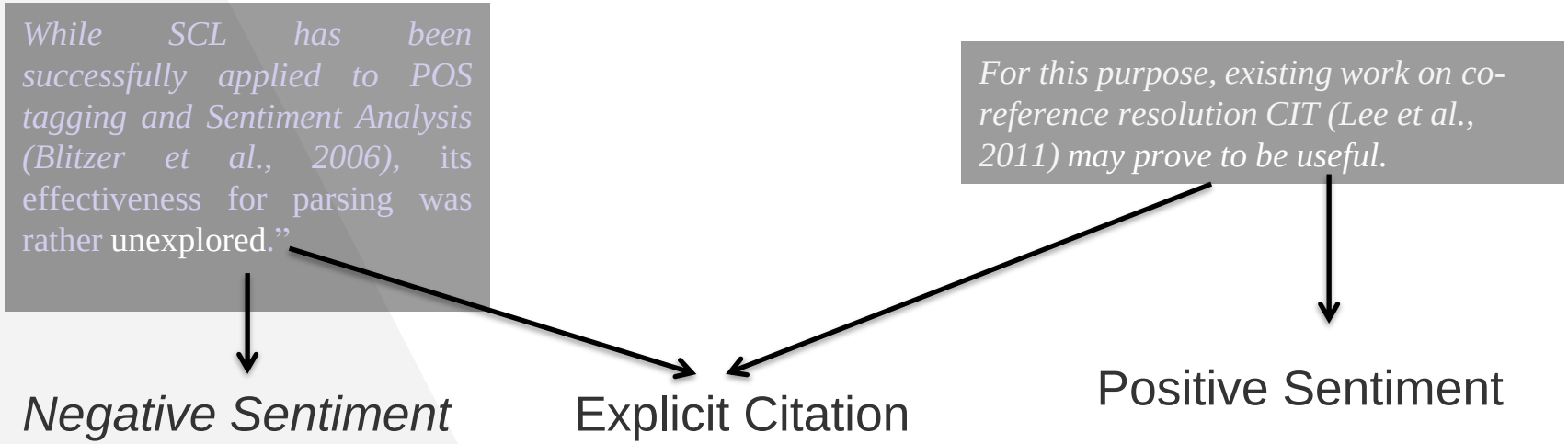
CITATION SENTIMENT ANALYSIS

- Sentence Level
- Context Level

# SENTENCE LEVEL CITATION SENTIMENT ANALYSIS

This involves categorizing explicit citation sentences into positive, negative and neutral, only by looking at the sentences in which the citation has been made.

The following example explains it.



# CONTEXT LEVEL CITATION SENTIMENT ANALYSIS

- Context level analysis means analyzing the citation context i.e. the text surrounding the explicit citations.
- Involves two major tasks namely- Context detection and Sentiment analysis of context.

**STEP ONE:**

CONTEXT DETECTION

- Detecting all the explicit and implicit references in the paper.
- Context is used as input for sentiment analysis.

**STEP TWO:**

SENTIMENT ANALYSIS OF CITATION CONTEXT.

- Classification of citation context into namely three classes- Positive (p) , Negative (n), Neutral or Objective (o).
- An ‘x’ class for sentences with no citations.

