



Universitat d'Alacant
Universidad de Alicante

Concit-Corpus: Context Citation Analysis to
learn Function, Polarity and Influence

Myriam Hernández-Álvarez



Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

TESIS DOCTORAL

Septiembre, 2015

Concit-Corpus: Context Citation Analysis to learn Function, Polarity and Influence

Myriam Hernández-Álvarez

UNIVERSIDAD DE ALICANTE

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

DEPARTAMENTO DE LENGUAJES SISTEMAS
INFORMÁTICOS

Concit-Corpus: Context Citation
Analysis to learn Function, Polarity
and Influence

Myriam Hernández-Álvarez

Memoria presentada para aspirar al grado de
DOCTORA POR LA UNIVERSIDAD DE ALICANTE

**MENCIÓN DE DOCTORA
INTERNACIONAL**

DOCTORADO EN APLICACIONES DE LA INFORMÁTICA

Dirigida por:

José Manuel Gómez Soriano

Dedicatoria

A mi madre, mi amado ángel guardián.

A mis hijos Andrés, Edgar y Mónica que son mi fortaleza y mi bendición.



Universitat d'Alacant
Universidad de Alicante

Agradecimientos

Ante todo doy gracias a Dios por haberme concedido la oportunidad de realizar esta investigación para contribuir en algo, en un campo del conocimiento fascinante como es el análisis de función, polaridad e impacto de literatura científica.

Mi corazón va hacia mi madre, compañera constante a lo largo de mi vida, que ahora sigo sintiendo a mi lado. Sin ella nada habría sido posible.

Mi profunda gratitud al Dr. José Manuel Gómez Soriano por su constante guía y apoyo. Gracias por estar siempre disponible para responder, corregir y ayudar. Haberlo elegido como mi Director de Tesis fue un gran acierto.

Mi reconocimiento a mi Alma Máter, la Escuela Politécnica Nacional por brindarme los medios para realizar esta investigación y a la Universidad de Alicante por darme la oportunidad de trabajar en el doctorado.

Finalmente, gracias a Diego Pérez, a Paúl Gualotuña, a José Luis Escobar por su soporte técnico, al Dr. Josafá Pontes por sus ideas respecto al tipo de anotación y por su participación en los experimentos para obtener los índices de acuerdo entre anotadores.

Abstract

Citation Context Analysis serves not only to classify citations according to qualitative criteria for obtaining information that can be used to improve impact assessment, but also has other applications as summary generation and development of better information retrieval techniques, among others. In the present study, we focus mainly in the application first mentioned.

In the Introduction section, we refer to many authors that have exposed the weakness of impact assessment using approaches with only citation counting, although these methods have proved to be very useful to valuate author and publishing media impact, they could be improved taking into account results from citation context analysis. Not all citations should have equal weight in the calculation of the impact, which is why we could use context citation analysis for considering quality criteria, such as the purpose and polarity, to differentiate citations.

Some authors have commented on the distortions that can occur when are used only counting methods to valuate impact. Marder et al. (2010) showed that controversial articles with incorrect or incomplete data get higher citation counts. This situation could generate unethical incentives to publish, considering that there is a lot of pressure over the academics, pressure produced by the importance of impact assessment on the researchers' career and on the image and relevance of the different scientific media. Therefore, it is necessary to differentiate the nature of each citation with complete citation criteria from citation context analysis.

Articles and other scientific documents are types of texts that have specific characteristics. We can mention the following: authors are not explicit about citation purpose; likewise, rarely they are clear about polarity, but rather they avoid criticism in order to evade adverse reactions from colleagues, and conceal negative feelings about work of others through language resources called "hedges"; many citations are mentioned for reasons outside research interest; and, finally there is specialized lexicon for each knowledge area (Verlic et al., 2008). Context citation analysis must take into account these features. We can see that this is a task with specific and complex challenges.

In Chapter 2, in the state-of-the-art, we noticed topics that are still non-resolved such as context size detection; implicit references recognition; definition of features to annotate; hedges detection to distinguish disguised negativity; and, over all, there is the necessity to overcome the absence of a common framework to facilitate research progress in collaborative conditions. This framework should include a standard classification, an annotation scheme, and an annotated corpus according such scheme. In fact, in the state-of-the-art of the present work, it was concluded that the biggest problem facing researchers in this field, is that there is no public available annotated corpus that responds to a medium or high granularity scheme in order to contain enough information for context analysis that can be used on a shared basis by scholars.

In this thesis, we addressed most of these unresolved issues. For instance, we defined a scheme for citation classification that takes into account function, polarity and impact. The scheme was designed to maintain a simple structure with six functions and three levels of polarity, that when combined with keywords and labels yield high granularity, comparable with complete ontologies as CiTO. Ciancarini (2014) noticed that this kind of ontologies present difficulty for annotation due to their complexity. However, with our proposed scheme and its structure, understanding and application of the scheme are facilitated. Anyway, the annotation task for high granularity remains challenging, but our scheme at least makes it easier that annotators use and take advantage of all possibilities of the scheme.

We applied the proposed scheme for annotating a citation corpus composed by 85 articles taken randomly from ACL Anthology with a total of 2195 bibliography cites. Using it, we could evaluate the impact that a citation has in a document. For this purpose, we propose two methods: in the first one, we take into account, in a directly way: negative, neutral or positive polarity to assign an impact category corresponding to *Negative*, *Perfunctory* or *Significant*, respectively. We justify this decision considering that authors have a more favorable disposition towards citations that have greater impact in their works. The second method is an algorithm developed based on criteria obtained from previous studies that link the impact with citation frequency; location in the document; number of sections in which citations appear; some functions and polarity. From this procedure, we obtain a classification using the same three impact categories as stated before.

The algorithm results are compared with data obtained from a survey applied to authors who rate citation impact in their own articles. Our method produces good

outcomes and has very similar results when we compared 161 impact citations obtained with our algorithm to the corresponding survey responses. We can see that our algorithm captures and relates the most important criteria to be considered for evaluate the impact, because when we match the algorithm results with data obtained in the mentioned survey, we observe a weighted average for *F-Measure* of 0.93, which is a very satisfactory value that demonstrates an excellent correlation between author's annotation and the results of the algorithm.

To continue with our experiments, and in order to use all the generated features, this procedure is applied to train a SVM with SMO classifier with entries that include every function, polarity, frequency, location and number of sections in which each citation occurs. Results have 0.98 weighted average for *F-Measures*. Consequently, we could train a classifier that uses the features of our corpus in order to recognize our algorithm and automatically classify impact in three levels: *Negative*, *Perfunctory* and *Significant*. The resulting classification has good values of Precision, Recall and F-Measure. We consider that this is a contribution that could be applied to incorporate quality assessment factors to impact valuation, in order to obtain holistic evaluation criteria where not all citations are considered equal for counting.

To annotate the corpus, we classify citation function and polarity according to the suggested scheme using an annotation methodology that includes a step of pre-annotation in which keywords and tags are detected to clarify and standardize an internal representation that a coder or annotator creates about citation context. With this method, the mental model is more likely to coincide with the ones produced for other coders and consequently we obtain a good rate of inter annotator agreement. With this pre-annotation step, we dramatically improve the agreement among annotators index, which is indispensable to validate rating reliability and reproducibility of the annotation scheme. We validate this index in Chapter 6, where we can see that the values of Fleiss' Kappa are as high as 0.862 for function and 0.912 for polarity. These values correspond to an almost perfect agreement in accordance to the scale of Landis and Koch (1977). Using keywords and labels, we obtain a notorious improvement, because without this stage, with the same annotators, there were low results for this index: 0.386 and 0.259 for function and polarity respectively due to the complexity of the annotation task for a high granularity scheme.

Regarding the context length for classification, in the annotation results, we noticed in that the context length chosen by coders largely corresponds to just one statement: the one with the citation. With less frequency appear length context of two or three sentences.

It is probably that the context should not include more than three sentences to cover all the necessary information about the reference.

In chapter 7, we evaluate the annotated features as useful entries in a classifier for citation function and polarity. Results rated high for Precision, Recall and F-Measure, which demonstrate suitability of the chosen features for those classifications. We chose algorithms after the recommendations of our initial state-of-the-art study. In our survey, the most suggested algorithms were SVM trained with SMO and Naïve Bayes. In our results, SVM – SMO has the best values: 0.833, 0.828 and 0.825 corresponding to Precision, Recall and F-Measure in function classification. These results are higher than the state-of-the-art. For polarity categorization, with the same algorithm again we obtained excellent values for Precision, Recall and F-Measure: 0.886, 0.882 and 0.88.

In Chapter 7, we also present experiments for establish the best combination of training and test sets that have to be independent samples. For our corpus size, is important that the test set has an appropriate size, especially for functions and polarity less frequent. For independence of training and test samples, it is better to choose the Percentage Split option of Weka, and for size considerations, we found that the best proportion between training and test samples is 66% vs 44%.

Due to the reliability that is obtained in our corpus annotation, we suggest that the data continue to be annotated manually using our methodology. We state that it is necessary to improve current automatic annotation techniques for obtaining reliable results when applied to an annotation scheme with medium or high granularity. Mandya (2012) categorizes annotation schemes in two classes: the one that uses manual classification and the one that has automatic feature extraction and classification. In that study, we observe that manual classification schemes have medium or high granularity while automatic processed schemes have low granularity. As we said, annotated corpora with a medium or high granularity provide valuable information indispensable to citation context analysis, but its annotation is a complex task, even for human annotators. Therefore, challenges for automatic annotation are big. According to our state-of-the-art study, the schemes with medium or high granularity showed in Table 1, are manually labeled by their authors; in studies that attempted automatic labeling of this kind of corpora, results come out not as good as it is necessary for having reliable data.

It remains as an important line for future work, the improvement of automatic labeling of corpora according to fine granularity patterns. The effort is justified because

this type of annotated corpus provides essential information that is required for citation context analysis.

In summary, among our contributions in the present work, we can mention the following: a proposed annotation scheme simple in its structure, but with high granularity; the annotation methodology, particularly as regards to the pre-annotation process to detect keywords and labels that are useful to create mental models, they also serve as input features in classification tasks; a public available annotated corpus that contains those features and is accessible for collaborative work; a method to evaluate citation impact using criteria exposed in other works and algorithms developed in our thesis; and, the experimental finding that the significant context around a citation usually takes no more than three sentences including the one with the mention.

The XML files for our annotated corpus will be available in the University of Alicante digital repository¹. Initially we have 85 annotated papers and 2195 citations, over time the corpus will continue to be populated with new annotated documents.



Universitat d'Alacant
Universidad de Alicante

¹ <http://hdl.handle.net/10045/47416>

Tabla de contenido

1. Introduction	1
2. State-of-the-Art.....	7
2.1 Citation Context Identification and Detection of Implicit Citations.....	7
2.2 Citation Function Classification	11
2.3 Citation Polarity Classification	16
2.4 Available corpora.....	18
2.5 Chapter conclusions.....	20
3. Esquema de clasificación	23
3.1 Criterios para clasificación.....	24
3.2 Comparación de nuestro esquema con la ontología CiTO.....	32
3.3 Conclusiones del capítulo	37
4. Evaluación del Impacto de una cita	39
4.1 Impacto relacionado directamente con la polaridad de las citas.....	40
4.2 Impacto calculado por ubicación, repetición, función y polaridad de las citas	43
4.3 Uso del aprendizaje automático para calcular el impacto	48
4.4 Conclusiones del capítulo	50
5. Metodología de anotación	53
5.1 Pre-procesamiento y definición del contexto.....	56
5.2 Pre-anotación.....	59
Ejemplo de pre-anotación 1	65
Ejemplo de pre-anotación 2	66
5.3 Clasificación de la función y la polaridad	67
Ejemplo de clasificación para la Función Based on, Supply.....	67
Ejemplo de clasificación para la Función Useful.....	68

Ejemplo de clasificación para la función Acknowledge, Corroboration, Debate.....	68
Ejemplo de clasificación para la Función Contrast.....	68
Ejemplo de clasificación para la Función Weakness, Correct.....	69
Ejemplo de clasificación para la Función Hedges	69
5.4 Procesamiento automático para valoración de impacto	70
5.5 Conclusiones del capítulo	70
6. Validez del esquema	73
6.1 Acuerdo entre anotadores.....	74
6.2 Organización de los experimentos y datos.....	74
6.3 Comparación de resultados con y sin pre-anotación de patrones.....	75
6.4 Conclusiones del capítulo	82
7. Clasificación automática usando el corpus anotado.....	85
7.1 Organización de los experimentos y datos.....	86
7.2 Resultados con SVM y Naïve Bayes.....	87
7.3 Análisis de resultados	106
7.4 Conclusiones del capítulo	114
8. Conclusiones y trabajo futuro.....	117
9. Publicaciones relevantes.....	121
References	123
ANEXO 1: Guía de anotación	133
i. Introducción y Esquema	133
ii. Procedimiento de anotación.....	136
iii. Etiquetas para la anotación y palabras clave de ejemplo.....	142
iv. Ejemplos de anotación para cada función	146
a. Based on, Supply.....	146
b. Useful.....	148
c. Acknowledge, Corroboration, Debate.....	151
d. Contrast.....	153
e. Weakness, Correct	155

f. Hedges	157
ANEXO 2: Archivos para el cálculo del acuerdo entre anotadores.....	161
ANEXO 3: Análisis de Factores para la valoración del Impacto con resultados de encuesta a autores.....	167



Universitat d'Alacant
Universidad de Alicante

Índice de figuras

Figura 1: Granularidad con una función.....	27
Figura 2: Criterios de clasificación de una cita de acuerdo a la función polaridad e impacto de sus referencias	27
Figura 3: Relación entre impacto y polaridad.....	41
Figura 4: Impacto de acuerdo a la polaridad de las funciones desagregadas	42
Figura 5: Algoritmo para valoración del impacto aplicando el Esquema de Clasificación propuesto	46
Figura 6: Proceso de anotación del corpus	55
Figura 7: Tamaño del contexto seleccionado por los anotadores	58
Figura 8 : Notación simbólica de los patrones.....	65
Figura 9 : Comparación de resultados sin y con pre-anotación de funciones	81
Figura 10: Comparación de resultados sin y con pre-anotación de polaridad de citas	81
Figura 11: Esquema de la Base de Datos MySQL.....	86
Figura 12: Citas por función	91
Figura 13: Conteo de citas por polaridad.....	92
Figura 14: Conteo de citas por función para cada polaridad	93
Figura 15: Valores de F-measure para función, en relación a los tamaños de las muestras de entrenamiento y pruebas con SVM.....	107
Figura 16: Valores de F-measure para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con SVM.....	108
Figura 17: Valores de F-measure para función, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes.....	109
Figura 18: Valores de F-measure para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes.....	110

Figura 19: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%.....	111
Figura 20: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%.....	112
Figura 21: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%.....	113
Figura 22: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%.....	114
Figura 23: Criterios de clasificación de una cita de acuerdo a su función, polaridad e impacto.....	135



Universitat d'Alacant
Universidad de Alicante

Índice de tablas

Tabla 1: Review of categories for citation functions and polarity classifications	13
Tabla 2: Available annotated corpora for citation polarity classification	19
Tabla 3: Esquema de anotación para funciones.....	28
Tabla 4: Esquema de anotación para polaridad.....	30
Tabla 5: Esquema de anotación para impacto	30
Tabla 6: Funciones desagrupadas por combinación función-polaridad.....	31
Tabla 7: Correlación entre las 41 propiedades de la ontología CiTO y nuestro esquema de clasificación	33
Tabla 8: Propuesta para valoración del impacto usando solo la polaridad.....	41
Tabla 9: Evaluación del impacto de las citas. Comparación entre anotación realizada por los autores y los resultados de nuestro algoritmo	47
Tabla 10: Matriz de confusión entre autores y nuestro algoritmo.....	47
Tabla 11: Valoración del impacto usando SVM con SMO en WEKA.....	49
Tabla 12: Matriz de confusión	50
Tabla 13: Tamaño del contexto seleccionado por los anotadores.....	57
Tabla 14: Número de sentencias previas y posteriores a la cita dentro del contexto	58
Tabla 15: Etiquetas	62
Tabla 16: Acuerdo entre anotadores sin pre-anotación correspondiente a la Función	78
Tabla 17: Acuerdo entre anotadores sin pre-anotación correspondiente a la Polaridad.....	79
Tabla 18: Acuerdo entre anotadores con pre-anotación correspondiente a la Función.....	79
Tabla 19: Acuerdo entre anotadores con pre-anotación correspondiente a la Polaridad ...	80
Tabla 20: Referencia para describir acuerdo entre anotadores según valor de Kappa	82
Tabla 21: Citas por función.....	90
Tabla 22: Citas por polaridad.....	91

Tabla 23: Conteo de polaridad en cada función	92
Tabla 24: Resumen de la Evaluación (test Split 66% vs 44%)	93
Tabla 25: Precisión detallada por clase	94
Tabla 26: Matriz de confusión	95
Tabla 27: Resumen de la evaluación (test Split 90% vs 10%)	95
Tabla 28: Precisión detallada por clase	96
Tabla 29: Matriz de confusión	96
Tabla 30: Resumen de la evaluación (test Split 66% vs 44%)	97
Tabla 31: Precisión detallada por clase	98
Tabla 32: Matriz de confusión	98
Tabla 33: Resumen de la evaluación (test Split 90% vs 10%)	99
Tabla 34: Precisión detallada por clase	99
Tabla 35: Matriz de confusión	100
Tabla 36: Resumen de la evaluación (test Split 66% vs 44%)	100
Tabla 37: Precisión detallada por clase	101
Tabla 38: Matriz de confusión	101
Tabla 39: Resumen de la evaluación (test Split 90% vs 10%)	102
Tabla 40: Precisión detallada por clase	103
Tabla 41: Matriz de confusión	103
Tabla 42: Resumen de la evaluación (test Split 66% vs 44%)	104
Tabla 43: Precisión detallada por clase	104
Tabla 44: Matriz de confusión	105
Tabla 45: Resumen de la evaluación (test Split 90% vs 10%)	105
Tabla 46: Precisión detallada por clase	106
Tabla 47: Matriz de confusión	106
Tabla 48: Valores de F-measure para función, en relación a los tamaños de la muestras de entrenamiento y pruebas con SVM	107
Tabla 49: Valores de F-measure para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con SVM.....	108

Tabla 50: Valores de F-measure para función, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes	108
Tabla 51: Valores de F-measure para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes.....	109
Tabla 52: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%.....	110
Tabla 53: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%.....	111
Tabla 54: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%.....	112
Tabla 55: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%.....	113
Tabla 56: Esquema de anotación para funciones	134
Tabla 57: Esquema de anotación para polaridad	134
Tabla 58 : Funciones desagrupadas por combinación función-polaridad	135
Tabla 59: Etiquetas	142
Tabla 60: Ejemplos de etiquetas y palabras clave asociadas a funciones agrupadas.....	143
Tabla 61: Ejemplos de palabras clave asociadas a polaridad.....	145
Tabla 62: Resultados de la anotación para cálculo del acuerdo entre anotadores.....	161
Tabla 63: Impacto calculado con nuestro método vs. Impacto según autores.....	167

1. Introduction

Citation analysis is a method of evaluating the impact of an author, a published work or scientific media. Sugiyama, Kumar, Kan and Tripathi (2010) suggested two types of citation analyses: citation counts (Garfield 1972) and citation context analysis. We accept this categorization as it is close to our thesis.

Currently, scientific literature is mostly available on the Web, where automatic citation indexes are combined with online searches using tools like CiteSeerX², Google Scholar³ and Microsoft Academic Research⁴. Indexes obtained by counting citations provide valuable information on the behavior of papers such as impact factor and H index (Hirsch 2005). The impact factor measures citation frequency for a journal in a particular year, while the H index is an indicator of author's reputation, evaluating his/her scientific production depending on the citation count. Small (1973) proposed co-citation analysis to add a similarity measure between works A and B by counting the number of documents that cite them. This approach is complementary to the concept introduced by Kessler (1963) regarding "bibliographic coupling" which states that documents that have similar references dealt with similar subjects.

Many authors have noted the weakness of citation counts because it is purely quantitative and does not differentiate between high and low citing papers. PageRank (Page, Brin, Motwani and Winograd 1999) partially solved this problem with a rating algorithm that takes into account not only counting of cited mentions but also the influence of citing journals. Sayyadi and Getoor (2009) went a step further and proposed a

2 <http://citeseerx.ist.psu.edu/>

3 <http://scholar.google.com/>

4 <http://academic.research.microsoft.com/>

FutureRank which is the expected future PageRank score, based on upcoming citations, with a procedure that tried to predict future reference counting, using authorship network and the publication time of the article. In the same field of Citation Count Prediction (CCP), Yan, Tang, Liu, Shan, and Li, (2011) suggested paper topics and author expertise among other features as input for several regression models to foretell the estimated citation counts of that article after a given period of time. Davletov, Aydin and Cakmak, A. (2014) dealt with CCP using citation count of the article when was just published, modeling it as a regression problem.

Mei and Zhai, (2008, June) extracted impact summaries, defining impact as influence of the paper in topics related to its research subject as reflected in its references.

Citation count analysis has valuable application in development of bibliometric measures. Tools as Science Citation Index⁵ or Journal Citation Reports⁶ use impact factor to rank scientific journals. These indexes are useful and have had a profound impact on science management. However, there are controversies regarding their simple approaches that consider all citations as equal regardless of the purpose or the polarity with which they were mentioned.

Radicchi (2012) showed that incomplete, erroneous or controversial papers have higher citation counts. This can generate perverse incentives for new researchers who may be tempted to publish incomplete or erroneous results as a means of receiving a higher number of citations (Marder, Kettenmann and Grillner 2010). In fact, this affects the quality of prestigious journals because it is known that accepting controversial articles is profitable and increases the overall citation numbers. Reviews, such as those conducted by the recently awarded Nobel Prize Winner (Sample 2013), emphasise this issue. In this context, Brazilian journals have used self-references to skew the Journal Citation Reports index (Van Noorden 2013).

Another limitation of quantitative citation analysis is that all references are interpreted as if an author were influenced by the work of another, without specifying the type of influence (Zhang, Ding, and Milojević 2013); this approach can omit the true impact of a citation (Ioannidis 2005; Young, Ioannidis, and Al-Ubay 2008; Marder, Kettenmann, and Grillner (2010); Nicholson and Ioannidis 2012; Radicchi 2012; Brembs and Munafò 2013; Schreiber 2013). To understand the influence of a scientific paper, it is advisable to account for other, more qualitative criteria, such as an author's disposition

5 <http://scientific.thomson.com/products/jcr/>

6 <http://purl.org/spar/cito>

towards a cited paper or the function of a citation in a paper. Not all citations are the same; for instance, a criticised quoted work does not have the same impact as a citation used as a conceptual starting point for a research paper. In the same sense, Reyhani, Kim, Lee and Kim (2013) asserted that the citation number obtained by a paper is not a sufficient indicator to understand its influence; more important is to know their contribution to citing papers, so, they proposed to calculate influence using similarity measures between citing and cited articles.

These problems add to the increasing importance of impact indexes in researchers' careers (Siegel and Baveye 2010). Pressure to publish seems to be the cause of increased fraud in scientific literature (Fang, Steen and Casadevall 2012). For these reasons, it is becoming more important to correct these problems and identify more complete metrics to evaluate researchers' relevance by taking into account other 'quality' factors related to citation context analysis. Citation context analysis seeks to find a contextual relationship between citing papers and cited papers using different methods that include natural language processing techniques, machine-learning systems, annotated datasets, statistical methods and dictionary approaches.

Citation context analysis applied to bibliometrics is reaching a point from which it will be possible to glimpse significant progress that could be applied to produce a more accurate index that includes qualitative and quantitative aspects for evaluating the relevance of science publications and researchers' work. For this type of assessment, additional information, such as the author's intention and disposition towards the cited work, is required. These criteria are related to citation functions and polarity. Automatic summarisation and survey generation of scientific text and more complete citation indexers for web browsing are also of interest. Hence, it is important to conduct a detailed study of the current state of this issue to guide future research. Therefore, we evaluate the recent developments in the most researched areas in citation context analysis: citation context identification, citation functions and polarity classification with their corresponding schemes and datasets.

Citation context identification is an important task in citation analysis. This approach is related to the identification of all parts of the text that are associated with a specific citation. Athar (2014) shows that including citation context in polarity citation analysis considerably increases the performance of citation classification algorithms.

According to Green, Ashley, Litman, Reed and Walker (2014), "argumentation mining, is a relatively new challenge in corpus-based discourse analysis that involves

automatically identifying argumentative structures within a document e.g., the premises, conclusion, and argumentation scheme of each argument..." Teufel and Moens (1997, 2002); Teufel et al. (1999); Teufel (2010) studied argumentative classification including context citation. We think that argumentation mining could serve as method to delimitate citation context, separating text that have arguments about a citation from text that are not related. It would be insightful to conduct studies with this focus so that detection Context definition is treated as a topic linked to argumentation mining. Thus, we suggest that, as another approach, the length of the citation context could be established using argumentation mining algorithms, to automatically detect the argument around the citation. Nevertheless, despite recent advances in discourse parsing and causality detection, the automatic recognition of argumentation structure of texts is still a very challenging task (Peldszus 2014).

The context is also related to the structure of discourse in the paper. Teufel (2010) and Athar (2014) state that accounting for the location of the citation in the article would aid the classification of the citation's function. Athar (2014) suggests that it is likely that the most relevant citations are referenced several times in the paper, particularly in the middle of the main body. This feature captures the number of citations in an article, and it is an important characteristic for modelling the influence of a cited article within the citing work.

Citation functions are associated with the task of the reference in the citing paper. Citation function classification uses different schemes and categories for different citation roles; note that there is not a consensus on a single scheme. These patterns are used to annotate corpus taken from scientific literature to train models for automatic citation classification.

Citation polarity is a more coarse-grained analysis that captures positive (favourable) or negative (unfavourable) dispositions towards a cited paper. Sometimes polarity can be mapped to a citation function.

In general, citation context analysis addresses special types of documents in which it is difficult to detect the function of the reference or the author's disposition towards a quote because:

- Authors do not always explicitly state the citation's purpose.
- Authors avoid explicit criticism, particularly when it cannot be quantitatively justified. They often accomplish criticism through implications and hedging, defined

as cautious language expressed in a vague form (Hyland 1996). MacRoberts and MacRoberts (1984) maintained that showing negative sentiment in a citation text is politically dangerous, and researchers prefer to disguise criticism toward the paper by praising its strengths in advance of a negative review if any.

- Many citations are simply mentioned as work that has been performed in the field. Ziman (1987) stated that many papers are cited out of ‘politeness, policy or piety’. They are passing references that are not really influential to the citing paper.
- There is a specific sentiment-related science lexicon that is primarily a compound of technical terms that diverge among scientific fields (Verlic, Stiglic, Kocbek and Kokol 2008). Often, academic expressions are specific to the knowledge in a field. Efforts to connect sentiments to technical terms and to develop general sentiment lexicons for science are underway (Small 2011).

For an integral methodology of citation context analysis, it is important to differentiate among references according to citation function and polarity; and, from there, using also additional information such as location and citation frequency it is possible to obtain an assessment of cited document influence on citing paper. A holistic impact index should include not only citation counts but also an individual valuation of influence; to achieve this it will be necessary to overcome the aforementioned challenges.

In this work, we present a study of citation context analysis to rank its function, polarity and impact using a corpus developed according to our own classification scheme. Understanding the use of references in a document is essential to conduct a successful research. Citation context analysis facilitates to achieve this understanding because it provides useful information such as purpose or function of the citation in the document; favorable or critical author’s disposition towards it; and, provides criteria for impact rating. This is the aim of the present Thesis, and, in this Chapter, we defined the problem and indicated the importance of solving it.

In Chapter 2, we present a survey of state-of-the-art in citation context analysis, context identification, function citation classification, polarity citation classification, and the available corpora for these tasks.

In Chapter 3, we explain our classification scheme focused on recognition of function, polarity and impact of citations. Annotators handle easily our scheme even though it has fine granularity when combining different classification criteria.

In Chapter 4, we propose two methods that contribute to valuate citation impact using concepts as function, polarity, location and frequency of citations in a document.

In Chapter 5, we describe our methodology for corpus annotation that includes a pre-annotation process. This step allows that annotators build similar mental models, to achieve a good agreement, condition that ensures reliability and reproducibility of the annotation. This pre-annotation process also serves to generate features that will be used in the citation classification according to the proposed scheme.

In Chapter 6, the scheme and annotation methodology is validated from the point of view of inter-annotator agreement. We compare results using and not using pre-annotation. We demonstrate that because of the task complexity, it is very difficult to get a good annotator agreement without this step. The significant improvement of inter-annotator agreement justifies the extra work done by manually annotate labels and keywords before citation classification.

In Chapter 7, we classified citations according to function and polarity using our scheme and annotated features. We use SVM algorithm trained with SMO incorporated in WEKA.

We present conclusions and trends of future work in Chapter 8. In Annex 1 we present a detailed annotation guide that functions in a completely stand-alone way. In Annex 2, we show a file for annotation results of three annotators that worked in the same data. With this information, we computed the inter-annotator agreement. In Annex 3, we present a file with information used to develop an impact valuation algorithm.

2. State-of-the-Art

2.1 Citation Context Identification and Detection of Implicit Citations

Citation context is defined as the text surrounding references in scientific literature that are related to them. Context in which a citation appears provide valuable data about it (Kataria, Mitra, and Bhatia 2010).

Several works have identified the optimal size of context windows for detecting the sentences referring to the citation. Kataria, Mitra, Caragea and Giles (2011) selected an adaptive window for context around a citation that statistically provides more information. They found window size through maximizing an objective function representing topical similarity between cited document and its corresponding context; where topics are semantically significant clusters of words presented as co-occurrences (Nallapati, Ahmed, Xing, and Cohen 2008).

Ritchie, Robertson and Teufel (2008, October) explored how to select context length to improve information retrieval of specific data. They experimented with variable window sizes ranging from one sentence that included the citation, to the entire text using linguistic motivations. They obtained better results with three sentences than with one, and stated that simple windows are more effective than ones with full sentences.

Recommendation systems use context detection techniques that can be applied in this domain. Using citation context they identify relevant papers that may lead to specific

recommendations of sections that need references (Livne, Gokuladas, Teevan, Dumais and Adar 2014).

He, Kifer, Pei, Mitra, and Giles (2011) analyzed methods to find citation context to make automatic recommendations to locate references. They used language models with 100 word-overlapping windows, each one with a context probability score. The score sequence identifies high probability regions that could be defined as citation context. They use contextual similarity around different parts of the document with topic relevance given citation context clusters and a probabilistic model to compute relevance scores.

As suggested before, we advocate for another approach in which detection of context should be related to the identification of the argument around the citation; thus, the context should be connected with the argumentation mining field. Manual processing or algorithms to automatically detect arguments around a citation could define its context. These arguments could be refuted or supported to infer different citation functions and polarity.

Current research is rather limited to experimental methods to find optimal fixed-sized windows of context. Most of the current efforts use machine-learning algorithms with supervised approaches. Unsupervised methods are less commonly used to define context, mainly because of the complexity of the task, which translates to a large set of rules that are difficult to handle; these methods are probably incomplete because they cannot cover all possible cases.

An important limitation of this approach is the size of the corpora used by qualitative researchers. These researchers tend to use relatively small datasets because generation of a specialised citation corpus is difficult and requires careful reading, professional knowledge and expert judgment (Zhang et al. 2013). Automatic annotations for generating large labelled corpus have generally produced poor accuracy.

Teufel, Siddharthan and Tidhar (2006) worked on 2,829 sentence citation corpus from 116 articles using a 12-class classification scheme by applying context in annotations to determine an author's purpose for citing a text. The process was manual and based on rhetorical information. The authors established that sentences containing the citation often do not contain information about the relationship among the citing and cited papers; thus, it is necessary to define contextual sentences before and/or after the citation. Notably, in the context, frequent citations are implicitly named.

Qazvinian and Radev (2010) classified sentences as contextual/non-contextual using three sentence-level features (similarity to reference, explicit citation, matches particular regular expressions) to train Markov Random Field and SVM models. The authors found the best results in a context window comprising four sentences before and after the reference. Note that their dataset consisted of 10 papers. Their F-macro value measure for the recognition of two classes, context and no context, was of 0.87.

Corpus size delimitation was also conducted by Small (2011) who manually processed 20 papers of the citation summary data from the Association for Computational Linguistics Anthology Network ACL AAN⁷ (Bird, Dale, Dorr, Gibson, Joseph, Kan, Lee, Powley, Radev and Tan 2008); to extract citation and co-citation contexts using a ‘bag’ of sample cue words to characterise dispositions towards cited work in maps of science. In their work, the context consisted of an average of 1.6 sentences around the point of reference.

Athar and Teufel (2012) worked on a supervised approach with context windows of different lengths. They developed a classifier using citations as features sets with SVM and n-grams 1-3 units in length and dependency triplets as features. The authors explored the effect of context window lengths and obtained the best performance for citation classification using an annotated context window of 4 sentences. They showed that ignoring the citation context would result in a classification with more non-identified polarity due to loss of information, especially the criticism towards a cited paper. They recognise the need for better algorithms to filter contextual sentences.

Abu-Jbara and Radev (2012) addressed the problem of identifying fragments of a citing sentence that are related to a target reference, i.e., the reference scope. Their intent was to identify which segment refers to a given citation in a sentence with multiple citations. They approach context identification differently because they considered a citing sentence as one that cites multiple papers. To date, little is known about different contributions of each co-citation in an article. The authors compared three methods: word classification with SVM, sequence-labelling CRF, and segment classification. CRF-based sequence-labelling performed significantly better than the word classification method. Their results also showed that segment labelling performs better than word labelling. They used a corpus formed by 3,500 sentences that each contained at least two references; the material was randomly selected from papers found in the ACL Anthology Network corpus (ACL AAN). However, this work only refers to one citing sentence, rather than a broader context.

⁷ Released Dec. 2013 <http://clair.eecs.umich.edu/aan/index.php>

Another supervised technique is conditional random fields (CRF) to tag, segment and extract information from documents. The technique has been applied by some authors to identify context. Angrosh, Cranefield and Stanger (2013) manually annotated text in 20 articles as a development dataset and implemented CRF to identify citation contexts. They established a model with 7 context types for citation sentences and six no-context sentences. They applied the technique to three datasets, where each was composed of 10 articles; they obtained the best results for citation context identification, with an F1 value of 89.5%.

A relationship exists between the location of a citation and its polarity and its importance within the article cited. Thus, it is necessary to take into account an established structure that normally has scientific papers, for example, IMRaD from Biber and Finegan (1994). Liakata, Saha, Dobnik, Batchelor and Rebholz-Schuhmann (2012) generated a tool with support vector machines and conditional random fields to train and compare 256 articles; 11 categories were recognised using Core Scientific Concepts (CoreSCs) to define context using text locations in various paper sections: Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion. As discussed below, their results should be improved upon.

As an example of unsupervised methods for citation context identification, we suggest the work of Kang and Kim (2012). Their corpus consisted of 56 articles from different academic journals that broadly belong to science and engineering fields; 1,048 citing sentences were yielded that used citation pattern rules with grammar presented in the Backus-Naur Form to detect citation contexts as phrases, clauses, sentences, multi-sentences and other characteristics (e.g., figures, equations and tables). Their manual experiments showed that multi-sentence citation context was only detected in 5% of the total citing sentences. This result differs from most in that the majority of experimental results suggest that to detect the polarity of references, the context size should be at least more than one statement (Jochim and Schütze 2012). The authors did not automatize the detection of citation contexts, but they define rules and suggest future work for using the contexts in a machine-learning model.

Athar and Teufel (2012) focused on features to automatically detect implicit reference mentions in citation contexts and showed that their inclusion improves the quality of classification. Features analysed were the use of acronyms, cue phrases that signal continuation of the topics related to the citations, and use of third-person pronouns,

among others. The resulting macro-average F-score was 0.754 for detecting implicit citations and 0.687 for polarity recognition.

The results attained with the different approaches of context citation detection considerably vary. However, the results are not actually comparable due to the diversity of the used measures and corpora.

The most common measures are macro and micro F-scores (Qazvinian and Radev 2008; Athar and Teufel 2012; Kang and Kim 2012) and average precision (Abu-Jbara and Radev 2012; Athar and Teufel 2012; Angrosh et al. 2013). Small (2011) used a log-likelihood method to measure probability and find cue words in groups that expressed dispositions towards cited work.

Moreover, the use of different corpora is a large problem because, even though different measures can provide, a slight indication of which systems are better, the use of incompatible corpora makes formal comparisons very difficult (Hernández and Gomez 2014). Standard resources are not available, and most of the corpora created or used in different studies have not been published.

There are some problems still to be resolved in the definition of citation context length. Automatizing detection of context length is a complex and yet incomplete task. In this field it could be applied context detection techniques developed for automatic recommendation systems or argumentation mining methods. However, until these approaches be fully developed, it is useful to apply fixed number of sentences as defined by experimental results.

2.2 Citation Function Classification

A citation's purpose reflects the author's intention when they added the reference. A citation function is related to the task of the reference in the citing paper. Here, the terms are often interchangeable.

Most citation classification approaches are not general applications but are rather heavily oriented towards specific science domains (Dong and Schäfer 2011). Almost all the research in this field corresponds to computer science and language technologies, while some research is associated with life science and biomedical fields.

Several authors have reviewed the contributions of the extraction and classification of citation functions in papers. Their objective was to identify citation functions or to detect if a citation was somewhat influential in the citing article (Teufel et al. 2006; Sugiyama et al. 2010; Dong and Schäfer 2011; Abu-Jbara and Radev 2012; Zhu, Turney, Lemire, and Vellino 2014) by considering that a reference is not only intended to acknowledge the influence of other authors but that it is also has non-scientific motivations. It is important to differentiate one citation function from another to define which references are relevant to evaluate impact indexes.

Studies have focused on three issues: manual annotation of a corpus using different category and function schemes with different approaches; automatic labelling of the corpus from training data; and addressing the problem of defining features that will achieve the best results in citation classification. However, some results of the various studies should be improved upon.

Additionally, how a citation function is categorised is important. Many proposals require human classification to differentiate citation function classes based on several types of criteria. A category review for citation function classification by several authors is shown in Table 1. Different annotation approaches present diverse levels of granularity in citation function definitions. These schemes define 3 to 35 different classes. Less granularity often refers to polarity (positive, negative, or neutral/objective), as we will discuss later. Garzone (1997) and Garzone and Mercer (2000) proposed a 35 category scheme. Teufel (2000) established a basic scheme with 3 categories and a full scheme with 4 more. Teufel, Siddharthan and Tidhar (2009) defined a 12 category pattern that also maps polarity and function. Most citations are objective/neutral, even though Teufel et al. (2006) claim that making a reference is itself an act of acknowledgment. In general, fine-grained categories are suitable for applications, such as summary extraction and information retrieval with qualified citation indexing.

Another interesting issue related to citation function is hedging. Mercer, DiMarco and Kroon (2004) conducted an introductory study that only applied a few citation context data as hedging cues that were studied by Hyland (1998), to detect the dissembled intent of citations. In Mercer et al. (2004) frequency analysis was used to demonstrate that hedging cues tend to occur within the citation context and may help determine the purpose of a citation. Experimental work in this field has yet to be conducted.

Jochim and Schütze (2012) used the model of Moravcsik and Murugesan (MM), 1975 with four facets to establish whether cited work is an idea or a tool, whether it is accurate

or faulty, whether it is fundamental or perfunctory, and whether the citing paper builds upon or is an alternative to the cited paper. The authors stated that each of the Moravcsik and Murugesan facets can be used according to the type of classification needed. The facet that established whether the work was accurate or faulty (confirmatory/negational) can be mapped directly to polarity. The Stanford Maximum Entropy classifier (Klein and Manning 2003) was applied in which context lengths of 1, 2 and 3 sentences are identified. The best results were obtained for sentence lengths greater than 1. The authors reported a good feature analysis for each facet and offered their annotated corpus to the scientific community (Table 2). In their work, they used a combination of features: unigrams, word-level linguistic features, comparatives, sentence location and a lexicon with positive and negative sentences. Their F1 scores varied as follows: 68.2 (discriminating between an idea and a tool); 51.1-52.9 (work that evolved from or was an alternative to a cited paper); 58.0 for organic (fundamental) or perfunctory research; and 51.1 for confirmational/negational facets.

Table 1: Review of categories for citation functions and polarity classifications

Author(s)	# of categories	Categories for the annotation scheme
Garzone 1997	35	7 negational; 5 affirmational; 4 assumptive; 1 tentative; 5 methodological; 3 interpretational/developmental; 1 future research; 2 conceptual; 2 contrastive; 4 reader alert.
Teufel 2000	7	Basic scheme: background, other, own. Full Scheme: basic scheme + aim, textual, contrast, basis.
Teufel et al. 2009	12	Negative polarity: weakness of cited approach; unfavorable contrast/comparison. Neutral polarity: contrast/comparison in goals of methods; contrast/comparison in results; contrast between two cited methods; neutral description of cited work, or not enough textual evidence for above categories or unlisted citation function. Positive polarity: author uses cited work as starting point; author uses tools/algorithm/data; author adapts or modifies tools/algorithm/data; this citation is positive regarding approach or problem addressed; author's work and cited work are similar; author's work and cited work are compatible/support each other.
Dong and Schäfer 2011	4	One dimension of Moravcsik and Murugesan (MM), organic vs. perfunctory divide into background, fundamental idea, technical basis, comparison.
Liakata et al. 2012	11	Core scientific concepts: hypothesis, motivation, goal,

		object, background, method, experiment, model, observation, result and conclusion.
Jochim and Schütze 2012	4	Apply Moravcsik and Murugesan (MM) scheme with four orthogonal facets that determined whether the reference is an idea or a tool, organic or perfunctory, building upon the cited work or presenting an alternative, confirmative or negational.
Angrosh et al. 2013	7 Non-citation; 7 Citation Sentences.	Non-citation sentences: background; issues; gaps; description; current work outcome; future work. Citation sentences: cited work identified gaps; cited work overcomes gaps; uses outputs from cited works; result with cited work; compare works of cited work; shortcoming in cited work; issue related cited work.
Li et al. 2013	3 polarity 12 function	Positive: based upon, corroboration, discover, positive, practical, significant, standard, and supply. Neutral: contrast, co-citation, neutral. One negative function.
Ciancarini et al. 2013	13	Agrees with, cites, cites as author, cites as authority, cites a data source, cites as evidence, cites as metadata document, cites as potential solution, cites as recommended reading, cites as related, confirms, corrects, critiques, derides.
Iorio et al. 2013	3	Positive, negative, neutral.
Meyers 2013	2	Corroborate and contrast

Dong and Schäfer (2011) tested textual, physical (location, number of references, etc.) and syntactical features in 120 articles from Association for Computational Linguistics Anthology Reference Corpus ACL ARC⁸ (Radev, Muthukrishnan and Qazvinian 2009) as training data. The results produced F1 macro values of 0.66 for citation function classification (fundamental idea/technical basis/comparison). They used a semi-supervised automatic data annotation with an ensemble-style self-training algorithm in approximately 170 training instances. The context was as many as three sentences.

Teufel et al. (2006) used cue words, part-of-speech (POS) tags and different length n-grams to train a WEKA⁹ machine learner IBK algorithm, and they obtained a macro F of 0.57 for replicating a human annotation for function classification, which is a difficult task. Their results for polarity classification presented a macro F of 0.71. As she noted, annotation is reliable at 0.72.

8 <http://acl-arc.comp.nus.edu.sg/>

9 <http://www.cs.waikato.ac.nz/ml/weka/>

Meyers (2013) classified two functions: corroborate and contrast. Corroborate describes two works that follow the same approach, while contrast describes two differing approaches or opinions. The author used a lexical dictionary and manually written rules to create an algorithm based on tree-modelled discourse. The results were acceptable, i.e., 216 correct answers for 291 citation relations were obtained. The algorithm was applied to 20 PubMed¹⁰ scientific articles, with Recall values of 67% for contrasting and 83% for corroborating papers. These results need to be proved in a larger corpus.

Ciancarini, Iorio, Di, Nuzzolese, Peroni and Vitali (2013) annotated a corpus for citation functions using citation contexts and 13 function categories. They did not report experimental results of their corpus, but they recognised that their work is preliminary.

Using a similar approach, Iorio, Di, Nuzzolese and Peroni (2013) implemented a citation-extraction algorithm through pattern matching. They detected polarity using CiTalO, a software developed using a combination of techniques of ontology learning from natural language, sentiment analysis, word-sense disambiguation (with SVM), and ontological mapping. The corpus was 18 papers published in the seventh volume of the Balisage Proceedings¹¹ (2011) regarding Web mark-up technologies. With the highest Recall of only 0.491, the results need improvement. CiTalO implements properties of CiTO¹² ontology.

Li, He, Meyers and Grishman (2013) used a maximum-entropy-based system to create a training set from annotated data and automatically classify functions into two levels of granularity; three polarity categories were linked to the sentiment of 12 citation functions. The authors used 91 annotated articles with 6,355 citation instances. The best F1 performance was 0.67.

Abu-Jbara, Ezra and Radev (2013) used a trained classification model and SVM classifier with a linear kernel and AAN dataset. They selected 30 papers that received a total of 3,500 citations within AAN in 1,493 papers. For development data, they created another dataset with 300 citations that cite five papers from AAN to determine the optimal citation context window. The purpose (function) classification results had a macro F of 58%. For automatic classification, the accuracy would probably increase as more training data becomes available (Abu-Jbara et al. 2013).

10 <http://www.ncbi.nlm.nih.gov/pubmed>

11 <http://balisage.net/Proceedings/vol7/cover.html>

12 <http://vocab.ox.ac.uk/cito>

In general, the best results for citation function or purpose classification were obtained using a SVM algorithm, with F1 values as high as 73.1, depending on the selected features.

The main unresolved problems in citation classification is the lacking of a standard categorization scheme for functions, and the development not yet accomplished of a public access golden corpus with sufficient features to perform citation classification that yield good results in Precision, Recall and F-Measures. Currently, low values for these results are obtained, because for most schemes their training annotated corpus is small and doesn't have the features needed to provide sufficient information to the classifier algorithms for their task.

2.3 Citation Polarity Classification

Polarity is defined as a coarse-grained approach to an author's sentiment towards a citation. It can be defined as positive, negative or neutral/objective. This type of citation classification is closely related to sentiment analysis.

Many algorithmic approaches can be defined for these tasks; machine-learning systems have proven effective in many text-classification tasks, particularly in the sentiment analysis area (Zhang, Yu and Meng 2007). A comparative study was performed by Sebastiani (2002) that concluded the best classifier applied to sentiment analysis was the super vector machine (SVM). Subsequent studies have confirmed these findings (Mullen and Collier 2004; Pang and Lee 2004; Wilson, Hoffmann, Somasundaran, Kessler, Wiebe and Choi 2005; Boldrini, Fernández Martínez, Gómez Soriano and Martínez Barco 2009; Prabowo and Thelwall 2009; Fernández, Boldrini, Gómez and Martínez-Barco 2011). These studies have been performed using general texts or have been extracted from Web 2.0, such as reviews, user comments, blogs and microblogs; their extrapolation to the domain of citations in scientific literature is still debated.

Polarity classification could be useful for improving bibliometric measures to evaluate the actual citation impact of a cited work in a citing paper. Citation polarity could be an additional criterion to improve measurements of a paper's impact by considering the differences between a citation that is positive from a citation that is criticised or refuted, which should have less weight or negative weight in the calculation of impact indexes (Abu-Jbara et al. 2013).

To classify polarity citations, different feature sets can be used with varying degrees of effectiveness. As we mentioned, in general, results from different studies are not comparable because of the diversity of corpora, but we may begin to assess the effects of different features in citation polarity classification using authors' results. Hedging should also be considered because it is used to disguise an author's negative disposition towards a cited work, but it is not yet used for this task.