# WORK DONE

# BASIC STEPS FOR CITATION SENTIMENT ANALYSIS

Extraction of Sentences

Manual Annotation

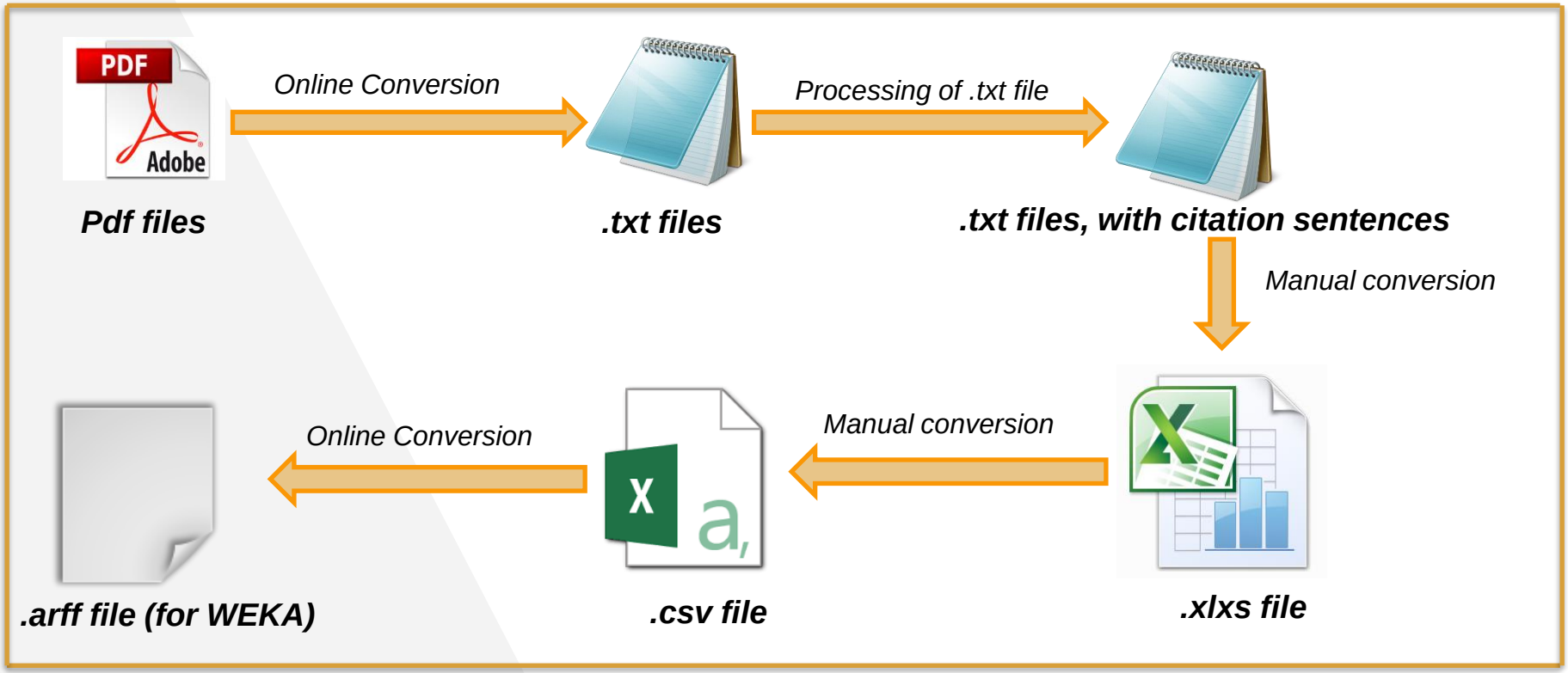
Sentiment Analysis

# CONSTRUCTION OF CORPUS

- **Requisites** : **Approach to be followed**
  - Two approaches: Manual and Automated
  - If Manual, choosing small corpus is better
  - If Automated, big corpus can be chosen too
  - We used manual and automated both.

- **Basic Ingredient** : **Data**
  - Here, research papers.
  - Choose an approach before choosing the data.
  - Choose a familiar topic; Helps in annotating sentiment
  - Conversion into accessible format, especially if automated.
  - We took 4 papers (Less, because manual work too)

# DIAGRAMATIC REPRESENTATION FOR CORPUS CONSTRUCTION



# PROCESSING OF TEXT FILE TO GET EXPLICIT CITATIONS

Extraction of explicit citation sentences:

- **Approach** : Using grep command with regular expression
- **Prerequisite knowledge** : Type of citations,(here harvard style)
- **How** : Use regular expression to search and extract citation sentences