Jochim and Schütze (2012) collected and classified different types of features used in previous work and evaluated their effects on the classification of different facets. For polarity classification, we are particularly interested in the confirmational/negational facet of the Moravcsik and Murugesan (MM) scheme because it may be mapped to positive and negative polarity. The authors detected that for this facet, a Bag-of-Word (BoW) classified as positive or negative produced better results than other features for identifying polarity, along with word-level linguistic features, such as modal auxiliary 'has' and main verbs, first- and third-person pronouns, comparatives, superlatives, and contrasting conjunctions. They also concluded that feature classes, such as lexical and frequency, are detrimental for classification of this facet.

In similar approaches, Abu-Jbara et al. (2013) conducted a BoW analysis using lists of negation, subjectivity, speculation and other similar cues to detect polarity. Athar (2011) used word-level linguistic features with POS tags to achieve polarity classification. Teufel et al. (2006) used verb tense, voice and so-called modal auxiliaries, such as 'may' and 'could', the overall location of a sentence and POS tagging. For machine learning, they used the K-nearest neighbours classifier, i.e., the IBk algorithm with k=3 in the WEKA tool kit, to replicate human annotation of a corpus classification of weak, positive and neutral categories (the macro F was 0.71).

Teufel et al. (2009), Dong and Schäfer (2011) and Jochim and Schütze (2012) showed a connection between citation functions and polarity in its location in a sentence, paragraph or section in the paper. Teufel et al. (2006) emphasised the relationship among sections of the paper where a citation was located near its function and polarity.

For polarity citation classification, Athar (2014) used a corpus consisting of 852 papers that cite the top 20 target papers. Athar obtained the best results for the SVM classifier (macro F = 0.808, micro F = 0.764) over the Naïve Bayes classifier (macro F = 0.471 micro F = 0.755) using 1-3 grams, dependency features, and window-based negation. Dependency structures are triplets, or relations (head, dependent) that capture the long-distance relationships between words. Window-based negation considers that all

words inside a k-word window ($k= 1$ to 15) of any negation term are suffixed to distinguish them from non-polar versions. Categories for different schemes for polarity classification and citation function are shown in Table 1.

Athar (2014) reported fewer citation occurrences with negative polarity. Some of these negative mentions cannot be detected because authors tend to hide negative disposition to cited papers through the use of hedging. Effective hedging detection is a necessary task, for which is needed to consider linguistic clues as detailed by Hyland (1996). Techniques should be used to address more abundant neutral citations that can skew classification results.

2.4 Available corpora

The most widely used corpora are subsets obtained from the ACL Anthology, which is a digital archive of research papers in computational linguistics. These subsets are ACL ARC (Radev, et al., 2009) and ACL AAN (Bird et al., 2008); they serve as basis for a number of experiments that use parts of these large bodies of data. Other corpora are annotated subsets from the Linguistic Data Consortium (LDC)¹³; this version adds to (ACL ARC), with several page images in both text and image forms. PubMed and the Lecture Notes in Computer Science (LNCS)¹⁴ series published by Springer are also used as basic corpora.

The datasets acquired from subsets of these large datasets were manually labelled according to the different approaches of the various tasks. ACL AAN was used to select papers for annotation and study by Teufel and Moens (1999); Athar and Teufel (2012); Abu-Jbara et al. (2013); and Tsai, Kundu and Roth (2013). (ACL ARC) was used as a basic dataset by Sugiyama et al. (2010), Athar (2011) and Abu-Jbara and Radev (2012). Some articles were chosen from PubMed by Li et al. (2013) and Meyers (2013). Additionally, some authors work with collections obtained from LNCS (Kaplan, Iida and Tokonuga 2009; Angrosh et al. 2013). The corpora mentioned do not seem to be available to the research community. Human annotation is often evaluated using an inter-annotator agreement kappa based on the difference between observed and expected agreements.

13 <https://www.ldc.upenn.edu/>

14 <http://www.springer.com/computer/lncs?SGWID=0-164-0-0-0>

We found only three public datasets containing citations with annotations of functions and polarity. The first dataset was DFKI citation corpus¹⁵ (Dong and Schäfer 2011), which takes subsets of the articles of the ACL Proceedings of 2007 and 2008. The second dataset is IMS Citation Corpus¹⁶ (Jochim and Schütze 2012), which uses articles from the ACL Proceedings of 2004. Both datasets are taken from the (ACL ARC) corpus. The DFKI Citation Corpus includes polarity annotations. The authors used the Moravcsik and Murugesan (MM) scheme, where positive entries correspond to confirmative and negative entries are negational citations. Jochim (2014) compiled these two corpora into CITD, which integrated unlabeled data from other years of ACL Proceedings, i.e., 1979-2003, 2005-2006 and 2009, with the Multi-Domain Sentiment Dataset MDSD Corpus¹⁷ (Blitzer, Dredze and Pereira 2007). This corpus compiles product reviews from Amazon.com for classifying quotes by applying the marginalized, stacked, denoising auto-encoders mSDA algorithm introduced by Chen, Xu, Weinberger and Sha (2012) with a linear SVM. In Table 2 we show the three previously mentioned annotated corpora for citation polarity classification.

Table 2: Available annotated corpora for citation polarity classification

Corpus	Number of citations	Positive	Negative	Neutral / Objective
DFKI	1,768	190	57	1,521
IMS	2,008	1,836	172	-
CCC	8,736	829	280	7,627

Manual annotation of a citation context analysis is a complex and time consuming task. For this reason annotated available corpora are small and not publicly available. The manual annotation process has problems as low inter-annotation agreement, due to the cognitive complexity of this job (Ciancarini, Di Iorio, Nuzzolese, Peroni and Vitali, 2014). A good agreement is a requirement for reliability and reproducibility of the annotation schemes (Arstein and Poesio 2008). For a manually annotated corpus, Teufel et al. (2006) presented a good inter-annotator agreement of 0.79.

Automatic corpus annotation for context citation analysis has not good outcomes. Some semi-supervised techniques have been applied to automatic labelling of corpus data. Liakata et al. (2012) obtained an automatically annotated corpus with citation context

15 https://aclbib.opendfki.de/repos/trunk/citation_classification_dataset/

16 <http://www.ims.uni-stuttgart.de/~jochimcs/citation-classification/>

17 <http://www.cs.jhu.edu/~mdredze/datasets/sentiment/>

based on discourse structure with accuracy around 0.50 when classifying the 11 CoreSCs. They used a CRF algorithm. Additional classes are more difficult to classify, and these results should be improved before this type of corpus is ready for use.

In this area there are tasks still to be improved such as: the definition of a golden corpus scheme for citation context analysis; manual annotation according to this standard with adequate coded features to yield good precision in automatic classification; and the development of methods for programmed marking of a dataset based on the aforementioned golden corpus scheme.

2.5 Chapter conclusions

The development of citation context-identification applications should be based on argumentative mining or similarity measure techniques; however, so far, the applications are mainly implemented as heuristic approaches. Detection of arguments around a citation, i.e., what is said about a cited work, should define the citation context. Most of the work for automatic context identification uses machine-learning algorithms with supervised approaches; preferred features are selected via Bag-of-Words, n-grams of length 1 to 3 and dependency triplets. Better results in automated context classification are achieved when a fewer number of classes exist. For example, for two classes, context and no context results have accuracy as high as 0.89. For more classes, as in the case of the 11-category CoreSCs, the performance accuracy is only 0.50.

For corpus data generation, manual annotation is difficult and lengthy. Automatic annotation especially for medium and high granularity schemes, using semi-supervised and supervised techniques has produced poor results. As a consequence, there is no corpus available with sufficient size and reliability that be accessible to the scientific community. Thus, researchers work with small datasets that they create themselves, and they use these datasets for manual experiments. Attempts to obtain robust automatic classifiers have not been successful.

The scope problem for a given sentence with multiple citations has been undertaken using a context of only one sentence to define what is said about each reference, but additional work is required to improve results.

Detection of implicit citations has had favourable outcomes, but techniques need to be tested in larger corpora.

Citation function classification aims to extract and categorise citation functions from scientific literature. One possible application is to detect if a citation is somewhat influential in a citing article. This type of information would be useful for improving citation indexes because influential references should have more weight than general or perfunctory citations. Similarly, if citing articles are built upon material from a cited paper, then this reference should have more weight than criticised references.

Of course, a classification scheme must first be defined to classify functions. Different schemes have been proposed with different granularity (35 to 3 classes). There is no standard defined for citation-classification function schemes. Citation polarity classification is a coarse-level sentiment analysis that could be seen as a citation function classification with 2 or three classes: positive, negative and/or neutral. Polarity classification also has bibliometric applications because criticised citations should have less weight in the evaluation of the impact indexes.

Another important issue is hedging as a means to disguise an author's motivation for making a reference. Experimental work should develop recognition of hedging to enhance detection and classification of citation functions and polarity.

Feature engineering is an essential problem for function and polarity detection. It seems that feature selection has more impact than algorithm selection in citation classification. Depending on classification scheme, different features become more appropriate and efficient. For citation function classification, features will depend on the categories defined. For polarity classification, it appears that the use of the location of the reference, semantic analysis at the word level (positive and negative implications) and word-level linguistic characteristics are the most efficient features.

Therefore, we elaborate that there are issues and interesting topics for research in these fields, such as:

- Detecting context windows of different lengths. We proposed that the window length should be related to argument mining such that the context includes most of what is said about a reference or that similarity measures are applied to detect relevant context around a citation.
- Identifying all references in the cited work, including those that are non-explicit, using natural-language-processing techniques and discourse analysis.

- Developing and applying domain-independent techniques and feature engineering for citation classification. Definition of annotated features is vital for achieving good classification results on Precision, Recall and F-measure.
- Detecting hedging to distinguish disguised negativity that has an impact in function and polarity classification.
- Developing a unique framework to execute experimental comparisons among different techniques. This framework should cover available annotated corpora and standard citation categories.

We proposed a survey of context identification, function classification, polarity analysis of citations; and, related research areas, such as feature engineering, schemes for categorisation of citation functions and polarity, manual and automatic creation of corpora, application of experiments to define citation contexts, and argumentative mining techniques for defining contexts.

As a result of this survey we concluded that the major problem facing researchers is that there is not an available corpus with sufficient size and standard and informative annotation schemes that can be used in a shared form to facilitate results that are comparable.

Under these conditions, it is difficult to evaluate the efficiency of different approaches and techniques. Obviously, it is necessary to establish conditions that allow and motivate collaborative work to achieve tangible goals in these fields. The results could be translated into the development of applications that facilitate studies of the dynamics of scientific literature and that promote integral bibliometric measures.

Once we have identified the most important problems in this field, in next chapters, we focus on the study of a classification scheme, the definition of an annotation method and the generation of a corpus for citation context analysis. The generated corpus will also be used to validate the scheme and for classification task of function and polarity. We also propose a function for assessing impact and influence.

3. Esquema de clasificación

Como se ha explicado en los capítulos anteriores, existe la necesidad de generar un corpus que sirva como base para el desarrollo de la investigación en el análisis de citas.

Este grupo de datos tendría como finalidades el análisis del contexto de citas y podría tener aplicaciones como la evaluación del impacto de las citas en un texto tomando en cuenta el propósito y la disposición con los que fueron realizadas, proporcionaría información que podría ser usada para obtener nuevos factores que podrán ser incorporados en el cálculo de un índice de impacto más completo, preciso y justo, que permita evaluar de mejor forma el impacto de los artículos citados y que evite incentivos perniciosos para la publicación en revistas científicas.

Este corpus deberá estar públicamente disponible para que permita: el avance en el estado de la cuestión por parte de todos los investigadores interesados; la evaluación objetiva e independiente de los sistemas de distintos investigadores; el fomento del trabajo colaborativo; el uso de un corpus gold-standard, de calidad, con el objetivo de evaluar y comparar las distintas técnicas utilizadas, con el tamaño suficiente para que las evaluaciones puedan dar resultados estadísticamente significativos.

Dicho corpus necesitará un esquema con una granularidad adecuada, que pueda ser manejada en forma sencilla por los anotadores, que proporcione suficiente información y dé flexibilidad a los sistemas sin que resulte demasiado engorroso a la hora de construirse o ampliarse (Hernández y Gómez, 2014). Además, también deberá anotar características útiles para la clasificación y detección automática de la función y polaridad de cada una de las citas. Esta información anotada ayudará a la realización y mejora de distintas partes de

los sistemas de obtención de nuevos factores de impacto sin necesidad de tener todo el sistema completo.

La construcción de un corpus de estas características, parte de un esquema de clasificación de citas apropiado para responder las preguntas de investigación que se realizan en el campo de análisis del contexto de citas bibliográficas, interrogantes que se relacionan con el propósito con que fueron realizadas (función); la disposición con la que el autor las cita (polaridad); y el impacto que tienen en el artículo que las menciona, es decir si son referencias importantes para el desarrollo del estudio, o si, por el contrario, simplemente se mencionan o pudieran ser fácilmente reemplazadas o incluso eliminadas sin que afectase a la calidad del trabajo.

Uno de los enfoques del presente trabajo es la construcción de este corpus. Una tarea que, como veremos, no ha sido nada fácil y en la cual algunos científicos han fracasado estrepitosamente dada su extrema dificultad. Para ello, no sólo nos hemos visto en la necesidad de construir dicho corpus sino que, además, se ha tenido que diseñar novedosos mecanismos y protocolos de anotación que propendan al logro de un mayor acuerdo entre anotadores.

Aunque no ha sido el objetivo del presente trabajo, pensamos que estos nuevos protocolos podrán servir también a la anotación de otros corpus muy distintos, en donde haya una baja tasa de acuerdo entre anotadores y que, por otros métodos más tradicionales, no puedan aumentarse. No obstante se requerirán nuevas investigaciones que confirmen o desmientan esta hipótesis.

3.1 Criterios para clasificación

El esquema de anotación de Concit se ha diseñado para categorizar citas bibliográficas presentes en un artículo científico, con el objetivo de reconocer su función y polaridad, así como también de valorar su impacto y relevancia. Es crucial, por tanto, definir una serie de categorías por cada uno de estos criterios a evaluar dentro de un esquema de clasificación consistente y útil.

Para determinar este esquema de clasificación es necesario considerar que hay, al menos, dos enfoques a la categorización de citas. El primero define que un esquema es útil en la medida en que es exhaustivo en su profundidad y granularidad. Un ejemplo de este enfoque es el esquema de 35 categorías de Garzone (1996) o el de la ontología CiTO con 41

caracterizaciones para la función de las citas. Estos diseños de alta granularidad tienen la ventaja de que proporcionan bastante más información para el análisis de función y a partir de allí de polaridad, pero son de muy difícil anotación por medios humanos o peor aún automáticos. Por esta razón, con estos esquemas no se han podido construir corpus reproducibles, confiables de tamaño suficiente para que den resultados estadísticamente significativos.

El segundo enfoque es el tomado por Teufel y Moens (1999) que cuestiona esta categorización de grano fino, pues asevera que la mayoría de instancias son difíciles de detectar porque no se encuentran pistas lingüísticas evidentes que puedan servir para clasificarlas, y aunque existan estas claves explícitas, detectarlas es un problema formidable. Teufel y Moens (1999) también expresan que juzgar la naturaleza de la cita conlleva un alto nivel de subjetividad y que hay una ausencia de medios para mapear esa naturaleza a los propósitos de la cita.

A lo anotado por Teufel y Moens (1999), se puede agregar un razonamiento adicional que se contrapone a la definición de un esquema de alta granularidad, y es que un esquema complejo con muchas categorías es muy difícil de anotar, por lo que los anotadores no utilizan todas las posibilidades del esquema y solamente aplican un subconjunto del mismo; además se alcanza un acuerdo entre anotadores muy bajo (Ciancarini, 2014), por lo que no se asegura la confiabilidad y reproducibilidad de esos esquemas. Sin embargo, en este punto cabe anotar que tampoco los esquemas de menor granularidad han logrado valores altos de acuerdo entre anotadores. Teufel et al. (2009) reporta un acuerdo que va desde 0,268 en algunas funciones a 0,79; con un promedio de 0,72, para una anotación de 548 artículos entre tres anotadores y un esquema compuesto por 12 funciones.

El reto es, entonces, diseñar un esquema que sea simple, pero que tenga la información que se requiere. Nuestro esquema consigue ambos objetivos clasificando en forma separada función y polaridad y agrupando funciones similares bajo la misma denominación, con la posibilidad de que estas funciones puedan ser desagregadas considerando la combinación con la polaridad. Además, usando los patrones definidos en el Capítulo 5, se proporcionan datos adicionales que agregan granularidad al esquema. De esta manera con solamente cinco funciones y 3 polaridades, más los patrones que definen al objeto como datos, herramienta, método, concepto, etc., se consigue individualizar con detalle las citas que se analizan de tal manera que nuestro esquema es comparable en granularidad a la Ontología CiTO como veremos más adelante. Sin embargo, en la mente de

los anotadores, la división de función, polaridad y patrones estructura un esquema simple, más fácil de comprender y manejar.

Nuestro esquema de anotación está especialmente diseñado para descubrir características que permitan definir el impacto de una cita mediante su función y polaridad y evaluar su impacto dentro del artículo a partir de esos criterios y de otros datos relevantes, como el sitio en el que se menciona la cita o la frecuencia en la que se la referencia, datos que también están codificados en el corpus. Estos criterios se deberán extraer teniendo en cuenta un contexto variable, calificado por su relevancia, que se define cuándo se realiza la anotación.

Para realizar una correcta anotación del corpus es necesario abordar los criterios descritos anteriormente, conjuntamente con los patrones que se describen en el Capítulo 5. De esta forma se puede tener un alto nivel de detalle para definir el propósito de las citas y deducir a partir de allí la función, la polaridad y simultáneamente el tipo de información a la que se enlaza con una granularidad muy fina.

Por ejemplo: las funciones podrían relacionarse con etiquetas como “CONCEPT”, “TOOL” o “DATA”, entre otras posibilidades. Al tener en cuenta estas relaciones se puede profundizar aún más en la categorización de las citas.

Esto se verá con más detalle en el Capítulo 5, sin embargo para ilustrar la granularidad que se logra, consideremos una cita con función *Useful* que tiene asociadas etiquetas que representan datos, conceptos, herramientas u otras; y además tienen palabras clave que pueden indicar una disposición positiva como por ejemplo la frase: “excelente rendimiento”, que será anotada dentro del corpus, o no aparece ninguna palabra clave de ese tipo lo que podría llevarnos a decir que tiene una polaridad neutral.

Entonces, podremos hablar de una cita “Useful” que proporciona datos, conceptos o herramientas, que ha sido acreditada en forma positiva o neutral. De este modo, como se puede apreciar en la Figura 1, con una sola función, en este caso *Useful*, podemos caracterizar citas con al menos seis combinaciones: *Useful Data Positive*, *Useful Concept Positive*, *Useful Tool Positive*, *Useful Data Neutral*, *Useful Concept Neutral* y *Useful Tool Neutral*.

En resumen, el esquema permite un nivel inicial de granularidad media que si bien es útil por sí mismo, puede ser completado a una descripción de alta granularidad, al procesar y enlazar las funciones con la polaridad y las etiquetas que se anotan. Y lo más importante es que esa alta granularidad se logra manteniendo la simplicidad; es mucho

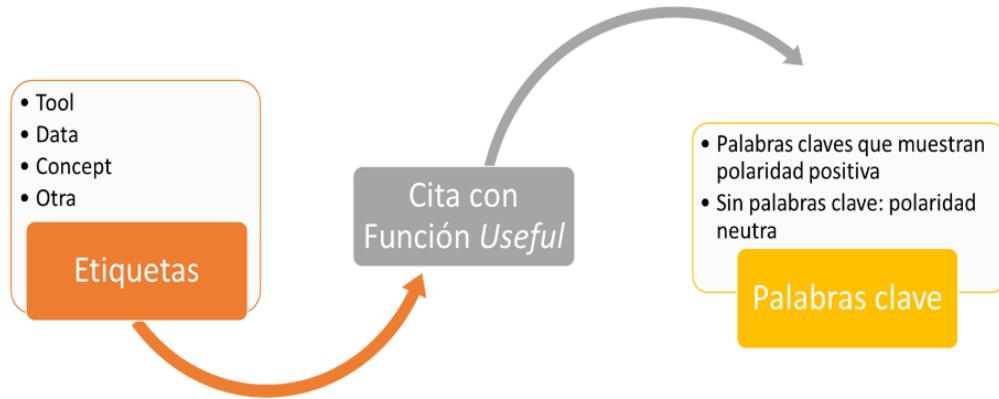


Figura 1: Granularidad con una función

más fácil para los anotadores recordar y aplicar en forma separada, cinco funciones y tres polaridades y reconocer patrones, que tener en mente 41 o 35 funciones independientes.

En la Figura 2 están representados los 3 criterios de impacto que hemos considerados relevantes para la evaluación de la relevancia de cada una de citas de un artículo: *función*, *polaridad* e *impacto*. A su vez, cada uno de estos criterios se divide en distintas categorías.

El criterio de *función* se refiere al papel que realiza la cita en el artículo que la



Figura 2: Criterios de clasificación de una cita de acuerdo a la función polaridad e impacto de sus referencias

menciona. El total de funciones son 6 y su descripción se resume en la Tabla 3.

Por sus características particulares es necesario realizar una mención especial a la función *Hedges* (Hyland 1996). *Hedges* es una forma de lenguaje cauteloso para ocultar una disposición negativa respecto a la cita, con el fin de evitar reacciones no deseadas por parte de los autores citados. Este recurso es muy común en literatura científica y utiliza distintas técnicas lingüísticas para suavizar posiciones acerca de las ideas de otros, como por ejemplo, empezar halagando un trabajo para que la subsiguiente crítica no se la tome duramente, o comparar el trabajo propio con el de otro y elogiar al propio como una forma de expresar desacuerdo. Detectar la presencia de *Hedges* permitirá descubrir referencias que implican posiciones y polaridades negativas encubiertas. Uno de los tipos de *Hedges* que detectamos y anotaremos, tiene la particularidad de que el análisis de la cita empieza con una característica positiva, seguida luego de una negativa, por ejemplo:

"The only recent work on citation sentiment detection using a relatively large corpus is by Athar (2011). However, this work does not handle citation context."

La característica positiva se refiere a que el trabajo citado es el único reciente que usa un corpus relativamente grande. La característica negativa es que el trabajo no maneja

Tabla 3: Esquema de anotación para funciones

Función de la referencia	Descripción
Based on, Supply	El artículo que referencia aplica el trabajo de la cita. El artículo que la referencia se construye a partir del trabajo de la cita (Based on) o el trabajo de la cita es usado como una fuente (Supply).
Useful	El material de la cita (concepto o herramienta) se reconoce como útil y se aplica en algún otro trabajo, no en el propio.
Acknowledge, Corroboration, Debate	La cita se menciona para reconocer algún trabajo previo. El artículo que la referencia puede: simplemente mencionar la cita (Acknowledge); estar de acuerdo con ella (Corroboration); o, discutir, disputar el trabajo de la cita (Debate).
Contrast	La cita se compara con otros trabajos, el resultado es un criterio que puede ser positivo, negativo o neutro.
Weakness, Correct	Se nota un error o debilidad de la cita (Weakness), se corrige un error o debilidad de la cita (Correct).
Hedges	Se usa un lenguaje cuidadoso para ocultar la crítica (Hedges).

contexto de citas. La característica positiva se expresa con las palabras clave “*using a relatively large*” y la negativa “*does not handle*”. En este caso se suaviza la crítica con el reconocimiento de algo positivo para la cita.

Consideremos otro ejemplo de *Hedge* en el que no se menciona la crítica exponiendo categóricamente un error, sino que se la disimula sin hacer una afirmación categórica, suavizándola con las palabras “*is rarely true*”:

“The first experiments in Argumentative Zoning used Naïve Bayes (NB) classifiers Kupiec et al., 1995; Teufel, (1999) which assume conditional independence of the features. However, this assumption is rarely true for the kinds of rich feature representations we want to use for most NLP tasks.”

La función *Hedges* es especialmente difícil de detectar y diferenciar de otras funciones incluso si estas funciones tienen una polaridad opuesta. Como apunta Hyland (1996), esta categoría, por su negatividad velada es difícil detectar a menos que se descubran las formas descritas en los párrafos anteriores.

Todas las agrupaciones del criterio de *función* se han realizado buscando el mejor acuerdo entre anotadores sin olvidar una de las aplicaciones básicas de este esquema, que es la de evaluar la función y relevancia de las citas. Además, hemos unido en la misma categoría a funciones que pueden ser separadas teniendo en cuenta el criterio de polaridad, separando sólo aquello imprescindible y que afecte a la relevancia de la cita. Por ejemplo, no hace falta que el anotador diferencie entre *Based on* y *Supply*, si las dos categorías expresan que el trabajo citado fue utilizado en la cita, y diferenciarlo en subcategorías hubiera dificultado el trabajo de los anotadores dada la sutil diferencia entre estas dos funciones, que se aclara con la polaridad pues *Based on* tiene polaridad positiva y *Supply* neutral. Por otro lado, la función *Acknowledge*, *Corroboration* y *Debate* se puede separar, usando su polaridad, en las 3 funciones. Con polaridad positiva se tendría una afirmación de los resultados de otro autor (*Corroboration*). Con polaridad neutral, simplemente, se mencionaría la cita como parte del estado de la cuestión (*Acknowledge*). Con polaridad negativa se mostraría desacuerdo (*Debate*). Se pueden desagrupar esas funciones gracias a la combinación con el criterio de *polaridad*, que sirve para hacer esta distinción.

En la Tabla 4 se han desglosado las tres categorías para el criterio de *polaridad*: *positive*, *negative* o *neutral* que expresan, respectivamente, una disposición favorable, no favorable o neutral al trabajo citado.

Tabla 4: Esquema de anotación para polaridad

Polaridad	Descripción
Positive	El autor tiene una disposición favorable hacia el trabajo citado.
Negative	El autor tiene una disposición no favorable hacia el trabajo citado.
Neutral	El autor no muestra una disposición ni positiva ni negativa hacia el trabajo citado.

Finalmente, en la Tabla 5, se describen las tres categorías del criterio de *impacto*, pues uno de los aspectos importantes es qué impacto ha tenido el trabajo citado en el artículo que lo referencia. Como veremos más adelante, este criterio se puede obtener de forma automática teniendo en cuenta otros factores y categorías de las citas, por lo que no necesitan ser anotados manualmente.

Si tenemos en cuenta todos los criterios, más los patrones que se describen en el Capítulo 5, se tiene una alta granularidad. Ya vimos como con una sola función se pueden tener al menos 6 caracterizaciones de clase de función. Como hemos visto anteriormente, combinando categorías de distintos criterios se afina la granularidad. Por ejemplo, solamente tomando en cuenta la polaridad y no los patrones, el criterio *Acknowledge*, *Corroboration*, *Debate*, cuando tiene polaridad *positiva* se referirá a *Corroboration*, cuando es *neutral* se relaciona con *Acknowledge* y cuando es *negativa* corresponde a *Debate*; y la función *Based on*, *Supply* cuando tiene polaridad *positiva* es *Based on*, y cuando es *neutral* es *Supply*.

Con este esquema se puede pasar de una granularidad media de la función de 6

Tabla 5: Esquema de anotación para impacto

Valor	Característica
Negative	No hay relación entre el artículo y la cita, en este caso el autor menciona la referencia con una disposición negativa o crítica.
Perfunctory	Citas triviales, relacionadas solo marginalmente con el artículo que la referencia. Polaridad neutral de la cita.
Significant	Citas importantes para el trabajo que las menciona, que están estrechamente relacionadas con el trabajo que hace la mención. Generalmente están vinculadas a una polaridad positiva hacia la cita.

categorías a una granularidad mucho más fina.

La función *Weakness*, *Correct* siempre tiene una polaridad negativa, lo mismo que *Hedges*. *Useful* no tiene polaridad negativa. *Based on*, *Supply* no tiene polaridad negativa.

Un resumen de las combinaciones entre ambos criterios se muestra en la Tabla 6.

Como ya mencionamos y veremos en detalle en el Capítulo 5, esta no es la única forma de aumentar la granularidad de nuestro esquema, sino que también se usarán otras características como palabras claves o construcciones sintácticas anotadas en el contexto de la cita se incrementará más la granularidad.

Tabla 6: Funciones desagrupadas por combinación función-polaridad

Función agrupada	Polaridad	Función desagrupada
Based on, supply	Positive	Based on
	Negative	N/A
	Neutral	Supply
Useful	Positive	Useful
	Neutral	Useful
Acknowledge, Corroboration, Debate	Positive	Corroboration
	Negative	Debate
	Neutral	Acknowledge
Contrast	Positive	Contrast
	Negative	Contrast
	Neutral	Contrast
Weakness, Correct	Positive	N/A
	Negative	Negative
	Neutral	N/A
Hedges	Positive	N/A
	Negative	Negative
	Neutral	N/A

Como hemos visto, nuestro esquema permite un alto nivel de granularidad por la simple combinación de dos características que se anotan en forma separadas: *función* y *polaridad*. De esta manera presentamos un diseño simple pero poderoso para especificar la tipología de las citas.

Los criterios que proponemos son completos en la medida en que permiten clasificar todas las instancias de las citas. Hemos demostrado durante la evaluación y valoración del modelo de anotación (Capítulo 5) que, con este esquema de clasificación, los anotadores tienen facilidad para encontrar una clara correspondencia entre las referencias encontradas con una categoría dentro del esquema propuesto.

3.2 Comparación de nuestro esquema con la ontología CiTO

La taxonomía de citas utilizada en el presente trabajo se puede correlacionar con otras taxonomías conocidas, como, por ejemplo, la ontología de citas definida en CiTO, con la cual comparamos nuestro esquema porque esa ontología proporciona una granularidad muy fina. Cada una de las 41 propiedades del objeto cite, pueden mapearse a los criterios de nuestro esquema que, sin embargo, resulta mucho más sencillo de etiquetar porque solamente tiene 6 funciones agrupadas y 3 valores de polaridad.

Ciancarini, et al., 2014, se refiere a la ontología CiTO expresando que caracteriza en forma completa la naturaleza de las citas y permite definir en forma precisa su clasificación, pero que debido al número de propiedades de las citas la mayoría de usuarios solamente usan algunas de estas clases, y que por las sutiles diferencias entre ellas se tiene un muy bajo un acuerdo entre anotadores. La combinación de los criterios de nuestro esquema puede servir para definir a cada una de las propiedades de esa ontología, lo que demuestra la granularidad fina de nuestro esquema, que además es mucho más sencillo, lo que facilita su anotación.

El objeto *cito:cites* de la mencionada ontología, define “*The citing entity cites the cited entity, either directly and explicitly (as in the reference list of a journal article), indirectly (e.g. by citing a more recent paper by the same group on the same topic)...*”. La relación entre la ontología descrita en CiTO para el objeto *cito:cites* y sus propiedades se muestra

en la Tabla 7, se usan los ejemplos descritos por los autores¹⁸ para demostrar la aplicación del esquema, donde [X] se refiere a una cita. La función desagregada corresponde a la clasificación de grano más fino que corresponde a una combinación de función y polaridad a la que adicionalmente se le puede agregar la información de las etiquetas que se explican en el Capítulo 5; de modo que, por ejemplo, se puede diferenciar una función *Based on*, *Supply* pues cuando tiene polaridad *positiva* realmente hace referencia a la función desagregada *Based on* y se puede añadir la información de una etiqueta semántica como <method>, <tool> o similares. Como se puede observar, con esos datos se especifica muy claramente la relación de la cita con el artículo que la referencia. Por esa razón, enfatizamos que nuestras 5 funciones agrupadas, junto con los tres niveles de polaridad, más las etiquetas y palabras clave descritas en el Capítulo 5 sirven para representar, al menos, las 41 propiedades de la ontología CiTO.

Tabla 7: Correlación entre las 41 propiedades de la ontología CiTO y nuestro esquema de clasificación

Propiedad objeto cito:cites	Ejemplo	Clasificación usando esquema propuesto			
		Función agregada	Pol.	Función desagreg.	Palabras clave o etiquetas
cites as authority	Newton asserted that we are like dwarfs standing on the shoulders of giants [X].	Acknowledge, Corroboration, Debate	Neu	Acknowledge “asserted”	
cites as data source	Italy has more than ten thousand kilometers of shoreline: see [X].	Based on, Supply	Neu	Supply	<data>
cites as evidence	We found an unquestionable demonstration of our hypothesis in [X].	Acknowledge, Corroboration, Debate	Pos	Corroboration	“unquestionable”
cites as metadata document	Basic bibliographic, entity and project metadata relating to this article, recorded in a structured machine-readable form, is available as an additional file [X] accompanying this paper.	Based on, Supply	Pos	Based on	“Relating to this article” <data>
cites as potential	This risk could be avoided using the approach shown	Useful (si no se usa en el	Pos	Useful	“risk could be avoided”

¹⁸ <http://purl.org/spar/cito>

solution	in [X].	artículo)				
		Based on, Supply ¹⁹ (si se usa en el artículo)	Pos	Based on	“using”	
cites as recommended reading	To our knowledge, [X] is the best source of exercises about UML, making it a valuable proposal for beginners.	Useful	Pos	Useful	“the best source”	
cites as related	An analysis similar to what we proposed here is presented in [X].	Contrast	Pos	Contrast	“similar”	
cites as source document	Several sections of this work are based on our literature review of the topic published as journal article [X].	Based on, Supply	Pos	Based on	<contribution> “are based on”	
cites for inf.	The grammar of Pascal was introduced in [X].	Useful ¹⁹ (Si no se usa en el artículo) Based on, Supply ¹⁹ (Si se usa en el artículo)	Neu	Useful	<tool>	
compiles	This book gathers interviews with academic researchers of several disciplines [X].	Acknowledge, Corroboration, Debate	Neu	Acknowledge	“gathers”	
confirms	Our findings are similar to those published in [X].	Contrast	Pos	Contrast	“similar”	
contains assertion from	We think that “to stand on the top of giants” [X] is a valuable principle to follow for our own research.	Based on, Supply	Pos	Based on	<quote>	
corrects	The result published in [X] is partially wrong, the correct result is 42.	Weakness, Correct	Neg	Correct ²⁰	“the correct result is” “wrong”	
credits	<i>Galileo was the first to observe Jupiter's satellites [X].</i>	Acknowledge, Corroboration, Debate	Pos	Corroboration	“was the first to observe”	
critiques	<i>The ideas presented in [X]</i>	Weakness,	Neg	Weakness	“badly”	

19 Según el contexto

20 En este caso la desagregación depende de las palabras clave y la estructura de la oración definidas en el contexto

are badly substantiated. Correct						
derides	The ideas published in [X] are incredibly stupid.	Weakness, Correct	Neg	Weakness	"incredible stupid"	
describes	Galileo's book [X] is a dialog among three scientists about Copernicus' eliocentric theory.	Acknowledge, Corroboration, Debate	Neu	Acknowledge	<contribution> "is"	
disagrees with	We do not share Galileo's opinion [X]: the Earth does not move.	Acknowledge, Corroboration, Debate	Neg	Debate	<author> "do not share"	
discusses	We now examine if Galileo is right when he writes [X] that the Earth moves.	Acknowledge, Corroboration, Debate	Neg	Debate	"examine if"	
disputes	We doubt that Galileo is right when he writes [X] that the Earth moves.	Acknowledge, Corroboration, Debate	Neg	Debate	<author> "doubt"	
documents	Herein we report in detail the complete set of ontological rules defined in the Overlapping Ontology [X].	Useful ¹⁹ (Si no se usa en el artículo) Based on, Supply ¹⁹ (Si se usa en el artículo)	Pos	Useful	"report" "the complete set"	
extends	We add to Galileo's findings concerning the Earth [X] that also the Moon moves.	Based on, Supply	Pos	Based on	"add"	
includes excerpt from	Oxford 01865 Oxshott 01372 Oxted 01883 Oxton 01578 is an excerpt from the UK Dialling Codes section of the Oxford Telephone Directory.	Acknowledge, Corroboration, Neu Debate	Acknowledge	Acknowledge	"excerpt from"	
includes quotation from	On June 4th 1940, Winston Churchill made a speech on the radio that has since become famous, that included the words: "... we shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills; we shall never surrender ..."	Acknowledge, Corroboration	Neu	Acknowledge	<quote>	
obtains	There is a need for more observational studies and	Based on,	Neu	Based on	"There is a need" <author>	

background from	studies using narrative causation to describe the potential contribution of information in problem-solving and decision-making [X]; our work addresses these needs.	Supply		"addresses these needs!"
obtains support from	Our ideas were also shared by Doe et al. [X].	Acknowledge, Corroboration	Pos	Corroboration <author> "shared by"
parodies	We act as giants on the shoulders of dwarfs [X]!	Hedges	Neg	Hedges "!"
plagiarizes	The conclusion of our dissertation can be summarised by the following motto, we created specifically for this purpose: we are like dwarfs standing on the shoulders of giants.	NA ²¹	NA ²¹	NA ²¹
qualifies	Galileo's masterpiece 'Dialogo sopra i due massimi sistemi del mondo' [X] is formally a dialog and substantially a scientific pamphlet.	Acknowledge, Corroboration, Debate	Neg	Debate "a scientific pamphlet"
refutes	We do not think that all their arguments in favour of their own and against the other strategies are equally convincing [X].	Hedges	Neg	<author> "think"
replies to	We will not investigate the issues of the approach proposed in [X] here, but rather we introduce yet another alternative.	Weakness, Correct	Neg	Correct 14 "introduce yet another alternative"
retracts	We wrote that the Earth moves in [X]; we now retire such statement.	Weakness, Correct	Neg	Correct <author> "retire such statement"
reviews	This paper discusses Toulmin's methodology in modelling argumentation [X], focussing on highlighting advantages and drawbacks of the application of such a methodology in the Social Web.	Acknowledge, Corroboration, Neu Debate	Acknowledge	<author> "discusses" <methodology>

21 No aplica porque por definición lo plagiado no contiene una cita.

ridicules	Galileo said that the Earth "moves" [X]; really? And where is it going?	Hedges	Neg	Hedges	"really?"
speculates on	We believe that if Galileo believed that Earth goes around the Sun [X], he also should believe that Moon goes around Earth.	Acknowledge, Corroboration, Debate	Neu	Acknowledge	"also should believe"
supports	We support Galileo's statement [X], that Earth moves.	Acknowledge, Corroboration, Debate	Pos	Corroboration	"support"
updates	Earth moves, said Galileo [X]; in addition, we can say now it moves very fast.	Acknowledge, Corroboration, Debate	Pos	Corroboration	"in addition"
uses conclusions from	Building upon Galileo's findings [X], we discovered that all the planets move.	Based on, Supply	Pos	Based on	"building upon"
uses data from	Using the information collected from our recent study [X], we can estimate that there are tens of millions of HTML forms with potentially useful deep-web content.	Based on, Supply	Pos	Based on ²²	<data>
uses method in	We follow [X] in using design patterns for testing.	Based on, Supply	Pos	Based on ²³	<method>

Universidad de Alicante

3.3 Conclusiones del capítulo

Luego de analizar alrededor de 1200 citas y sus respectivos contextos, en este capítulo se definió un esquema para categorización de citas que toma en cuenta su propósito o función, la disposición del autor al mencionar la referencia o polaridad y el impacto que tiene en el documento que las menciona. En el siguiente capítulo se definirán dos métodos para asignar valores de impacto de acuerdo a nuestro esquema para impacto del artículo citado.

22 Asociado a la etiqueta DATA

23 Asociado a la etiqueta METHOD

Se ha presentado el diseño de un esquema simple que combina 6 funciones agrupadas, 3 niveles de polaridad y la asociación con palabras clave y etiquetas detectadas en el contexto relevante, para lograr una granularidad fina que se puede mapear semánticamente a las 41 propiedades de una ontología conocida como es CiTO.

Las funciones agrupadas se desagregan tomando en cuenta la polaridad y también las construcciones sintácticas que se encuentran en el contexto alrededor de la cita, por ejemplo si se refiere a datos, a métodos, a herramientas, etc., esta es información valiosa que está anotada en el corpus y que puede ser utilizada en distintas aplicaciones relacionadas con el análisis de citas bibliográficas. La simplicidad del esquema es importante porque facilita a los anotadores el etiquetado manual del corpus.



4. Evaluación del Impacto de una cita

La lógica del esquema de clasificación visto en el capítulo anterior, define criterios que separan claramente las diferentes clases, que cubren las posibilidades de categorías que se pueden presentar y que, además, se van a relacionar con medidas de impacto. Este esquema guiará el proceso de anotación del corpus para análisis de citas.

En este capítulo proponemos usar la información de este esquema para resolver el problema presentado en la introducción del presente trabajo, en donde se establecía que, debido a que no todas las citas tienen la misma importancia o relevancia para el artículo que las menciona, para definir su impacto no es suficiente contar sus ocurrencias sin hacer una distinción entre ellas pues no todas las citas deben tener el mismo peso, sino que se deben calificar con valores de impacto diferenciados de acuerdo a la relevancia que tengan en el trabajo que las menciona. Se presenta, por tanto, dos algoritmos distintos para el cálculo de la relevancia con el objetivo de realizar una evaluación preliminar. Esta evaluación nos permitirá conocer si a partir de las anotaciones del corpus se puede aproximar una función a los criterios de impacto que los propios autores tienen de las referencias que citan. Con estos datos pretendemos demostrar que con los factores definidos en la sección anterior podremos, en el futuro, utilizarlos para medir la significancia de las citas. El algoritmo que cumpla este requisito no tiene por qué ser óptimo pues no es el objetivo de esta tesis realizar una búsqueda exhaustiva, sino demostrar que, al menos, hay un algoritmo que nos pueda dar unos resultados aceptables.

Los dos métodos presentados usan la información que se etiqueta o se puede obtener en el corpus que se genera en el presente trabajo. La valoración del impacto de una cita en

un artículo es muy importante porque si se pudiera evaluar en forma consistente el impacto de un artículo sobre otro, se generaría un criterio que podría ser el inicio de una nueva visión para juzgar las distintas contribuciones científicas en una manera integral. Esta sería una substancial mejora a las técnicas actuales que se basan en conteos de citas consideradas todas iguales. Para poder evaluar el resultado de ambas aproximaciones, se les envió una encuesta a los primeros autores de los 85 artículos anotados en el corpus para que valorasen, según su criterio, la relevancia de las citas que utilizaron en sus artículos en las categorías de *Negative*, *Perfunctory* y *Significant*. Desgraciadamente sólo contestaron 6 autores a la encuesta. No obstante obtuvimos 161 citas anotadas que presentamos en la Tabla 65 del Anexo 3. Los nombres de las funciones y las polaridades los pusimos abreviados, por ejemplo en lugar de *Acknowledge* se escribe *Ack*, en lugar de *Positive* se usó *Pos*. Cabe destacar que la gran mayoría (123 de las 161) de las citas fueron etiquetadas manualmente como superficiales (*Perfunctory*), 36 fueron marcadas como importantes (*Significant*) y sólo 2 fueron clasificadas como negativas (*Negative*). Este patrón suele ser bastante habitual (como veremos en el Capítulo 7) pues la gran mayoría de citas mencionadas en un artículo suelen ser superficiales y suelen estar en la sección de Introducción (que incluye a la del estado de la cuestión). Sin embargo hay mucha reticencia a criticar el trabajo de otros autores. Las medidas utilizadas para obtener los resultados son *Precision*, *Recall* y *F-Measure* descritos en la Sección 7.2. Para el cálculo del intervalo de confianza se ha utilizado una normal tipificada con un nivel de confianza del 95% y cuya formulación también se puede encontrar en la Sección 7.2. En los siguientes apartados se describen ambos métodos y se evalúa su eficacia. También hemos visto interesante comprobar si con un algoritmo de aprendizaje podemos aproximarnos al algoritmo que nos da mejores resultados.

4.1 Impacto relacionado directamente con la polaridad de las citas

En el primer método, el impacto de una cita se relaciona directamente con la polaridad o disposición del autor hacia la cita.

En general especificamos tres valores de impacto, como se muestra en la Tabla 8. En este punto la evaluación de impacto se realiza directamente tomando en cuenta la polaridad, es decir, se evaluará el impacto con un valor de *Negative* si la polaridad es negativa, *Perfunctory* si la polaridad es neutral y *Significant* si la polaridad es positiva.

Tabla 8: Propuesta para valoración del impacto usando solo la polaridad

Polaridad Impacto	
Negative	Negative
Neutral	Perfunctory
Positive	Significant

Este enfoque se basa en los siguientes razonamientos: i) si las citas son criticadas tienen una polaridad negativa y probablemente no tendrán una cercanía al trabajo del autor que hace la referencia, por esa razón les asignaremos un valor de impacto igual a *Negative*; ii) son triviales y están relacionadas solo marginalmente con el artículo que las referencia (polaridad neutral), para ellas determinamos el valor de *Perfunctory*; y iii) están más relacionadas con el trabajo que hace la mención, por lo que el autor muestra una disposición favorable hacia ellas (polaridad positiva), a este tipo de citas les fijamos el valor de *Significant*.

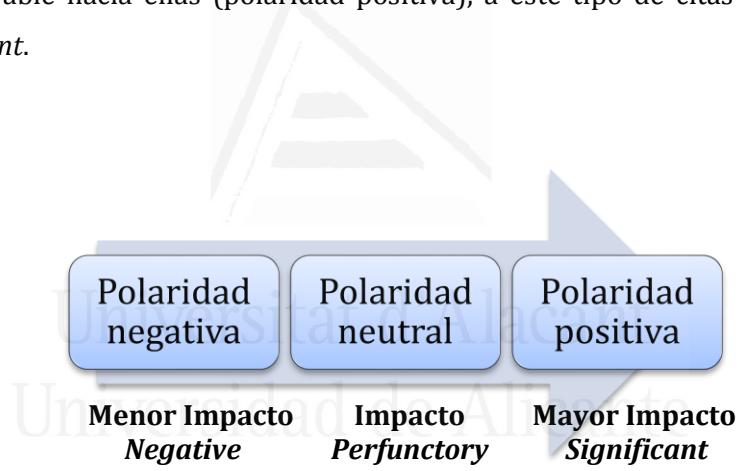


Figura 3: Relación entre impacto y polaridad

Como se muestra en la Figura 3, a la polaridad negativa correspondería asignarle el menor valor de impacto; las citas que se realizan con una disposición positiva tendrían un factor de impacto mayor que si se realizan neutral o negativamente. Para menciones repetidas, el valor de impacto total de la cita en el artículo, corresponde al de mayor frecuencia, o en caso de empate entre distintas polaridades, corresponderá al valor más alto.

Pero, como hemos visto en el capítulo anterior, existen una serie de funciones segregadas que equivalen a cada una de las polaridades. Por lo tanto, en la Figura 4, se presenta el valor de impacto según dicha función desagregada.

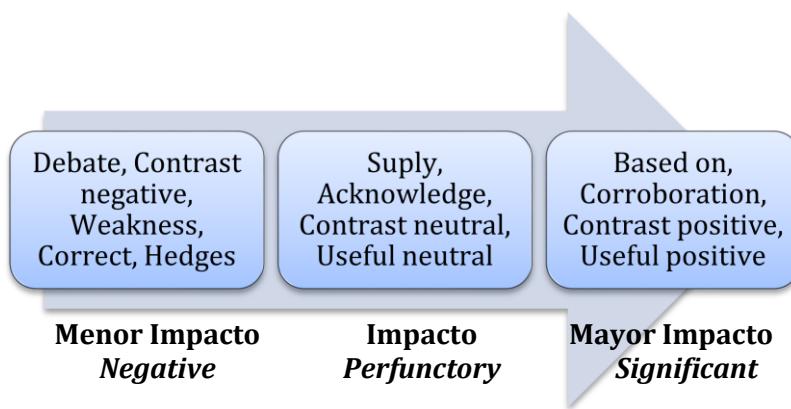


Figura 4: Impacto de acuerdo a la polaridad de las funciones desagregadas

En la Tabla 65 del Anexo 3 se presentan los resultados de una encuesta a los autores de los artículos sobre el impacto de cada cita en su documento. Para esta primera aproximación al cálculo del impacto de citas se utilizará la columna *Polaridad*, pues en ella aparece la polaridad de la cita para todos aquellos casos en los que aparece dicha cita. La columna *Impacto calculado con nuestro método* es el resultado del método utilizado en el apartado 4.2. Los resultados de esta comparación se aprecian en la Tabla 9.