***.txt files, with all sentences*** *(D.txt)*

***% $ grep ' [(]*[[A-Z a-z]*[, . et. al.]* [[:digit:]][[:digit:]][[:digit:]][[:digit:]]]*)\| ([[:digit:]] [[:digit:]][[:digit:]][[:digit:]])\|[(]*[A-Z a-z ]*[, . "et. al.]* [[:digit:]]][[:digit:]]][[:digit:]] [[:digit:]]; ' D.txt > E.txt***

***.txt file with explicit citation sentences*** *(E.txt)*

# MANUAL ANNOTATION OF CORPUS

The thoughts that I publish in what follows are the precipitate of philosophical investigations which have occupied me for the last sixteen years. They concern many subjects: the concepts of meaning, of understanding, of a proposition and sentence, of logic, the foundations of mathematics, states of consciousness, and other things. I have written down all these thoughts as remarks, short paragraphs, sometimes in longer chains about the same subject, sometimes jumping, in a sudden change, from one area to another. ⊖ Originally it was my intention to bring all this together in a book whose form I thought of differently at different times. But it seemed to me essential that in the book the thoughts should proceed from one subject to another in a natural, smooth sequence.

After several unsuccessful attempts to weld my results together into such a whole, I realized that I should never succeed. The best that I could write would never be more than philosophical remarks; my thoughts soon grew feeble if I tried to force them along a single track against their natural inclination. ⊖ And this was, of course, connected with the very nature of the investigation. For it compels us to travel criss-cross in every direction over a wide field of thought. ⊖ The philosophical remarks in this book are, as it were, a number of sketches of landscapes which were made in the course of these long and meandering journeys.

*Sentences*

*Reads the text*

*Assigns polarity to each sentence*

Sentence 1:……..(p)
Sentence 2:……..(o)
Sentence 3:……..(p)
Sentence 4:……..(n)
Sentence 5:……..(n)
Sentence 6:……..(o)
Sentence 7:……..(n)

*Annotated Sentences*

# MANUAL ANNOTATION OF CORPUS

- As shown in Fig, Manual Annotation means to manually assign sentiment polarity to the corpus, i.e. classify the citation context as Positive, Negative and Objective.
- Why manual annotation used because :
  - More accurate than automated systems
  - Sets a baseline for automated systems; Thus, helps in analysis.
  - As training set when using SA Classifiers
  - Comes in handy when testing a new theory