Aunque la media de *Precision*, *Recall* y *F-Measure* con este sencillo método están alrededor de un $78,5\% \pm 6,3$, se puede apreciar que para el impacto *Significant* los resultados están, con un 95% de confianza, entre el 63,6% y el 30,96%, claramente unos resultados insuficientes. Esto se debe a que los autores no tienen por qué mencionar

Tabla 9: Comparación entre el impacto anotado por los autores y su cálculo a partir de la *Polaridad* de la cita.

Impacto	Precision	Recall	F-Measure	Intervalo de confianza
Negative	1,0000	1,0000	1,0000	--
Perfunctory	0,8309	0,9187	0,8726	$\pm 0,0592$
Significant	0,6842	0,3611	0,4727	$\pm 0,1631$
Weighted Avg.	0,8002	0,7950	0,7848	$\pm 0,0625$

expresamente de forma positiva a las citas que consideran relevantes. Sin embargo, con los impactos *Negative* y *Perfunctory* parece que hay una clara correlación entre la polaridad de la cita y su impacto negativo o superficial. La mayoría de las citas neutras

serán superficiales y la totalidad de citas negativa tendrán un impacto negativo. Esto era bastante esperable debido a las definiciones en nuestro modelo de anotación de estos dos impactos. No obstante hay que ir cuidado pues como se ve la Tabla 10 no es posible valorar el intervalo de confianza debido al escaso número de muestras negativas.

En la matriz de confusión se confirman estos hechos pues la mayoría de problemas se presentan entre los impactos *Significant* y *Perfunctory*. Sobre todo al clasificar correctamente las citas importantes para el autor que muchas de ellas son clasificadas como superficiales. Únicamente con el criterio de polaridad no hay suficiente información para saber si la cita es relevante o superficial aunque para las negativas parece que sí aunque en este último caso es necesario más datos para poder concluir con un mínimo de confianza.

Tabla 10: Matriz de confusión para la evaluación del impacto mediante la polaridad

		← clasificado como	
		Negative	Perfunctory
← evaluado como	Negative	2	0
	Perfunctory	4	113
	Significant	0	23
		0	13

4.2 Impacto calculado por ubicación, repetición, función y polaridad de las citas

El segundo enfoque que proponemos para evaluar el impacto, aparte de la polaridad que hemos visto en el apartado anterior, también tiene en cuenta factores como el sitio en el que se encuentra la cita dentro de un artículo científico, su función y las veces en que se la menciona. McCain y Turner, 1989 introdujeron la noción de que el sitio del artículo en el que aparece la cita fija su función e importancia en la publicación; por ello propusieron que cada cita tenga un valor de impacto tomando en cuenta la sección del artículo en la que aparece, usaron algunos criterios relacionados con la naturaleza de las secciones, asignando diferentes pesos de acuerdo a sus funciones retóricas.

Cano, 1989 estudió la relación entre tipo de citas, nivel reportado de utilidad y ubicación de la cita. Prabha, 2007, en un estudio empírico de comportamiento de las citas, llegó a la conclusión de que menos de la tercera parte de las referencias mencionadas en un artículo fueron consideradas esenciales para quienes las hicieron y que la mayor parte de citas se las menciona en la *Introducción* del artículo. Se puede ver, entonces que estudios anteriores ya han definido la importancia de la ubicación de la cita dentro del artículo para determinar si es o no una cita esencial, principio que usaremos dentro de este enfoque para valoración del impacto.

Para definir la ubicación de la cita en el artículo usamos la estructura IMRaD (Biber y Finegan, 1994), esto es *Introduction, Methods y Materials, Results and Discussion*. Esta es una estructura usada frecuentemente en literatura científica (Sollaci, y Pereira 2004), que estandariza la organización del contenido para facilitar la lectura y el manejo de la gran cantidad de información que se encuentra disponible. Cuando los artículos presentan otros títulos en sus secciones, se puede realizar una correspondencia con esta estructura básica, realineando cada parte de acuerdo con el significado de cada sección de IMRaD: la *Introducción* explica el alcance y objetivo del estudio; *Materiales y Métodos* describen cómo se realizó el estudio; los *Resultados* reportan lo que se encontró; y, la *Discusión* explica el significado de los resultados y proporciona lineamientos para futuros trabajos.

La repetición o no de la mención es otro criterio para calificar el impacto. Herlach (1998) prueba que cuando hay más de una mención a una referencia dentro del mismo artículo de investigación, en distintas secciones, esto indica que hay una relación cercana y útil entre el artículo que cita y el citado. En la investigación de Herlach (1998), la cercanía y utilidad de la relación entre artículos fue determinada mediante juicio de usuarios y es un criterio próximo a la medición del impacto.

Además, como vimos en el punto anterior, la polaridad es otro elemento que se puede tener en cuenta para definir el impacto de una cita. En este enfoque este criterio se combina con los otros mencionados.

Una misma cita que tiene varias menciones en un artículo puede ser clasificada de diversas maneras a lo largo del texto, con distintas funciones y polaridad. Para entrenar el algoritmo, catalogamos cada cita en un artículo tomando en cuenta el impacto de cada cita por su función y polaridad; considerando la función y la polaridad de la cita en cada sección, y el número de secciones en la que aparece. Se supone que una cita es más influyente en un artículo si se la menciona varias veces, sobre todo si es en distintas secciones, en especial en las centrales de la publicación, esto es *Métodos, Resultados y*

Discusión, dejando de lado la *Introducción* porque en ella principalmente se establecen los antecedentes del estudio. Teniendo en cuenta los factores señalados se definen valores de impacto de *Negative*, *Perfunctory*, *Superficial*, tal y como se determinó en la Sección 4.1 y de acuerdo a nuestro esquema de clasificación.

Para automatizar la evaluación del impacto de una cita en un artículo, se usa SVM con SMO en WEKA. Para el algoritmo mencionado se tienen como entradas las siguientes características que se encuentran anotadas en el corpus o se pueden obtener de él: número de veces que se menciona la cita en cada sección IMRaD del artículo, las sumas parciales y totales de ocurrencias por sección y en todo el artículo repartidas por función, por polaridad; como entrada adicional se considera el número de secciones en la que aparece la misma cita en el artículo. Esta última característica se puede inferir estudiando el comportamiento de los autores, pues Herlach (1998) establece que una cita es más influyente si se la nombra a lo largo de todo el trabajo y su mención no está circunscrita a una sola sección; en ese caso se puede decir que se tiene a la cita como una referencia en los distintos puntos de desarrollo del trabajo y por lo tanto tiene una alta probabilidad de ser cercana al estudio que la refiere.

Se desarrolló una función algorítmica para valorar el impacto, tomando en cuenta los criterios explicados en párrafos anteriores. El diagrama de flujo se presenta en la Figura 5, en donde se aprecia los pasos y las decisiones que se tienen en cuenta para decidir cuándo una cita es *Negative*, *Perfunctory* o *Significant*. Como se puede observar en dicho diagrama, si una cita se etiqueta como *Base on* al menos una vez se considera como importante para el trabajo que la cita pues se basa en dicha referencia para continuar con su investigación y, por lo tanto, es importante para el autor. Si la cita sólo se menciona una única vez y en la introducción o en el estado de la cuestión, entonces el autor está reconociendo que conoce dicho trabajo pero no es crucial para el avance del suyo y, por lo tanto, es una cita superficial. Sin embargo, si se menciona positivamente en otra sección del artículo más importante como en la metodología, resultados o conclusiones, la cita será importante. Sin embargo, si la polaridad es negativa o neutra, el impacto será *Negative* y *Perfunctory* respectivamente.

Resaltar que un impacto *Negative* no quiere decir que la cita no tenga relevancia para el autor o que este impacto sea realmente negativo, sino que el impacto será menor que si fuera *Perfunctory* o *Significant*. De esta forma se incentiva a los investigadores a que sus trabajos estén bien hechos y con resultados válidos para recibir mejor valoración de

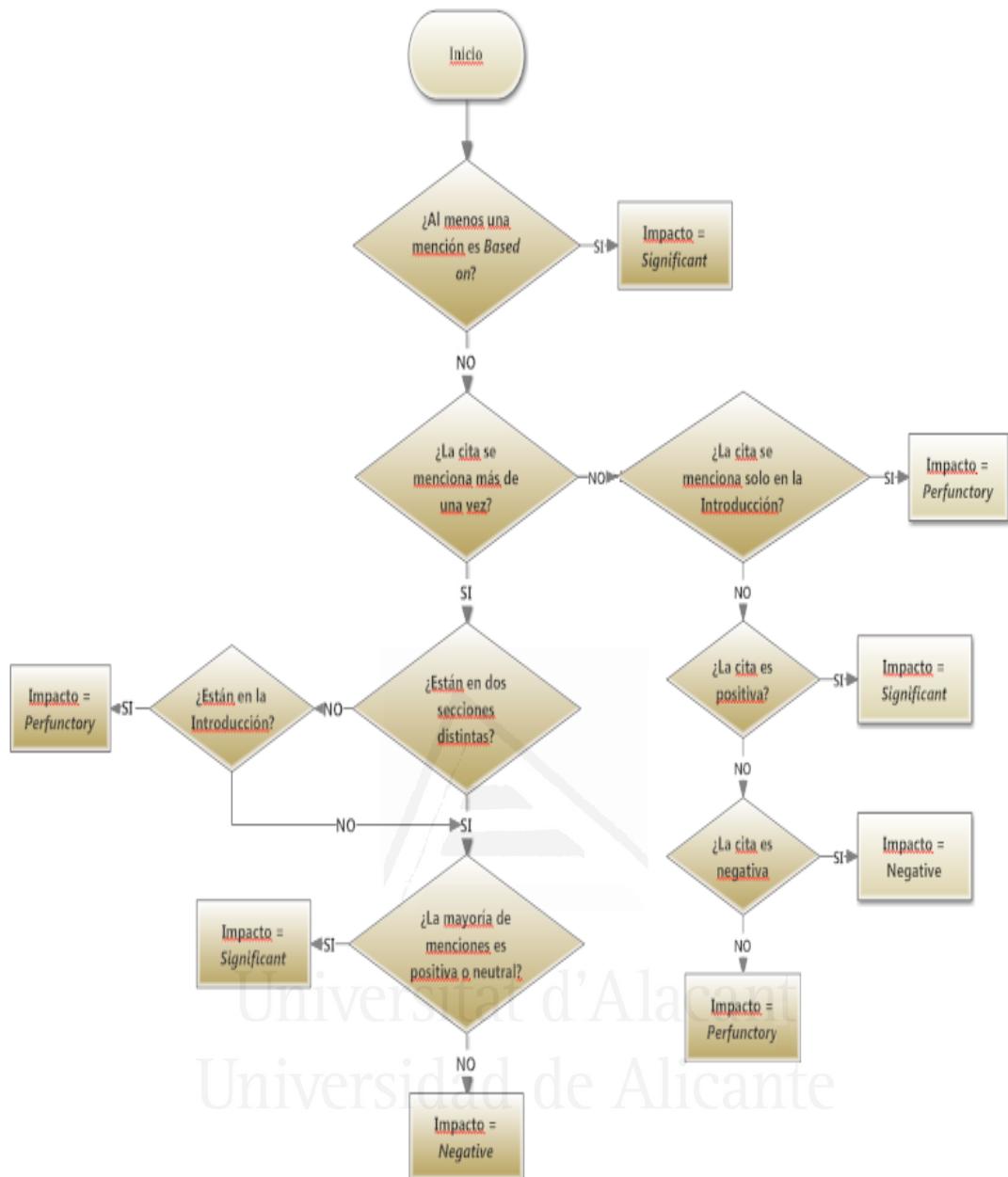


Figura 5: Algoritmo para valoración del impacto aplicando el Esquema de Clasificación propuesto

impacto y no se utilice las citas negativas expresamente para reducir los índices de un autor.

Siguiendo con el diagrama de la Figura 5, pero esta vez sobre las citas que se mencionan en más de una ocasión, observamos que si todas las citas están en la sección de introducción, serán superficiales pero si están en secciones distintas serán importantes o

negativas dependiendo de la polaridad de la mayoría de las referencias de cada una de las distintas citas.

En la Tabla 11 comparamos las calificaciones de los primeros autores con las obtenidas con nuestro enfoque y alcanzamos un valor medio de *F-Measure* de $92,2\% \pm 4,1$ una mejora significativa de un 14,9% con respecto al método que sólo utilizaba la polaridad para medir el impacto de la cita. Pero lo más destacado es la sustancial y significativa mejora (un 44,3%) de la categoría *Significant* al incluir más factores al cálculo del impacto.

Tabla 11: Comparación entre anotación realizada por los autores y los resultados de nuestro algoritmo de valoración de impacto

Impacto	Precision	Recall	F-Measure	Intervalo de confianza
Negative	1,0000	1,0000	1,0000	--
Perfunctory	0,9580	0,9268	0,9421	$\pm 0,0413$
Significant	0,8378	0,8611	0,8493	$\pm 0,1169$
Weighted Avg.	0,9316	0,9130	0,9221	$\pm 0,0414$

Estos resultados apoyan, no sólo la validez del algoritmo para calcular el impacto de las citas, sino también la calidad de las anotaciones pues si cualquiera de estos dos factores hubiera sido incorrecto, los resultados no hubieran sido tan positivos. Además, que la tasa de acierto esté tan próxima a los valores de acuerdo mutuo de los anotadores nos indica que la contribución de la tasa de error del algoritmo es baja.

La matriz de confusión para este ejemplo se despliega en la Tabla 12. Se puede apreciar que, en proporción, los errores más comunes son etiquetar como citas superficiales las importantes y viceversa pero ninguna importante se etiqueta como

Tabla 12: Matriz de confusión entre autores y nuestro algoritmo

		← clasificado como		
		Negative	Perfunctory	Significant
2	0	0		Negative
3	114	6		Perfunctory
0	5	31		Significant

negativa ni una negativa como importante.

Sin embargo, dada la poca cantidad de citas negativas y su poca significación estadística se requiere conseguir más datos para estar seguros que existe esta correlación tan fuerte entre la polaridad negativa y un impacto negativo. No obstante estos resultados preliminares son muy prometedores y avalan los resultados de los experimentos realizados en el Capítulo 7 sobre la calidad y utilidad del corpus.

4.3 Uso del aprendizaje automático para calcular el impacto

Una aproximación interesante hubiera sido realizar los experimentos de la sección anterior utilizando un sistema de aprendizaje automático en vez de un algoritmo pero dada el pequeño número de autores que nos contestaron hace difícil aplicar algún método supervisado. Pese a este impedimento, nos preguntamos si, en el caso de tener un corpus más grande, podría ser factible el uso de estos sistemas. Aunque el corpus anotado por los propios autores es pequeño, el corpus realizado en la presente tesis se compone de 2092 citas pero, por supuesto, estas citas no tienen valorada manualmente su impacto sino la función, polaridad, contexto, palabras claves y patrones que se describirán en los próximos capítulos. Para superar este escollo y dado los buenos resultados que hemos obtenido, decidimos utilizar el algoritmo de la sección anterior para etiquetar de forma automática todo nuestro corpus y entrenar un SVM con SMO para comprobar si sería posible utilizar un algoritmo de aprendizaje en vez de un algoritmo *ad hoc*. Además, aunque nuestro algoritmo tiene una cierta complejidad, no tiene en cuenta todas las posibles combinaciones de los distintos parámetros que nuestro modelo anota. Sin embargo, a un SVM se le puede añadir todas estas características anotadas del corpus y que sea el propio algoritmo de aprendizaje el que seleccione y potencie las más importantes de forma automática. También nos interesaba conocer si este sistema era capaz de aproximarse a los buenos resultados que obtuvimos con nuestro algoritmo para la valoración de impacto en las tres categorías consideradas y estimar su posible uso futuro cuando tengamos más citas con anotaciones de impacto.

Como se ha comentado, para realizar este experimento, se usó el corpus de artículos etiquetados y con los respectivos campos se forman los archivos .arff para WEKA, en donde los atributos son el número de ocurrencias por funciones desagrupadas, que ya

involucran una combinación de función y polaridad, por cada una de las cuatro secciones de IMRaD; las ocurrencias totales por funciones desagrupadas en todo el artículo y el número de secciones en las que está presente la cita. De las 2092 citas del corpus se escogió aleatoriamente 348 para entrenar el algoritmo anotado con el algoritmo de impacto de la sección anterior y, después, se utilizó 1744 citas para evaluar el modelo y así conseguir una buena significación estadística.

En la Tabla 13 se puede apreciar los resultados de *Precision*, *Recall* y *F-Measure* para cada una de las categorías de impacto y la media.

Tabla 13: Valoración del impacto usando SVM con SMO en WEKA

Impacto	Precision	Recall	F-Measure	Intervalo de confianza
Negative	0,97	0,891	0,929	±0,0372
Perfunctory	0,981	1	0,99	±0,0057
Significant	0,984	0,963	0,973	±0,0163
Weighted Avg.	0,98	0,981	0,98	±0,0066

Como se aprecia en la tabla, los resultados son muy prometedores, alcanzando, como media una tasa de acierto del 98% con una significación estadística acumulada de ±0,6%, aunque los valores de la categoría *Negative* son significativamente más bajos, llegando a un *Recall* de 89,1%. Estos resultados se debieron a que el número de citas *Negativas* es considerablemente menor y por lo tanto un el algoritmo de aprendizaje obtuvo muy pocas muestras de esta categoría para su entrenamiento.

Si observamos la matriz de confusión de la Tabla 14 podemos observar que el mayor número de errores se produjeron entre los *Negative* y los *Perfunctory* o de *Significant* por *Perfunctory*. No obstante, utilizando el SVM se puede apreciar que sí que hay confusión entre extremos, como confundir los *Negative* por *Significant* y viceversa, cosa que no ocurría con nuestro algoritmo. Una cosa a destacar que en la detección de las citas superficiales no comete ningún error y solamente lo comete al confundir las *Significant* y *Negative* por otras categorías.

Con esta evaluación podemos concluir que sin lugar a dudas podríamos sustituir el algoritmo de la Sección 4.2 por un sistema de aprendizaje automático y obtendríamos, como mínimo, similares resultados. Queda para trabajos futuros, cuando se obtenga más

anotaciones de los autores comprobar si con un algoritmo de aprendizaje se podría mejorar estos resultados tan prometedores.

Tabla 14: Matriz de confusión

Negative	Perfunctory	Significant	← clasificado como
163	14	6	Negative
0	1181	0	Perfunctory
5	9	366	Significant

4.4 Conclusiones del capítulo

El impacto puede evaluarse a partir de las anotaciones en el corpus realizadas con las especificaciones definidas en el capítulo anterior. Se proponen dos métodos para valorar el impacto. El primer método relaciona directamente el impacto con los tres niveles de polaridad: negativa, neutral y positiva para asignarles una categoría de impacto que puede ser *Negative*, *Perfunctory* o *Significant*, respectivamente. Esta consideración se justifica tomando en cuenta que los autores tienden a tener una disposición más favorable hacia citas que son más cercanas y alineadas con su propio trabajo.

El algoritmo desarrollado se valida con el análisis de los resultados obtenidos en una consulta a autores sobre las citas que referenciaron en sus estudios. Se muestra una alta correlación (weighted average para *F-Measure* de 0,98) entre la anotación realizada por los propios autores con la efectuada con nuestro método. Estos resultados validan el método desarrollado para clasificar los tres niveles de impacto: *Negative*, *Perfunctory* y *Significant*.

Usamos el algoritmo para etiquetar el impacto en forma automática en los artículos en el corpus. Con base en la información del corpus guardada en la base de datos, transformamos a un archivo .arff que ingresamos a un clasificador SVM con regresión SMO. Con el aprendizaje automático obtenemos excelentes resultados, con los que se demuestra la factibilidad de usar esta clasificación en distintas aplicaciones, entre ellas la

de enriquecer los métodos actuales de medición de impacto de un autor, que actualmente se basan únicamente en conteo de citas.



5. Metodología de anotación

El principal escollo para generar un corpus estándar para análisis de citas, es la dificultad para lograr un buen acuerdo entre anotadores; sin embargo, este acuerdo es un requerimiento indispensable para certificar que la anotación es confiable y reproducible.

El proceso de etiquetado es una tarea compleja que involucra errores por lo que generalmente se logra un bajo acuerdo, aún cuando sea realizado por anotadores conocedores de los temas que hayan pasado por un entrenamiento apropiado. Esta clase de codificación requiere una lectura minuciosa pero, aún así, los modelos mentales de los distintos participantes generalmente no convergen en una opinión compartida. Debido a esta situación se produce un bajo índice de acuerdo entre anotadores (Ciancarini, Di Iorio, Nuzzolese, Peroni, Vitali 2014). Esta situación es más problemática para esquemas de anotación de grano fino, en los cuales los etiquetadores no aplican todas las posibilidades sino solamente algunas de ellas y las diferencias entre las clases son muy sutiles o porque las clases que tienden a escoger les resultan más claras o familiares. De todas maneras, cada persona se creará un modelo mental particular que guiará hacia una clasificación que probablemente no coincida con la de los otros anotadores.

La forma de trabajo para anotar un corpus para análisis de citas, basándose en un esquema de clasificación, y sin aplicar una estrategia adicional para mejorar el acuerdo, consiste en los siguientes pasos: primero, los codificadores deben seguir un procedimiento de familiarización y entrenamiento en la aplicación del esquema; luego les corresponde leer el artículo y tratar de entender cada pedazo del texto, dentro del contexto, para inferir la función y la polaridad con las que fue realizada la cita; y, por último, deben decidir qué clases son las que mejor calzan para interpretar este propósito dentro del esquema. Como se mencionó, para cada paso el anotador se crea un modelo mental (Davidson, Dove, Weltz

1999) que le permite tomar decisiones, basándose en el conocimiento del esquema y el respectivo mapeo con el entendimiento del texto que leyó. Las guías de anotación y el entrenamiento de los codificadores tienen como objetivo disminuir las discrepancias entre el modelo del esquema de clasificación y el modelo mental que se crea cada anotador.

Según Davidson, et al., 1999, el modelo mental es una representación interna que se construye usando la información disponible, es inestable y sujeto a cambio, sirve para definir estrategias de solución de problemas y se nutre de la realimentación que proporcionan los resultados de las decisiones. Si no se proporciona a los codificadores un mecanismo que facilite y estandarice la construcción de ese modelo mental, los resultados de la codificación van a ser muy variables entre anotadores y van a haber diferencias hasta en la clasificación realizada por una misma persona en diferentes momentos.

Estas situaciones las constatamos en las primeras fases del presente trabajo, cuando, aplicando un esquema de clasificación de funciones de grano mediano, con un grupo de 3 anotadores conocedores del tema de los artículos que se estaban anotando, se presentaban grandes divergencias entre ellos, puesto que no coincidían en la percepción de la función de los mismos textos, lo que llevaba a obtener consistentemente bajos índices de acuerdo, incluso para la clasificación de los tres niveles de polaridad, que podría catalogarse como un problema menos complejo que el de clasificación de la función. El desacuerdo empezaba con la identificación de la longitud del contexto relevante a las citas y continuaba al tratar de clasificar la función y la polaridad de la cita en ese contexto.

La falta de un acuerdo entre anotadores limita la aplicación de cualquier esquema, independientemente de cuán bien diseñado se encuentre. Por esta razón vimos la necesidad de crear una estrategia para facilitar que los anotadores elaboren un modelo mental apropiado que coincida con el de quienes diseñaron el esquema y con el de los otros anotadores que están realizando la tarea. Esta estrategia consiste en añadir una etapa de pre-anotación a las tradicionales de pre-procesamiento y clasificación. Se agrega además un procesamiento automático para detectar características relacionadas con el impacto de un artículo en un documento, como se puede observar en la Figura 6.

La diferencia entre una metodología tradicional de anotación y la nuestra se encuentra en el proceso de pre-anotación de etiquetas y palabras clave, para la formación de un modelo mental estandarizado que lleva a un buen acuerdo entre los anotadores del corpus destinado al análisis del contexto de citas bibliográficas. Las etiquetas y palabras clave se relacionarán directamente con determinadas funciones y polaridades. Este pre-etiquetado proporciona información sobre el contenido y la estructura del contexto de las

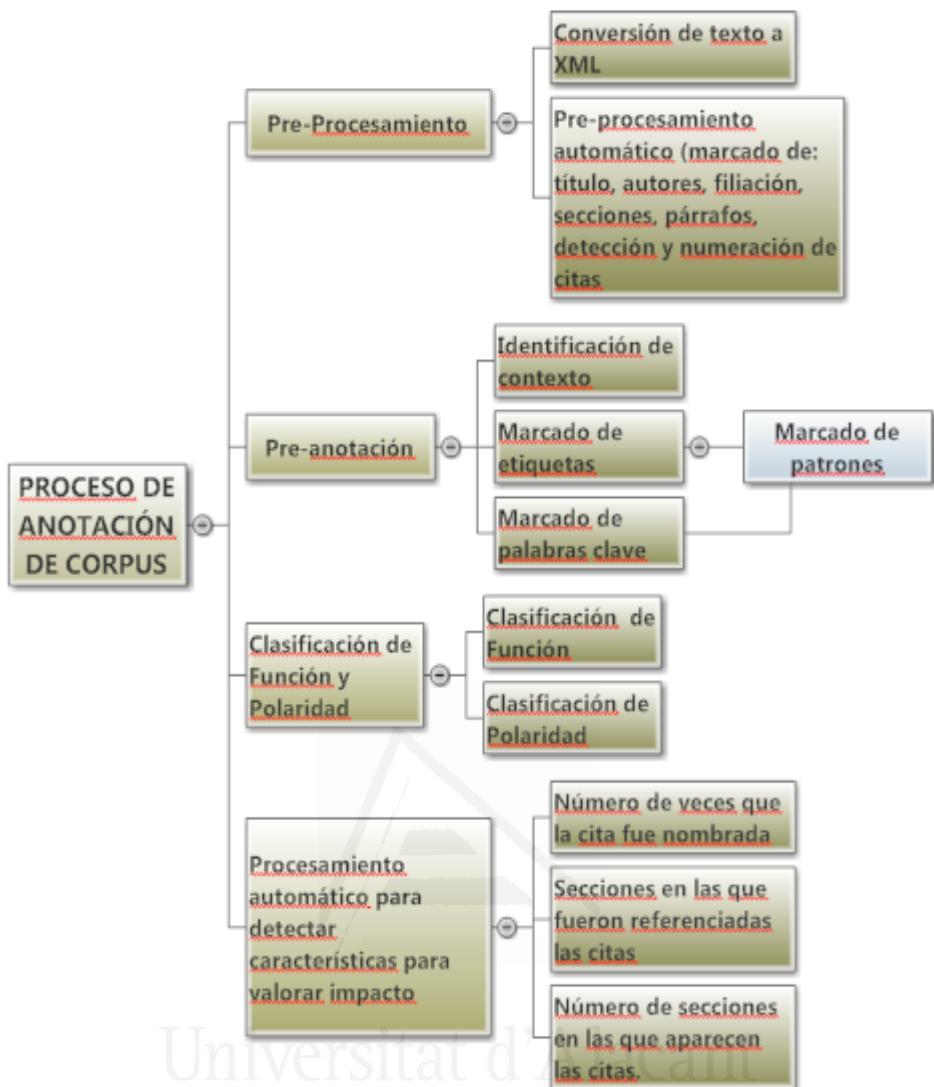


Figura 6: Proceso de anotación del corpus

citas. Esta información puede ser usada en muchas aplicaciones de análisis de contexto de citas.

El proceso de pre-anotación permite la creación de modelos mentales coincidentes entre los distintos codificadores, mediante una forma no gráfica de obtener mapas conceptuales para detectar ideas y relacionarlas en forma estructurada dentro del contexto de la cita. Este procedimiento requiere un trabajo adicional en el etiquetado, pero tiene como ventaja que involucra el marcado semántico del texto alrededor de la cita, que proporciona información valiosa que puede ser usada en muchas aplicaciones de análisis del contenido del contexto de referencias bibliográficas.

Para la codificación se usó el entorno de desarrollo NetBeans IDE 8.0.2 que soporta la escritura de archivos XML. Los archivos anotados son procesados para obtener los

contextos de las citas con sus respectivas anotaciones. Esta información se la guarda en una base de datos, para facilitar su manipulación; la información en la base de datos se vuelve a procesar para conformar las entradas a los algoritmos que se usarán. Los archivos anotados en XML, tienen la propiedad de que mantienen toda la información inicial del artículo, y, por lo tanto, fácilmente puede volver a obtenerse el texto original; con esto se cumple uno de los principios de la anotación de un corpus, de acuerdo a Leech (1993).

La metodología de anotación consta de una fase de pre-procesamiento automático que prepara el texto, una segunda fase en la que se efectúa la anotación humana y una final en la que se obtienen automáticamente características adicionales que son útiles en la valoración del impacto.

5.1 Pre-procesamiento y definición del contexto

El pre-procesamiento consta de dos pasos: en el primero, el corpus se convierte desde texto a XML. En el segundo paso: se etiquetan el título de los artículos y la filiación de sus autores, se dividen los artículos en secciones y párrafos. Luego se pasa a una fase de reconocimiento y enumeración de las referencias en el texto. Se desarrolló un programa que reconoce expresiones regulares para detectar las citas bibliográficas en el formato oficial de la Antología de la ACL.

Una vez que los artículos están pre-procesados, para aplicar nuestro esquema de anotación, es necesario definir primero cuál va a ser el contexto de la referencia. Este contexto debe contener solamente el texto relevante adyacente a la cita para poder discriminar su función y polaridad. El contexto no puede ser muy amplio para que el coste de anotación no sea excesivo y que el modelo mental no se vuelva muy complicado.

En los experimentos previos se pudo observar que si no se proporcionaba una norma específica para definir el contexto, los anotadores definían su longitud en forma muy variable, que iba desde considerar solamente la oración que contiene la cita, a varios párrafos alrededor de ella. Se pudo apreciar que resultaba muy difícil para estas personas lograr un acuerdo para demarcar las oraciones relacionadas directamente con la cita, solamente aplicando el criterio de que el contexto debe contener la información necesaria para clasificar función y polaridad.

Para disminuir la complejidad que presenta la delimitación del contexto de una referencia, se resolvió establecer, como primera aproximación, el límite exterior del contexto como la longitud del párrafo que contiene la cita. Por esta razón, a los anotadores se les entrega el corpus con una pre-anotación inicial en la que se definen los párrafos con la etiqueta <paragraph>. Para tomar la decisión de que los límites del contexto estén dentro de un párrafo, partimos del hecho de que, por definición, un párrafo es el conjunto de oraciones que expresan una idea o argumento completo, por lo tanto, un párrafo tiene una buena posibilidad de incluir la argumentación relevante sobre una referencia. Por lo tanto, se le pide al anotador, que dentro del párrafo, detecte la información relevante sobre la cita, para ello usa la etiqueta <context>. A continuación presentamos un ejemplo de lo que se tiene luego del pre-procesamiento y anotación del contexto.

<paragraph>Recent years have witnessed a significant growth of research into weakly supervised ML techniques for NLP applications. <context>Different approaches are often characterised as either multi- or single-view, where the former generate multiple redundant (or semi-redundant) 'views' of a data sample and perform mutual bootstrapping. This idea was formalised by <cite id="5">Blum and Mitchell (1998)</cite> in their presentation of co-training.</context> Co-training has also been used for named entity recognition</paragraph>

Otra ventaja de fijar un párrafo como el límite externo del contexto es que facilita el trabajo de los anotadores que no tienen que leer varios párrafos alrededor de la cita para poder delimitar el contexto. En el resto de ejemplos que se presentarán en este capítulo se tomará en cuenta solamente el contexto de cada cita, sin embargo no se mostrará la etiqueta <context>.

En los resultados de la anotación se pudo apreciar que la longitud del contexto

Tabla 15: Tamaño del contexto seleccionado por los anotadores

Tamaño del contexto	Número de ocurrencias
Una sentencia	1502
Dos sentencias	377
Tres sentencias	127
Cuatro sentencias	56
Cinco o más sentencias	30

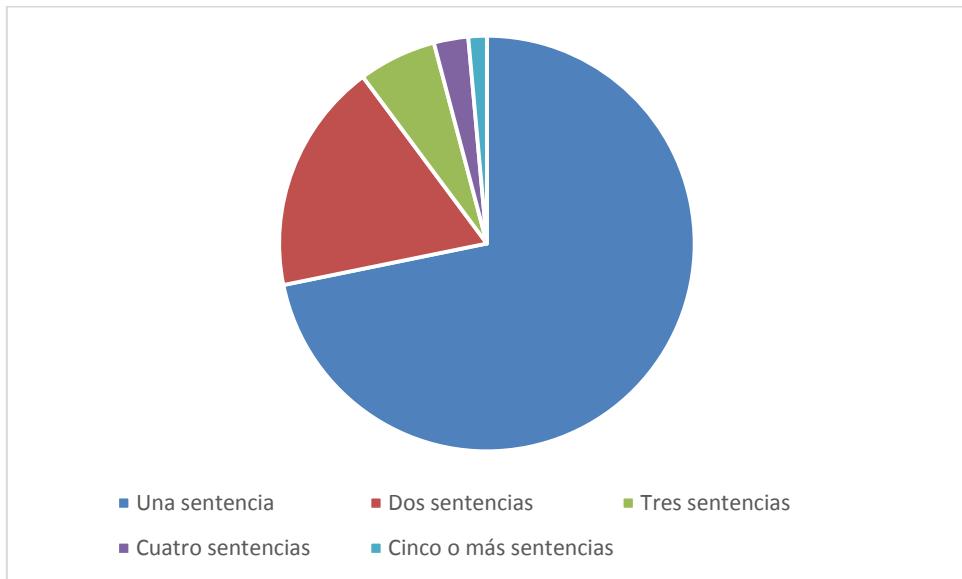


Figura 7: Tamaño del contexto seleccionado por los anotadores

escogido por los anotadores mayormente corresponde a una sentencia: la que contiene la cita. Se presentan con menor, pero aún significativa frecuencia, contextos con longitud de dos y tres sentencias. Es decir que probablemente el contexto no necesita incluir más que esos valores para contener toda la información que se requiere para el marcado con miras a la clasificación de las citas.

En la Figura 7 podemos ver los porcentajes de citas según el tamaño de su contexto. En esta gráfica apreciamos que con sólo las 3 primeras sentencias cubrimos el 95,6% del contexto de las citas.

En caso de que el contexto tenga más de una sentencia, el número de oraciones antes y después de la cita es balanceado. Los resultados se muestran en la Tabla 16. Se puede

Tabla 16: Número de sentencias previas y posteriores a la cita dentro del contexto

Número de Sentencias en el contexto	# de contextos que tienen oración antes de la cita	# de contextos que tienen dos oraciones antes de la cita	# contextos que tienen tres oraciones antes de la cita
Sólo la sentencia que tiene la cita	0	0	0
Dos sentencias	165	212	0
Tres sentencias	91	36	0
Cuatro sentencias	22	27	7

apreciar que las oraciones en el contexto se distribuyen balanceadamente antes y después de la cita. En este sentido no se detecta un patrón claro.

5.2 Pre-anotación

Este proceso se inicia con la definición de la sección a la que pertenece la referencia según el apartado donde se encuentra en el artículo usando el esquema de Introducción, Métodos, Resultados y Discusión (IMRaD). En el caso de que las secciones estén divididas con otros nombres, el anotador deberá decidir su correspondencia con secciones equivalentes dentro del esquema IMRaD.

En la pre-anotación el codificador marca el texto con etiquetas y palabras clave, información que tiene contenido semántico que será usado con dos objetivos:

- La formación de un modelo mental claro, de algún modo estandarizado, que posibilite la coincidencia el modelo mental de quien diseñó el esquema y de los otros anotadores pero sin condicionarles la respuesta.
- Como características de entrada para clasificadores con aprendizaje automático supervisado.

Este etiquetado inicial está compuesto de las siguientes etapas:

1. *Reconocimiento del contexto relevante para las citas*, para encontrar el texto que directamente habla sobre ellas y que contiene la información necesaria para su clasificación. El contexto estará contenido en el párrafo en el que se encuentra la cita.
2. *Identificación de los patrones en el texto formados por etiquetas y palabras clave* consistentes de n-gramas con una longitud de hasta cinco palabras. Más adelante se presentan ejemplos de estos patrones para cada función.

Un patrón, en este entorno, es una ordenación de palabras que nos sirve para revelar el propósito y la disposición del autor al realizar la referencia que se relaciona con la función que tiene la cita dentro del artículo que la menciona; y, con la polaridad favorable o no hacia la cita. Gracias al proceso de pre-anotación, aumenta considerablemente el acuerdo entre anotadores, logrando niveles altos que acreditan la validez de la anotación.

Adicionalmente, los patrones que se van obteniendo a partir del etiquetado manual se usan para mejorar la guía de anotación, obteniéndose un listado que ayuda a resolver

cualquier duda de los anotadores. Además, proporcionan información para realizar generalizaciones respecto a la forma sintáctica de lo que se dice sobre la referencia y su relación con función y polaridad.

Por otro lado, estas generalizaciones podrían ser usadas como modelos de entrada para automatizar el proceso de anotación.

El proceso de pre-anotación sirve para facilitar la construcción de modelos mentales coherentes y compartidos entre los anotadores, pues detectar los patrones macro del contexto sirve para aclarar a los humanos la estructura de lo que se dice respecto a las citas y para adquirir el significado que se les atribuye. Los patrones son estructuras que pueden correlacionarse con mapas conceptuales que ayudan a ordenar las relaciones entre los conceptos presentes en el texto.

Los patrones están compuestos por una parte variable que corresponde a conceptos generales como: herramienta, datos, método, etc., y una parte fija constituida por secuencias de palabras en la forma de n-gramas formados por palabras clave relacionadas con las distintas polaridades y funciones. Por ejemplo, se anotarían como palabras clave: "a lot of effort", "manually or randomly", "state-of-the-art" o "however".

Agregamos esta información usando un esquema de etiquetas en XML. El conjunto de etiquetas mostrado contiene conceptos que se han ido definiendo en el proceso de anotación de acuerdo a su necesidad, tales como: trabajo citado, autor, teoría, acción, método, conceptos, tarea, resultados, experimento, característica positiva, característica negativa, etc. Las etiquetas se diseñaron de tal manera que cubren, con la menor ambigüedad posible, el mayor número de posibilidades. Por ejemplo, la etiqueta <cite></cite> se refiere a las referencias y tendrá atributos como identificación, función y polaridad, la etiqueta para palabras clave tendrá esta forma: <kw>secuencias_de_palabras o n-grama</kw>. La etiqueta XML para el concepto trabajo citado <cite> lucirá así:

```
<cite id="número_de_cita_en_el_artículo" function="función_de_la_cita" polarity="polaridad_de_la_cita">instancia_de_la_cita</cite>
```

Por ejemplo, si se tiene el texto: "Our classifier is based on the detailed previous work by Dong and Schäfer, 2011", la correspondiente anotación será:

```
"<author>Our</author> <tool>classifier</tool> <kw>is based on </kw> the
<posfeature>detailed </posfeature>previous work by <cite
id="número_de_cita_en_el_artículo" function="bas" polarity="pos">Dong and
Schäfer, 2011</cited>
```

Y el patrón que se obtiene es: AUTHOR TOOL KW POSFEATURE CITED.

AUTHOR, TOOL y CITED son la parte variable y las secuencias: "is based on" y "detailed" son las palabras clave o keyword. A la hora de decidir por parte del anotador las categorías de función y polaridad a la que pertenece esa cita, se le mostrará el siguiente patrón:

AUTHOR TOOL is based on POSFEATURE CITED

Es decir, podrá ver las palabras claves, sin colapsar en su etiqueta, pues éstas dan mucha información sobre el propósito y polaridad de la cita en el texto. Con este patrón que se muestra a los anotadores, es muy fácil deducir que "*la herramienta del autor está basada en una característica positiva del artículo citado*" y por lo tanto, la cita tendrá la función *Based on, Supply* y la polaridad *Positive*.

En este ejemplo es bastante claro, porque las palabras clave lo establecen, que la cita corresponde a la función *Based on, Supply* ya que la herramienta citada se ha usado en el trabajo del autor que cita. La polaridad es positiva porque se califica como "detailed". Por lo tanto se desagrega la función como *Based on*.

Cualquier sentencia puede ser producida con este patrón. Por ejemplo, para el siguiente texto:

Our algorithm is based on the detailed Vector Space Model – VSM (Salton et al., 1975)

La anotación sería:

```
<author>Our</author><tool>algorithm</tool><kw>is based on</kw> the  
<posfeature>very complete</posfeature> Vector Space Model – VSM  
(<cited>Salton et al., 1975</cited>)
```

El patrón que se consigue es: AUTHOR TOOL KW POSFEATURE CITED.

En este ejemplo se aplica el mismo análisis puesto que el patrón sugiere que una herramienta del artículo que cita fue desarrollada usando como base el trabajo citado, al que además se le distingue positivamente con la secuencia de palabras con sentido positivo: "very complete".

Las palabras clave, las calificamos como n-gramas porque son cadenas consecutivas de hasta cinco palabras. Durante el proceso de anotación se han evaluado distintos tamaños máximos para estas secuencias y hemos constatado que la longitud de cinco palabras permite que estas secuencias tengan la información relevante sin que se vuelvan muy específicas, de este modo se las podrá encontrar en otros textos en relación con una determinada función o polaridad.

En la Tabla 17 se muestran todas las etiquetas de nuestro modelo de pre-anotación y sus conceptos asociados. En total se compone de 18 etiquetas que se corresponden con 18 conceptos distintos.

La etiqueta para *cita* se refiere a la que marca la referencia en el texto. Esta etiqueta tiene al número de identificación, como atributo generado automáticamente en el pre-procesamiento. Presenta también dos atributos que se marcan en el corpus en la etapa de anotación que son función y polaridad.

Tabla 17: Etiquetas

Conceptos	Etiquetas XML	Etiquetas
Trabajo citado	<cited>	CITE
Autor que cita	<author>	AUTHOR
Teoría	<theory>	THEORY
Acción	<action>	ACTION
Método	<method>	METHOD
Datos	<data>	DATA
Herramienta	<tool>	TOOL
Concepto	<concept>	CONCEPT
Tarea	<task>	TASK
Resultado	<result>	RESULT
Persona(s)	<person>	PERSON
Experimento	<experiment>	EXPERIMENT
Campo del conocimiento	<field>	FIELD
Artículo	<paper>	PAPER
Característica	<feature>	FEATURE
Característica positiva	<posfeature>	POSFEATURE
Característica negativa	<negfeature>	NEGFEATURE
Característica negativa	<negfeature>	NEGFEATURE

La etiqueta para *author* corresponde a una mención explícita o implícita que el autor realiza sobre sí mismo, puede ser un pronombre, por ejemplo: "we", "our" o el nombre de una herramienta que ha sido desarrollada por el autor, o expresiones como "our work", "our algorithm", etc. La etiqueta para autor se requiere que esté presente en el contexto para la función *Based on, Supply* de modo que se defina que la fuente ha sido usada por el autor que hace la referencia, lo que puede inferirse del texto, aunque no esté explícitamente expresado. La etiqueta para *autor* también puede corresponder a una referencia a sí mismo, al referirse a su trabajo en algún otro artículo.

Theory representa a un grupo coherente de afirmaciones que aún no se tienen como hechos comprobados, pero que ya han sido confirmadas por un número substancial de experimentos. Ejemplo: "*computational theory*", "*semantic theory*".

Action se califica a algo que se hace, se marcan verbos activos que expresan lo que un sujeto realiza y tienen un objeto que recibe esa acción. Por ejemplo: "*develop*", "*annotate*", "*provide*". No se aplicaría a verbos que expresan un estado como "*remain*", "*result*", "*keep*", "*be*".

Method se refiere a una técnica que se usa para resolver un problema: un diagrama, una representación de conocimiento, un tipo de análisis de datos de un problema o una función matemática. Ejemplos: "*Argumentative Zoning*", "*sentiment analysis methods*", "*co-citation*", "*hierarchical and syntax augmented models*". *Data* pueden ser un corpus, un conjunto de n-gramas, etiquetas de *Part_Of_Speech*, datos anotados, texto simple, etc. Ejemplos: "*ACL Anthology Network*", "*British National Corpus*", "*blog*", "*OntoNotes*".

Contribution puede ser una fuente, un descubrimiento, la especificación de un problema, una actividad realizada, información relevante a un problema de investigación, una detección importante, etc. Ejemplos: "*insights derived here*", "*new solution presented in figure 1*".

Tool puede referirse a un software o a un algoritmo para procesamiento de datos como un parser, un clasificador, una implementación en un lenguaje de computadora, un modelo computacional entrenado, un traductor, etc. Ejemplos: "*SProUT system*", "*SVM classifier*", "*MEDIE service*".

Concept se etiqueta a una noción general, una idea abstracta o concreta, un principio explicitorio de un sistema científico, un almacenamiento de información. Pueden ser conceptos abstractos o concretos, por ejemplo: "*information*", "*database*", respectivamente. Pueden estar relacionados con el total y las partes; "*paper*", "*section*", "*paragraph*".

"citation context". Otros ejemplos: *"Impact factor"*, *"rhetorically-based schema"*, *"Maximum entropy (ME) models"*, *"classification scheme"*.

Task es un trabajo por realizar o ya efectuado. Ejemplos: *"Analysis of citing sentences"*; *"mining de web for bilingual texts"*; *"shallow parsing"*; *"table extraction"*; *"classifying sentences"*.

Result se etiqueta para producto final, consecuencia, conclusión. Ejemplos: *"high precision data"*; *"ample evidence"*; *"findings"*, *"outcomes"*.

Person se codifica cuando se menciona a individuos o grupos humanos cuyos miembros tienen alguna característica en común. Ejemplos: *"researchers"*, *"clinicians"*, *"they"*, *"writers and readers"*, *"right wing representatives"*.

Experiment se codifica para operación destinada a descubrir, comprobar o demostrar determinados fenómenos o principios científicos. Ejemplos: *"analysis"*; *"research space"*; *"first study"*; *"research in"*.

Field se anota para el ámbito de una actividad o de un conocimiento. Ejemplos: *"citation context analysis"*, *"speech recognition"*, *"image management"*, *"social sciences"*.

Paper se refiere a un artículo científico que no se nombra explícitamente o que su mención se realizó anteriormente. Ejemplos: *"mentioned work"*, *"previous article"*, *"some of this production"*.

Feature propiedad o comportamiento que distingue a algo. Ejemplos: *"menu items"*, *"semantic characteristics"*.

Posfeature se anota cuando se tienen características positivas. Ejemplos: *"in-depth approach"*, *"deeper analysis"*, *"improved performance"*, *"converges much faster"*.

Negfeature se codifica cuando se tienen características negativas. Ejemplos: *"have shortcomings"*, *"main problem"*, *"lead to the wrong prediction"*, *"does not handle"*, *"is not flawless"*.

Una notación simbólica de los patrones, constituidos por palabras clave y etiquetas que resume nuestro esquema de pre-anotación es la que se presenta en la Figura 8, donde las etiquetas de los patrones corresponden a conceptos que se muestran en la Tabla 17.