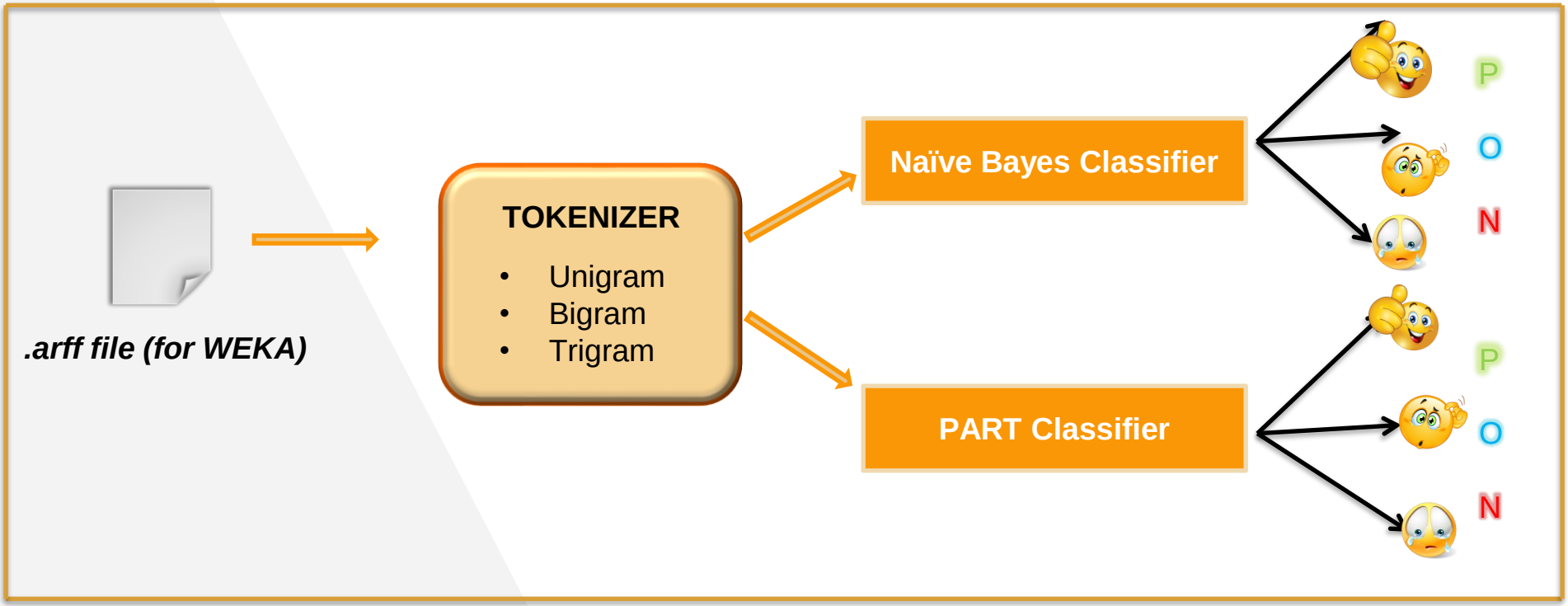
- We used manual annotation for explicit citation and used it to train the classifiers. The sentences were classified as follows:

| **Polarity** | Positive (P) | Negative (N) | Objective (O) |
|---|---|---|---|
| **No. of Sentences** | 27 | 8 | 104 |

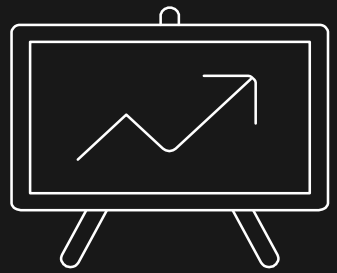
# SENTIMENT ANALYSIS

- Carried out by WEKA tool.
- Data pre-processed by WEKA. Stop words and Stemmers removed.
- Feature used for Tokenizing the sentences is N-grams – All Unigrams, Bigrams, Trigrams used.
- 10 Cross-fold Validation used for training the classifier.
- Classifiers used : (1) Naïve Bayes, and (2) PART Classifier
- Metrics used for analysis : Precision, Recall, F-Score.