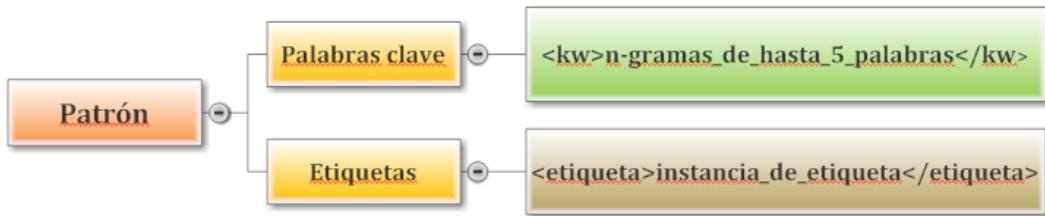


Figura 8: Notación simbólica de los patrones

El grupo de etiquetas ha sido desarrollado experimentalmente de acuerdo a las necesidades de anotación de forma que se minimice la ambigüedad. Al mismo tiempo, de acuerdo a los resultados de las anotaciones, se ha ido depurando la lista presentada. Al definir estas etiquetas, sin embargo, hemos advertido que los anotadores deben tener cuidado con frases nominales que pueden llevar a equívoco. Por ejemplo, la frase nominal “*Maximum Entropy Implementation*”, aunque el término “*Maximum Entropy*” por sí mismo puede referirse a un método, en este caso está con la función sintáctica de adjetivo que es usado para la calificación del nombre “*Implementation*”. En este caso se refiere a un algoritmo, y por lo tanto, corresponde a una TOOL y no a un METHOD.

Como un importante valor añadido, cuando se combinan estas etiquetas con las funciones y polaridad definidas en el esquema explicado en el Capítulo 3, añaden granularidad a la clasificación como se puede ver en la Tabla 6 y el ejemplo de la Figura 1.

Para ilustrar el etiquetado en la pre-anotación previa a la clasificación de funciones y polaridad, en a continuación presentaremos ejemplos. Hasta el momento sólo hemos considerado referencias bibliográficas en el idioma inglés, puesto que es el lenguaje común de la ciencia; sin embargo, opinamos que los principios de la metodología tales como la identificación de patrones formados por etiquetas y palabras clave; y, el proceso de clasificación que usa información como características , podrían ser aplicados también en otros idiomas; por supuesto sería necesario tomar en cuenta particularidades especiales de los distintos idiomas y multilingüismo.

Ejemplo de pre-anotación 1

We compare our zone classifier to a reimplemention of Teufel and Moens's NB classifier and features on their original Computational Linguistic corpus. Like Teufel (1999) we model zone classification as a sequence tagging task. Our zone classifier achieves an F-score of 96.88%, a 20% improvement.

Anotación: We <kw>compare</kw> <author>our zone classifier</author> to a reimplemention of <tool>Teufel and Moens's NB classifier</tool> and features on their original Computational Linguistic corpus. Like <cite>Teufel

(1999)</cite> we model zone classification as a sequence tagging task. <author>Our zone classifier</author> <kw>achieves</kw> an F-score of 96.88%, a 20% <posfeature> improvement</posfeature>.

Patrón: KW AUTHOR CITE AUTHOR KW POSFEATURE

Comentario: Aquí se evidencia una comparación de una herramienta del autor del artículo con la de otro autor que es citado; el resultado de la comparación es favorable para el autor que referencia. Es importante enfatizar que sólo se etiquetan los patrones relevantes para la clasificación. Las palabras clave son “*compare*” y “*achieves*”. El grupo de funciones que se anota es “*Contrast*” y su polaridad es “*Negative*” pues la cita pierde en la comparación, nótese que la polaridad es negativa aunque se tenga una etiqueta de <posfeature>, esta aparente contradicción vuelve complicada la detección de esta polaridad; para el anotador humano no presenta demasiada dificultad, porque puede verse que la característica positiva se refiere al trabajo del autor y no a la cita, pero para la clasificación automática esta forma de expresarse puede plantear retos para la identificación de la clase.

Por lo tanto los atributos de la referencia que corresponden a la clasificación de función y polaridad son: <cite id="número_de_la_cita" function="con" polarity="neg">Teufel 1999</cite>.

Ejemplo de pre-anotación 2

Tomemos la siguiente oración como un segundo ejemplo:

The baseline score shown in bold, is obtained with no context window and is comparable to the results reported by Athar (2011) .

Anotación: <author>The baseline score</author> shown in bold, is obtained with no context window and <kw>is comparable to</kw> the results reported by <cite>Athar (2011)</cite>.

Patrón: AUTHOR KW CITE

Comentario: La etiqueta semántica es AUTHOR, las palabras clave corresponden al n-grama “*is comparable to*” que denota comparación. Con este patrón, los anotadores pueden detectar que los autores están comparando resultados propios con los de la cita. Que los resultados de “*The baseline score*” son propios del autor y se puede deducir porque se encuentran en la parte de presentación de sus descubrimientos experimentales que, en

este caso, aunque no se muestra, están dentro del contexto de la cita. Los patrones ayudan a determinar el propósito o intención del autor al realizar la referencia que en este caso pertenece a la función “*Contrast*”, y la polaridad para la referencia es “*Positive*” porque la usa como referencia para calificar sus resultados como exitosos. Por lo tanto los atributos de la referencia que corresponden a la clasificación de función y polaridad son: <cite id="número_de_la_cita" function="con" polarity="pos">Teufel 1999</cite>.

5.3 Clasificación de la función y la polaridad

Con la información que se ha marcado en la pre-anotación, al anotador le resulta más fácil tomar la siguiente decisión que es la de clasificar la función y polaridad de la referencia. Gracias a los modelos mentales que se han creado usando las etiquetas y las palabras clave, en buena medida se habrá eliminado la ambigüedad y el anotador estará en capacidad de coincidir, en la clasificación de función y polaridad, con el resto de personas que están realizando esta tarea.

La función y la polaridad se ingresan como atributos en la etiqueta de cada cita, en formato XML. A continuación presentamos algunos ejemplos de aplicación para obtener las funciones a partir de las etiquetas y palabras clave.

Ejemplo de clasificación para la Función Based on, Supply

Our classifier uses the maximum entropy implementation described in Curran and Clark (2003).

Anotación: <author>our classifier</author> <kw>uses</kw><method>maximum entropy implementation</method> <cite id= "number" function="bas" pol="neu">Curran and Clark 2003</cite>

Patrón: AUTHOR uses METHOD CITE.

Función: *Based on, Supply*; **polaridad:** *Neutral*; **función desagregada:** *Supply* porque el autor usa el método de la cita, no hay palabras clave que denoten polaridad positiva o negativa.

Ejemplo de clasificación para la Función Useful

In scientific texts, knowing the type of information that a zone represents (e.g., background knowledge, hypothesis, experimental observation, conclusion, etc.) allows for automatic isolation of new knowledge claims (Sandor and de Waard, 2012).

Anotación: <method>background knowledge, hypothesis, experimental observation, conclusion, etc. </method> <kw>allows</kw> <task>automatic isolation of new knowledge claims</task> <cite id="number" function="use" polarity="pos">Sandor and de Waard 2012</cite>

Patrón: METHOD allows TASK CITE.

Función: “*Useful*” porque el método se usa en la realización de una tarea; **polaridad:** “Positive” porque “*allows*” tiene una connotación positiva.

Ejemplo de clasificación para la función Acknowledge, Corroboration, Debate

Some example applications include part-of-speech (pos) tagging (Ratnaparkhi, 1996), parsing (Johnson et al., 1999), language modelling (Rosenfeld, 1996), and text categorisation (Nigam et al., 1999).

Patrón: TASK CITE TASK CITE TASK CITE TASK CITE

Anotación: <tool>part-of-speech (pos) tagging</tool> <cite id="number" function="use" polarity="neu"> Ratnaparkhi, 1996</cite> <tool>parsing</tool> <cite id="number" function="use" polarity="neu"> Johnson et al., 1999</cite> <tool> language modelling </tool> <cite id="number" function="use" polarity="neu"> Rosenfeld, 1996</cite> <tool> text categorisation </tool> <cite id="number" function="use" polarity="neu"> Nigam et al., 1999</cite>

Función: “*Acknowledge, Corroboration, Debate*”; **polaridad:** “*Neutral*”; **función desagregada:** “*Acknowledge*” porque en las diferentes citas simplemente se está reconociendo la existencia de esas tareas.

Ejemplo de clasificación para la Función Contrast

Gasperin (2009) presents a full annotation of anaphora and coreference in biomedical text, but only noun phrases referring to biomedical entities are considered. In contrast, Cohen et al. (2010) build a corpus of 97 full-text journal article in the biomedical domain where every co-referring noun phrase is annotated (CRAFT - Colorado Richly Annotated Full Text).

Patrón: CITE presents DATA but only In contrast CITE ACTION DATA where every

Anotación: <cite id="number" function="con" polarity="neg">Gasperin (2009)</cite> <kw>presents</kw> a full annotation of anaphora and coreference in <data>biomedical text</data>, <kw>but only</kw> noun phrases referring to biomedical entities are considered. <kw>In contrast</kw>, <cite id="number" function="con" polarity="pos">Cohen et al. (2010)</cite> <action>build</action> a <data>corpus</data> of 97 full-text journal article in the biomedical domain <kw>where every</kw> co-referring noun phrase is annotated (CRAFT - Colorado Richly Annotated Full Text).

Para la cita 1 la **función:** *Contrast*; **polaridad:** *Negative*. La clave para la clasificación como *Contrast* son las palabras clave *In contrast*, la función es negativa como lo denotan las palabras clave *but only*.

Para la cita 2 la **función:** *Contrast*; **polaridad:** *Positive*. Igualmente la clave para la clasificación con la función *Contrast* son las palabras clave *In contrast*, la función es positiva como lo revelan las palabras clave *where every*.

Ejemplo de clasificación para la Función Weakness, Correct

From this we can see that the n-grams (unigrams and bi-grams) have by far the largest impact – but these features were not directly implemented by Teufel and Moens (2002).

Patrón: but METHOD were not directly implemented by CITE.

Anotación: <kw>but</kw> <method>n-grams (unigrams and bi-grams)</method> <kw>were not directly implemented by </kw> <cite id = "number" function="wea" polarity="neg">Teufel and Moens (2002)</cite>

Función: “Weakness, Correct”; **polaridad:** “Negative”. La clave para la clasificación es el uso de la palabra “but”. La **función desagregada** es “Weakness”, las palabras clave no corresponden a “Correct”.

Ejemplo de clasificación para la Función Hedges

Argumentative Zoning (Teufel, 1999; Teufel and Moens, 2002) attempts to solve this problem by representing the structure of a text using a rhetorically-based schema. We used another technique.

Patrón: CITE attempts to solve

Anotación: <cite id="number" function="hed" polarity="neg">Teufel and Moens, 2002</cite> <kw>attempts to solve</kw>

Función: "Hedges"; **polaridad:** "Negative". En este último caso, al usar la secuencia de palabras "attempts to solve" se podría indicar que *intentaron* resolver un problema, pero no lo consiguieron, el uso del verbo "attempt" muestra velada negatividad.

5.4 Procesamiento automático para valoración de impacto

Una vez realizado el etiquetado, se procesan los artículos, para definir información relacionada con impacto: el número de veces que las citas fueron nombradas; las secciones en las cuales se las mencionó: introducción, método, resultados, discusión o sus equivalentes; y el número de secciones en las que aparecen. Estos datos serán parte de la entrada para definir el tipo de impacto de cada cita en el documento.

Como se explicó en el capítulo 4, las valoraciones del impacto se obtienen de forma automática con un proceso que toma en cuenta los criterios obtenidos anteriormente para cada referencia a las citas.

5.5 Conclusiones del capítulo

Anotar manualmente un corpus para análisis de referencias bibliográficas es una tarea complicada que, hasta el momento, no había tenido resultados suficientemente satisfactorios que aseguren la calidad del corpus en cuanto a confiabilidad y reproducibilidad. Por la complejidad de la información semántica que hay que tomar en cuenta, si la anotación manual resulta compleja, muy probablemente, la anotación automática tendrá salidas no muy limpias con las técnicas disponibles actualmente.

Las decisiones de los anotadores tienen que ver con los modelos mentales que se construyen a partir de la información de que disponen, estos modelos mentales tienden a ser cambiantes, a no coincidir con los modelos del esquema y a diferir entre los distintos codificadores. En el proceso de anotación es muy difícil conseguir un buen acuerdo entre

anotadores aún con un entrenamiento adecuado y una guía completa para la tarea; o, incluso, sin que importe si el esquema es sencillo o no (Ciancarini, et al. 2014). Esto sucede porque los modelos mentales que se forman los humanos en el proceso de anotación, son muy cambiantes y diversos. Este es un problema grave que limita la calidad de los resultados del proceso de anotación manual y complica la generación de corpus anotados que puedan ser usados para realizar avances en el campo del análisis de contexto y clasificación de referencias bibliográficas.

Para poder continuar con nuestro trabajo, tuvimos que presentar una solución a este problema diseñando una estrategia que consiste en la integración al proceso de etiquetado de un procedimiento de pre-anotación de patrones. Los patrones están formados por grupos de palabras clave en forma de secuencias simples de palabras o n-gramas y etiquetas que dan información sobre la estructura del texto relevante a las citas. Con la información obtenida al etiquetar estos patrones, los anotadores construyen su modelo mental que, de este modo, tiene una mayor probabilidad de corresponder al modelo del esquema y facilita su decisión respecto a la clasificación de función y polaridad de la cita. Este procedimiento requiere un trabajo adicional en el etiquetado, pero tiene como ventaja que involucra el marcado semántico del texto alrededor de la cita, información valiosa que puede ser usada en muchas aplicaciones de análisis del contenido del contexto de referencias bibliográficas.

Los patrones que proponemos tienen una parte fija que corresponden a secuencias de palabras consecutivas o n-gramas con una longitud máxima de cinco palabras, y una parte variable que corresponde a etiquetas definidas y depuradas en las primeras etapas de nuestra experimentación.

Para realizar la anotación, primeramente se pre-procesa al corpus y se convierte a un formato XML, se etiqueta automáticamente para señalar: título, autores, secciones, párrafos; y, para detectar y enumerar las citas. Se realiza un marcado del contexto dentro del párrafo y se identifica la sección en la que se encuentra la cita viendo la correspondencia a un componente de la estructura IMRaD para artículos científicos. Se realiza la pre-anotación de los patrones para obtener los modelos mentales. Y finalmente, se clasifican polaridad y función de la cita. Se presentaron algunos ejemplos ilustrativos de cada parte del proceso de anotado del corpus.

En resumen, el proceso de pre-anotación es una contribución de nuestro trabajo. En los capítulos siguientes demostremos que constituye una metodología que resulta muy útil, pues involucra que los anotadores construyan un modelo mental sobre la estructura

del texto alrededor de la cita, lo que les permite llegar a conclusiones compartidas y les facilita clasificar en forma coherente la función y la polaridad de las citas.

Los patrones anotados tienen dos usos adicionales. Primero, demostrarán que también son muy informativos como características de entrada para algoritmos de clasificación automática, y así los usaremos en el Capítulo 7; y, segundo, planteamos como trabajo futuro que los patrones podrían ser aprovechados para modelado semántico de los contextos para la automatización del proceso de anotación de un cuerpo de datos para análisis de contexto de citas.

La función y la polaridad junto con otros criterios adicionales que se obtienen a partir del corpus anotado servirán además para desarrollar una función que puede ser usada para valorar el impacto de las citas en el documento que las menciona.



6. Validez del esquema

Un esquema de clasificación codificado manualmente como el que presentamos en este trabajo, necesita demostrar que los datos obtenidos son confiables, reproducibles y precisos.

Un acuerdo entre anotadores suficientemente alto puede asegurar que los resultados serán consistentes a lo largo del tiempo, aún si se tienen diferentes anotadores. Un valor aceptable para el acuerdo entre anotadores certifica que los datos son confiables; y, que diferentes anotadores obtengan un etiquetado muy parecido trabajando en forma independiente, asegura reproducibilidad. Que los resultados que genera la anotación sean adecuadamente parecidos a un corpus gold-estándar validado (si es que existe) tiene que ver con la precisión (Arstein y Poesio, 2008).

Por otro lado no basta que un corpus anotado sea confiable, reproducible y preciso, también es necesario que sirva para realizar clasificaciones de forma automática. La demostración de que nuestro corpus tiene esa utilidad la realizaremos en el siguiente capítulo. La evaluación de la confiabilidad y reproducibilidad se realizará en este capítulo. No existe un corpus gold-standard para comparar nuestras anotaciones y definir su exactitud de acuerdo a la definición expresada en el párrafo anterior, por lo que se usará una validación con un set de pruebas codificado manualmente.

6.1 Acuerdo entre anotadores

Como ya se comentó, un requerimiento indispensable para respaldar la calidad de un modelo como el discutido, requiere demostrar la confiabilidad y la reproducibilidad de los datos obtenidos. Estos criterios se relacionan con la confiabilidad del proceso de anotación. Para medir esta confiabilidad con varios anotadores, es necesario medir el acuerdo obtenido en una pequeña muestra del corpus que se comparte y se revisa por los mismos anotadores. A partir de esos resultados, se evalúa la confiabilidad, es decir, si se podrán generalizar los resultados obtenidos en esta muestra a todo el proceso, en el que probablemente van a intervenir nuevos anotadores y no solo los que codificaron la muestra (Artstein y Poesio, 2008).

Según Krippendorff (2004) la confiabilidad y reproducibilidad de la anotación se asegura cumpliendo tres requerimientos: un esquema claro, instrucciones detalladas y criterios específicos para escoger anotadores. Para medir reproducibilidad se deben tener al menos tres anotadores que deben trabajar independientemente entre sí.

En nuestro experimento se cumplieron estos requerimientos. Se propuso un esquema claro, detallado y con suficientes ejemplos de aplicación; los anotadores son personas que lo han revisado cuidadosamente y tienen conocimientos del área de lingüística computacional; y, por último, para anotar la muestra las tres personas trabajaron en forma separada. Se pidió a los anotadores que sigan en forma consistente un procedimiento claramente establecido. En el primer experimento no se realiza pre-anotación con patrones, y en el segundo se efectúa primeramente una pre-anotación para obtener un modelo mental con patrones, de la forma como se explicó anteriormente.

De esta manera evaluaremos el efecto de la pre-anotación con patrones para verificar si con ella se logra una mejora en el acuerdo entre anotadores.

6.2 Organización de los experimentos y datos

Se usaron artículos del archivo de la Asociación de Lingüística Computacional (ACL por sus siglas en inglés), escogidos de forma aleatoria. Los textos se pre-procesaron para marcar párrafos dentro de los cuales las personas detectan el contexto. Para validar el modelo y

calcular el acuerdo entre anotadores se realizó el etiquetado con tres personas trabajando en forma independiente; estos codificadores recibieron un entrenamiento previo y se les proporcionó la documentación guía para la anotación.

El proceso sin pre-anotación de patrones se lo hizo sobre 101 citas, una variable para función y una variable para polaridad. El proceso con pre-anotación se hizo sobre 108 citas, una variable para función y una variable para polaridad.

El corpus que se entregó a los anotadores contenía el etiquetado base con títulos, autores, secciones, párrafos delimitados, citas detectadas y enumeradas.

Para el experimento sin pre-anotación de patrones, se pidió a los anotadores que lean cuidadosamente cada párrafo que contiene citas y que directamente clasifiquen su función y polaridad.

En el experimento con pre-anotación de patrones se pide que los anotadores, igualmente lean cuidadosamente el texto dentro de cada párrafo con citas, pero que antes de clasificar, marquen las etiquetas y las secuencias de palabras que consideren relacionadas con la función y la polaridad.

Para valorar el acuerdo, las anotaciones de cada codificador se extraen del archivo XML y se guardan en archivos de texto separados, cuya primera línea corresponde a los nombres de las variables, es decir la función, la polaridad y el impacto, separados por una tabulación; y, las siguientes líneas son los resultados de las anotaciones para cada cita. Estos archivos se cargan en el programa desarrollado por Geertzen, J²⁴, 2012, para calcular el nivel de acuerdo entre anotadores.

6.3 Comparación de resultados con y sin pre-anotación de patrones

Artstein y Poesio, 2008 dicen que los datos son confiables si se muestra que los anotadores entendieron las categorías asignadas y por tanto producen en forma consistente resultados similares. De este modo, la confiabilidad es un requisito para demostrar la validez de un esquema. Si los codificadores no muestran acuerdo entre sí, puede deberse a que algunos de ellos están equivocados, que el esquema de anotación no

24 <https://mlnl.net/jg/software/ira/>

es apropiado para los datos o que lo que se pretende anotar tiene una complejidad y ambigüedad tan alta que no permite realizarlo con los métodos tradicionales.

Adicionalmente, la confiabilidad implica la habilidad de distinguir entre las clases de acuerdo a las especificaciones del esquema de clasificación.

Con la evaluación del acuerdo entre anotadores, sin usar patrones y usándolos, vamos a definir si es correcta nuestra hipótesis de que realizando una pre-anotación con patrones, con una guía clara de lo que se quiere detectar, se facilita la construcción de modelos mentales similares entre los anotadores, lo que les lleva a emitir clasificaciones afines, que además coinciden con los modelos mentales de quienes plantearon el esquema de clasificación. En definitiva, se va a comprobar si detectar patrones relacionados con las funciones y la polaridad sirve para aclarar el contexto y clasificar las citas conforme a la intención del esquema, lo que redunda en una coincidencia entre anotadores.

El acuerdo entre anotadores puede evaluarse utilizando coeficientes que tienen en cuenta la corrección de la probabilidad de que los codificadores estén de acuerdo en un ítem simplemente por azar. Arstein y Poesio (2008) sugieren que los coeficientes se escojan de acuerdo a la tarea. En el caso de que la anotación tenga un sesgo hacia algunas categorías, de acuerdo a la recomendación de los autores mencionados, para la evaluación de confiabilidad se debe usar el coeficiente de Krippendorff. Con fines de comparación, obtenemos valores para varios coeficientes incluido el mencionado. Se analizaron las citas para clasificar su función y polaridad, sin pre-anotación y otras diferentes citas con pre-anotación. Se calculó el acuerdo entre anotadores en cada caso.

Arstein y Poesio (2008), resumen el cálculo del acuerdo entre anotadores utilizando los coeficientes más conocidos: π (Scott 1955), κ kappa (Cohen 1960) y α (Krippendorff 2004). Los dos primeros coeficientes parten de la fórmula:

$$\pi, \kappa = \frac{Ao - Ae}{1 - Ae} \quad \text{Ecuación 1}$$

Donde Ao es el acuerdo observado, es decir la proporción de clasificaciones en la que dos codificadores están de acuerdo, calculada como el promedio de la sumatoria de los acuerdos entre dos codificadores, donde 1 refleja coincidencia y 0 una diferencia en la categorización. Ae es el acuerdo por azar, es decir la probabilidad de que dos codificadores estén de acuerdo en una categoría para un ítem en forma aleatoria. El valor $Ao - Ae$ se refiere entonces a cuánto acuerdo se logra más allá del azar. El valor $1 - Ae$ mide el acuerdo sobre lo casual que se puede lograr. El radio entre $Ao - Ae$ y $1 - Ae$ indica la proporción de un acuerdo más allá del azar.

En la fórmula anterior, Ao es igual para los dos coeficientes; pero, Ae varía debido a que lo que se asume que lleva al cálculo de la probabilidad de que el codificador i escoja la categoría k para el caso en el que estuviera escogiendo las categorías al azar. Para el cálculo de π se asume que se obtendría la misma distribución para cada codificador bajo el criterio de que la asignación aleatoria de categorías realizada por cualquier codificador se gobierna por la distribución de ítems en el mundo real. Para el cálculo de Ae en κ se supone que se tendría una distribución separada para cada anotador basándose en que cada clasificación reflejará el sesgo individual del anotador.

Este tipo de cálculo con sus variaciones se aplica para dos codificadores. La generalización para varios codificadores aplicada en π , la realiza Fleiss (1971) con el coeficiente multi- π que asume que la probabilidad de que dos codificadores arbitrarios asignen un ítem a una categoría particular $k \in K$ se toma como la probabilidad conjunta de que cada codificador realice esta asignación en forma independiente. El acuerdo esperado es la suma de estas probabilidades conjuntas.

$$\check{P} = \frac{1}{i_c} n_k \quad \text{Ecuación 2}$$

$$A_e^\pi = \sum_{k \in K} (\check{P}(k))^2 \quad \text{Ecuación 3}$$

$$A_e^\pi = \sum_{k \in K} \left(\frac{1}{i_c} n_k \right)^2 \quad \text{Ecuación 4}$$

$$A_e^\pi = \frac{1}{(i_c)^2} \sum_{k \in K} n_k^2 \quad \text{Ecuación 5}$$

Donde i = número de ítems, C = número de codificadores, n_k = total número de asignaciones realizadas por todos los codificadores, K = número de categorías.