# CLASSIFICATION USING NAÏVE BAYES AND PART CLASSIFIER



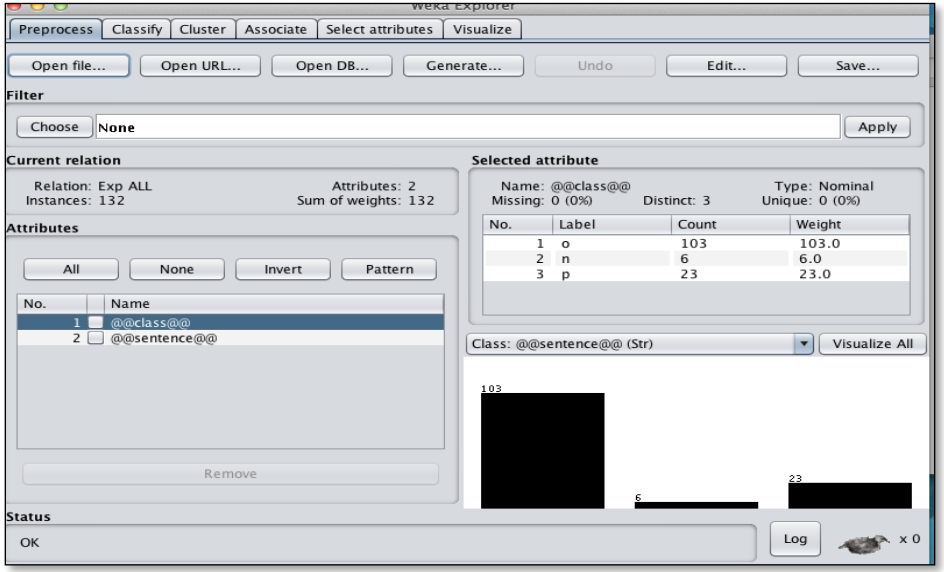
# RESULTS



# RESULTS OF SENTIMENT CLASSIFICATION

The Results obtained are:

| **Polarity** | Positive (P) | Negative (N) | Objective (O) |
|---|---|---|---|
| **No. of Sentences** | 23 | 6 | 103 |



# ANALYSIS BY WEKA

***The better classifier out of the two, according to our corpus, is PART classifier with an accuracy of 80.303% and F-Measure of 0.891.***

- In Naïve Bayes classifier, best accuracy was obtained in the case of bigrams and trigrams.

| Features | Correctly Classified sentences | Precision | Recall | F-Measure |
|---|---|---|---|---|
| **Unigram** | 77.2727 % | 0.779 | 0.990 | 0.872 |
| **Bigram** | 78.0303 % | 0.780 | 1.000 | 0.877 |
| **Trigram** | 78.0303 % | 0.780 | 1.000 | 0.877 |

- In case of PART classifier, accuracy increases when going from unigrams to bigrams, but decreases , when trigrams were used.

| Features | Correctly Classified Sentences | Precision | Recall | F-Measure |
|---|---|---|---|---|
| **Unigram** | 71.9697 % | 0.791 | 0.883 | 0.835 |
| **Bigram** | 80.303 % | 0.810 | 0.990 | 0.891 |
| **Trigram** | 76.5152 % | 0.777 | 0.981 | 0.867 |

# FUTURE SCOPE

- **Better annotations** will result in better results.
- Sentiment detection can be improved.
- The detection of context can be optimized for different window sizes.
- For efficient detection of context, **features** can be added such that they perform extraction of relevant context and also, work equally well on any corpus domain.
- Need for detecting **citations outside the context window**.
- Automatic summarization, it is another promising avenue for further research.

# CONCLUSION

- This study about Citation Sentiment Analysis, was intended for finding out sentiment behind citations. Since, nowadays discussions are getting more and more expressive, studying sentiment closely for such reference becomes more vital.
- What we infer from those results is the fact that, explicit citation sentences are mostly objective in nature and most of the sentiment is covertly coiled in the surrounding sentences.
- Besides, by comparing the results of Weka with the literature survey about the types of classifier used, we can say that the classifiers and features are highly corpus dependent. Citation sentiment analysis is a budding field which needs to be explored thoroughly to fully utilize its power in the research arena

# REFERENCES


- Athar, A. and Teufel, S., 2012, June. 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* (pp. 597-601). Association for Computational Linguistics.
- Athar, A., 2014. Sentiment analysis of scientific citations. Technical Report, University of Cambridge, Computer Laboratory,(UCAM-CL-TR-856).
- Murray, J.,. Finding Implicit Citations in Scientific Publications: Improvements to Citation Context Detection Methods. Master's Thesis at CSC. Dept. of Computer Science, KTH Royal Institute of Technology,. Sweden, 2015.
- Ying Ding, Guo Zhang, and Tamy Chambers, 2014. Content-Based Citation Analysis: The Next Generation of Citation Analysis, *Journal Of The Association For Information Science And Technology* 65(9):1820–1833.
- G. Vinodhini, RM.Chandrasekaran. 2012 June. "Sentiment Analysis and Opinion Mining: A Survey". *International Journal of Advanced Research in Computer Science and Software Engineering* ,Volume 2, Issue 6.

# THANKS!

*Presentation by:*
*Vasundhra Dahiya*
*Urvashi Choudhary*