La generalización para kappa puede adaptarse de esta propuesta de Fleiss. Artstein y Poesio 2008 declaran que a menudo la generalización de kappa para múltiples anotadores se confunde con la propuesta de multi- π de Fleiss, sin embargo, la mayoría de autores denominan a este coeficiente con el término Kappa de Fleiss y nosotros también acogeremos esta denominación.

Krippendorff α se basa en una fórmula que se expresa en términos de desacuerdos. Este coeficiente sirve para múltiples codificadores que tienen diferente magnitud en sus desacuerdos y toma en cuenta pesos para calificarlos porque, por ejemplo, no es lo mismo una diferencia entre “contrast” negativo y “weakness” que entre “based on, supply” y “weakness”. El desacuerdo entre este último par debería tener un peso mayor porque la

separación entre las clases es mayor. El desacuerdo observado, el esperado y α se definen con las siguientes fórmulas:

$$D_o^\alpha = 2s_{within}^2 = \frac{1}{ic(c-1)} \sum_{i \in l} \sum_{j=1}^k \sum_{l=1}^k n_{ik_j} n_{ik_l} d_{kjkl} \quad \text{Ecuación 6}$$

$$D_e^\alpha = 2s_{total}^2 = \frac{1}{ic(ic-1)} \sum_{j=1}^k \sum_{l=1}^k n_{kj} n_{kl} d_{kjkl} \quad \text{Ecuación 7}$$

$$\alpha = 1 - \frac{D_o}{D_e} \quad \text{Ecuación 8}$$

Donde cada valor usado por los codificadores pertenece a una categoría $k \in K$, $n_i k$ es el número de veces que a un ítem i se le asigna un valor k , esto es el número de codificadores que hacen ese juicio. Por cada par ordenado de valores $k_a, k_b \in K$ hay $n_i k_a$ $n_i k_b$ par de juicios para el ítem i , en cambio por valores no distintos hay $n_i k_a (n_i k_a - 1)$ pares de juicios.

Se utilizó el software de Geertzen, J. (2012) para el cálculo de los coeficientes. A_obs es el acuerdo observado, A_exp es el acuerdo esperado, D_obs es el desacuerdo observado y D_exp es el desacuerdo esperado.

La Tabla 18 presenta los resultados de los acuerdos entre tres anotadores en la codificación del corpus de prueba sin pre-anotación para la clasificación de la Función, estos resultados son bajos, similares a los reportados para este tipo de tarea por Ciancarini (2014).

Tabla 18: Acuerdo entre anotadores sin pre-anotación correspondiente a la Función

Fleiss	Krippendorff	Pairwise avg.
A_obs=0.554	D_obs = 0.446	% agr = 55.4
A_exp=0.274	D_exp = 0.729	Kappa = 0.386
Kappa=0.386	Alpha = 0.388	

Para la clasificación de la Polaridad los resultados son todavía peores, como se puede apreciar en Tabla 19, aunque el número de clases sea inferior y supuestamente una tarea más sencilla. Por alguna razón que todavía desconocemos, los anotadores tuvieron más dificultad para reconocer la polaridad, es decir, si las citas eran positivas, negativas o

neutrales que la función que realizaba la cita pese a que ésta última se componía de 5 categorías.

Tabla 19: Acuerdo entre anotadores sin pre-anotación correspondiente a la Polaridad

Fleiss	Krippendorff	Pairwise avg.
A_obs=0.6	D_obs = 0.0.4	% agr = 60
A_esp=0.464	D_esp = 0.539	Kappa=0.276
Kappa=0.254	Alpha = 0.259	

Los valores bajos en los coeficientes nos indican que los anotadores tuvieron problemas para distinguir entre categorías y en coincidir entre ellos, esto sucedió sin pre-anotación de patrones. Como se ha comentado, estos resultados coinciden con los que se encuentran en la literatura científica, como lo expresa Ciancarini (2014). Esta situación se repite para esquemas más o menos complejos, pues sucede que es muy difícil obtener un buen acuerdo entre anotadores cuando analizan la categorización de citas bibliográficas de acuerdo a su función. En nuestro experimento, los codificadores leyeron en forma cuidadosa los artículos, pero sin el proceso de pre-anotación los resultados no fueron suficientemente satisfactorios porque se tienen que tomar en cuenta demasiados detalles y aún con una lectura minuciosa, la estructura conceptual del texto es difícil de apreciar.

La Tabla 20 muestran los resultados del acuerdo para la función con los mismos anotadores pero con pre-anotación.

Tabla 20: Acuerdo entre anotadores con pre-anotación correspondiente a la Función

Fleiss	Krippendorff	Pairwise avg.
A_obs=0.911	D_obs = 0.089	% agr = 91.1
A_esp=0.354	D_esp = 0.648	Kappa=0.862
Kappa=0.862	Alpha = 0.862	

Con el proceso de pre-anotación se alcanzó un valor de Kappa de 0,86, el cual se considera una anotación casi perfecta.

Un resultado todavía mejor se consiguió cuando se anotó la polaridad como se observa en la Tabla 21.

Tabla 21: Acuerdo entre anotadores con pre-anotación correspondiente a la Polaridad

Fleiss	Krippendorff	Pairwise avg.
A_obs = 0.98	D_obs = 0.02	% agr = 98
A_exp=0.776	D_exp = 0.225	Kappa=0.913
Kappa=0.912	Alpha = 0.912	

Cuando se pide a los anotadores que etiqueten realizando primero un proceso de pre-anotación, los resultados son excelentes, lo que prueba la validez del esquema y la utilidad del proceso de pre-anotación de patrones. Con este resultado podemos concluir que los patrones realmente sirven para crear un modelo mental que coincide con el esquema propuesto y que permite una anotación coherente cuyos valores comparten los diferentes anotadores. Los valores bajos sin pre-anotación de patrones se pueden explicar por la complejidad del proceso de tratar de inferir la función y la polaridad de una cita analizando el contexto de la misma, incluso detectar el contexto es en sí mismo una tarea que implica un reto.

Hubo una gran mejora en los resultados y los índices de acuerdo entre anotadores fueron altos usando la metodología propuesta que incluye el proceso de pre-anotación con patrones para la creación de un modelo mental claro, compartido entre anotadores y cercano al esquema de categorías planteado. El proceso de extraer y etiquetar conceptos y n-gramas como palabras clave permite que los anotadores distingan más claramente la estructura semántica y facilita la toma de decisión con respecto a la clasificación de función y polaridad. El resultado es una mejora muy significativa en los distintos índices que expresan el acuerdo entre anotadores. Con estos resultados positivos se valida el uso de la propuesta de una metodología de pre-anotación de patrones para la creación de un modelo mental coincidente entre anotadores y coherente con el esquema de clasificación.

Los valores que se obtienen validan esta metodología que puede ser un aporte importante para mejorar los resultados de los procesos de anotación de este tipo de corpus. Cabe anotar que de todos modos es importante que los anotadores pasen por un proceso previo de entrenamiento.

En las Figura 9 y Figura 10 se obtienen los coeficientes: Kappa de Fleiss (Fleiss, 1971), Alpha de Krippendorff (Krippendorff, 2004) y Kappa para el promedio por pares que serían los valores sin tomar en cuenta una corrección del acuerdo por azar.

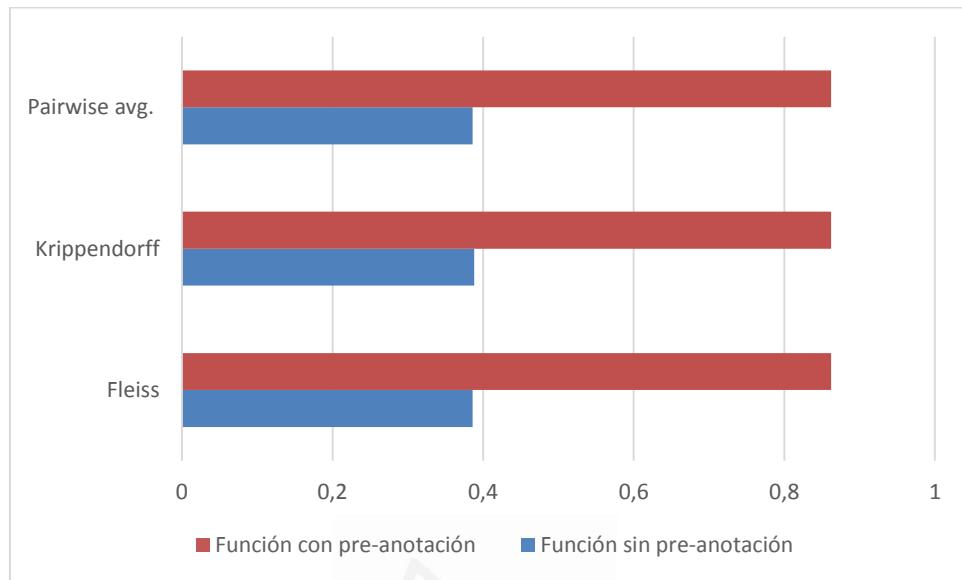


Figura 9 : Comparación de resultados sin y con pre-anotación de funciones

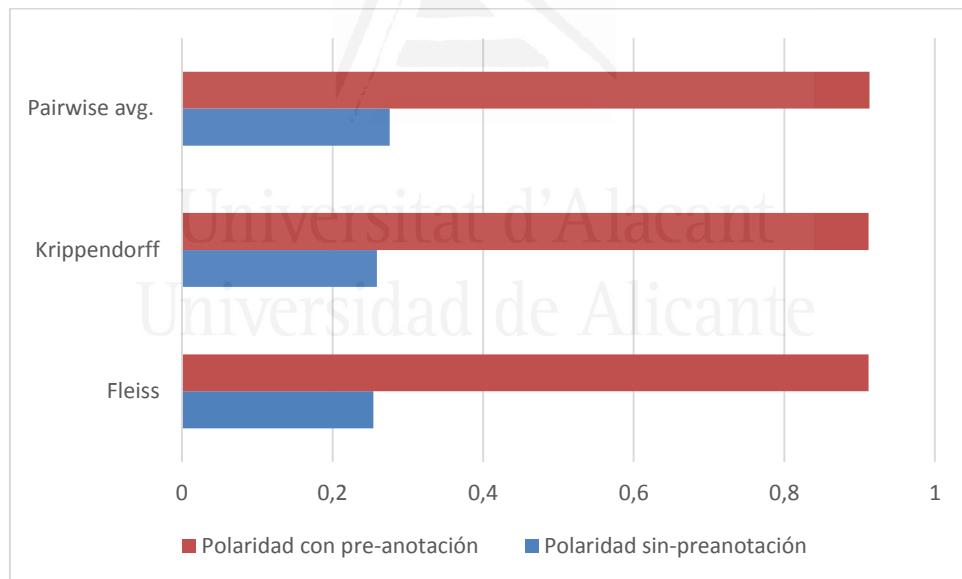


Figura 10: Comparación de resultados sin y con pre-anotación de polaridad de citas

En la clasificación de funciones de las citas, el porcentaje de acuerdo, sin tomar en cuenta una corrección por azar, es de 91,1%. Con la respectiva corrección incluida en el cálculo de Kappa, se tiene un $K = 0,862$. De acuerdo a Landis y Koch (1977), se puede concluir que esta magnitud de Kappa corresponde a un substancial acuerdo entre anotadores, nivel que se considera para Kappa entre 0,6 y 0,8. El acuerdo entre anotadores es incluso mayor para la clasificación de polaridad. La polaridad tiene un $K = 0,912$ y la función un $K = 0,859$. Los mismos autores (Landis y Koch, 1977) califican a los resultados

para Kappa, que van entre 0,8 y 1,0, como perfectos, estos autores asignan los valores que se muestran en la Tabla 22 como referencia para describir el acuerdo entre anotadores calculado con el índice Kappa.

Tabla 22: Referencia para describir acuerdo entre anotadores según valor de Kappa

Valor de Kappa	Acuerdo entre anotadores
<0	pobre
0,00 – 0,20	leve
0,21 – 0,40	débil
0,41 – 0,60	moderado
0,61 – 0,80	substancial
0,81 – 1,00	casi perfecto

6.4 Conclusiones del capítulo

Un requerimiento indispensable para respaldar y probar un modelo como el discutido, es la demostración de confiabilidad y reproducibilidad en los resultados obtenidos. La confiabilidad de los datos tiene que ver con la confiabilidad del proceso de anotación. Para medir la confiabilidad en la anotación del esquema con varios codificadores, es necesario medir el acuerdo obtenido en una pequeña muestra del corpus que ha sido revisado por las mismas personas. De esta manera, el modelo del esquema se valida a través del resultado que lo califica como reproducible.

En nuestros experimentos, realizamos primero el proceso de anotado sin pre-anotación y obtuvimos resultados bajos, parecidos a los reportados en la literatura científica para esta tarea (Hernández y Gómez, 2014). Teufel, 2006 reporta resultados similares en promedio, solo un poco más altos en ciertos casos en los que se diferencian ciertas clases que resultaron más claras para los anotadores, pero el resto de sus resultados fueron similares a los nuestros sin pre-anotación.

Estos valores bajos de acuerdo entre anotadores, sin pre-anotación se pueden explicar debido a la complejidad de la tarea de anotación de funciones aún con un esquema sencillo

compuesto por cinco funciones y tres tipos de polaridad. Se pidió que los anotadores se tomen su tiempo para leer cuidadosamente los textos pero, sin el proceso de pre-anotación, los resultados fueron pobres porque los codificadores tenían que tomar en cuenta demasiados detalles y, generalmente, la estructura que los autores usan en los textos resulta difícil de revelar. Más aún cada anotador tiende a formarse un modelo conceptual diferente de la estructura del texto, modelo que no siempre es coherente con el que tenían en mente los diseñadores del esquema, ni tampoco coincide con el de los otros codificadores. El resultado es que el acuerdo entre anotadores es bajo.

Posteriormente, realizamos el proceso de pre-anotación para extraer conceptos semánticos y palabras clave que permitan ver más claramente la estructura de las oraciones y facilitar la toma de decisiones al momento de realizar la clasificación. Esto es posible con la formación de modelos mentales guiados por lo que se está buscando para realizar la clasificación, a través del ejercicio de distinguir el contexto relevante y, dentro de él, los patrones significativos para la categorización.

El resultado de aplicar este proceso de pre-anotación es una mejora muy significativa del acuerdo entre anotadores, lo que valida el uso de la metodología propuesta y permite continuar con el trabajo de generación de un corpus base. Los valores de Kappa de 0,862 para función y 0,912 para polaridad, corresponde a acuerdos prácticamente perfectos, con mayor razón si se considera que el esquema tiene cinco funciones y tres polaridades, que al combinarse proporcionan una alta granularidad. Estos valores constituyen una mejora respecto a resultados iniciales que ya eran favorables, reportados en Hernández y Gómez (2015). La mejora fue alcanzada con un proceso minucioso de capacitación y entrenamiento a los anotadores, condición indispensable para la aplicación de cualquier esquema de clasificación y, por supuesto, con la aplicación de la metodología de pre-anotación. En nuestro trabajo este proceso inicial de capacitación y entrenamiento a los codificadores, tomó alrededor de 20 horas.

Ciancarini et al., 2014 menciona que el principal problema de la anotación por agentes humanos es que difícilmente concuerdan los modelos mentales que cada uno de ellos se forma al leer un texto, y por lo tanto, los resultados del etiquetado tienden a diferenciarse mucho entre uno y otro anotador. Este obstáculo ha impedido que se realicen anotaciones manuales lo suficientemente reproducibles y confiables que sirvan de base para el desarrollo de un corpus gold-standard para el análisis de citas; de ahí la importancia de la metodología que proponemos, porque con ella se podrá superar este escollo que impide

conseguir buenos acuerdos entre anotadores y la generación de un corpus confiable y reproducible.

A partir de estos resultados positivos, el esquema de clasificación propuesto y la metodología de anotación fueron validados a satisfacción y los resultados se califican como confiables y reproducibles. Para continuar con la validación del esquema planteado, el siguiente punto que queremos demostrar es que, el corpus que se genera aplicando este esquema, se pueden emplear con éxito algoritmos para clasificación automática. Este tópico se trata en el Capítulo 7.



Universitat d'Alacant
Universidad de Alicante

7. Clasificación automática usando el corpus anotado

Un esquema de clasificación y una metodología de anotación deben ser herramientas para generar un corpus base que sirva como punto de partida para estudios en análisis de contexto de citas.

El corpus anotado manualmente con etiquetas, palabras clave, funciones y polaridad de acuerdo a nuestro esquema de clasificación, y siguiendo el proceso de anotación propuesto, debe ser evaluado no solamente desde el punto de vista de la confiabilidad en su codificación, sino también desde el enfoque de su aplicación para clasificar con niveles de *Precision* (Ecuación 9), *Recall* (Ecuación 10) y *F-Measure* (Ecuación 11) aceptables. Que la definición de patrones mejora el acuerdo entre anotadores, se comprueba en el capítulo anterior; la segunda afirmación se comprobará en el punto 7.2.

Usando algoritmos para aprendizaje automático, probaremos que los patrones etiquetados sirven no solamente para facilitar la anotación del corpus apoyando a la creación de modelos mentales, sino también para alimentar a un clasificador con las etiquetas y palabras claves como *features* o características de entrada para categorizar las citas. Con esto validaremos que se han elegido las características más idóneas para ser anotadas en el corpus, con miras a implementar una aplicación que clasifique función y polaridad de citas bibliográficas.

7.1 Organización de los experimentos y datos

El corpus que se utiliza para los experimentos se lo obtuvo del archivo de la Association for Computational Linguistics (ACL), escogidos de forma aleatoria a los que se aplicó la metodología explicada en el Capítulo 5, para obtener un corpus anotado.

Las anotaciones de cada codificador se guardaron en formato XML y se procesaron con un programa en JAVA para recuperar los fragmentos con citas y sus respectivos patrones y palabras clave. Esta información se guardó en una base de datos relacional MySQL que, además, mantiene la información referente al orden en el que aparecen los patrones y palabras claves.

A partir de las anotaciones almacenadas se creó un corpus anotado que sirve como entrada al clasificador WEKA. El corpus tiene una entrada por cada cita y por cada contexto. Cada una de estas muestras tendrá como características los patrones y palabras claves asociadas a las referencias. En caso de tener varias citas en el mismo contexto, se repiten los datos por cita para que se constituya en una entrada separada.

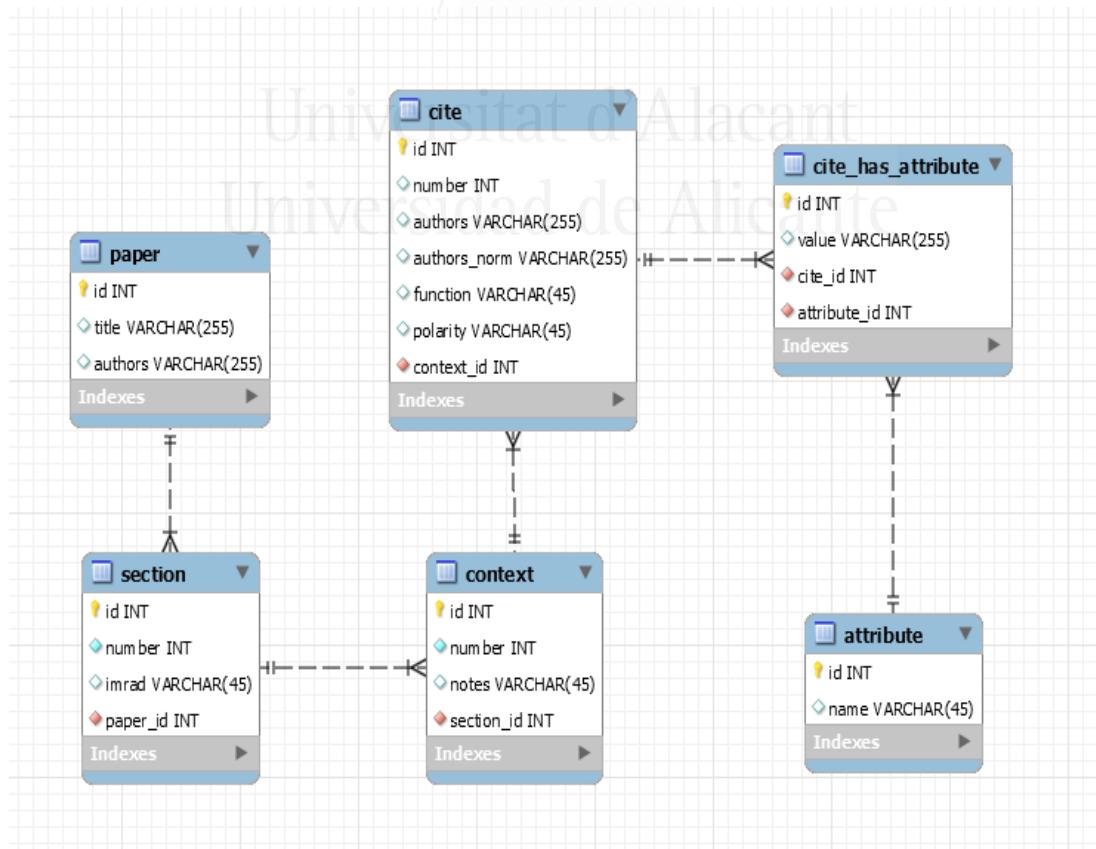


Figura 11: Esquema de la Base de Datos MySQL

El contenido de la base de datos se procesa para obtener archivos tipo .arff que es el formato de entrada para WEKA. Los atributos se refieren a las características que vamos a ingresar y son: cita, función, polaridad, palabras clave, etiquetas. En la carga del archivo .arff, se mantiene el orden en el que se presentan las palabras clave y las etiquetas porque ésta es una información sobre la estructura del contexto que resulta de utilidad a los algoritmos de clasificación que se van a emplear.

7.2 Resultados con SVM y Naïve Bayes

Se hicieron pruebas para distintos algoritmos y, coincidiendo con el estudio inicial del estado de la cuestión presentado en el Capítulo 2, los mejores resultados se obtienen con el algoritmo de Sequential Minimal Optimization (SMO) para entrenar un Support Vector Machine (SVM) (Platt, 1999), usando como datos los patrones y palabras clave y el orden de aparición de cada uno de ellos; también se obtienen resultados satisfactorios con Naïve Bayes (Duda y Hart, 1973).

Para probar que el corpus etiquetado con los patrones presenta condiciones para automatizar exitosamente la clasificación de función y polaridad, y lograr buenos índices de exactitud, realizamos experimentos en los que se procesaron 2092 citas con 73 atributos, usando WEKA. Los algoritmos probados fueron SVM entrenado con SMO, y el algoritmo de Naïve Bayes que también tiene un buen rendimiento para esta aplicación.

Naïve Bayes asume que los valores de las instancias de cada característica o atributo, en una clase, son estadísticamente independientes de los valores de las otras características de la misma clase. Por supuesto, las palabras no son independientes entre sí, porque pueden presentarse múltiples ocurrencias de ciertas secuencias a lo largo del texto; sin embargo, como veremos, a pesar de que nuestra aplicación tiene características funcionalmente dependientes, este algoritmo proporciona buenos resultados. De acuerdo a Rish (2001, Agosto) la precisión del algoritmo de Naïve Bayes no está correlacionada directamente con el grado de dependencia entre características sino con la cantidad de información sobre la clase que no se pierde debido a la suposición de independencia.

En nuestros experimentos, los clasificadores obtienen los siguientes atributos: *function, polarity, cite, kw, kw1, method, data, action, data1, tool, author, kw2, task, action1, action2, method1, method2, kw3, concept, feature, tool1, paper, experiment, author1, concept1, posfeature, task1, task2, theory, experiment1, experiment2, kw4, concept3, result,*

result1, person, kw5, data2, action3, data3, action4, action5, kw6, method3, kw7, method4, kw8, method5, kw9, kw10, concept2, concept3, action6, action7, contribution, negfeature, posfeature1, field, paper1, negfeature1, field1, feature1, people, author2, posfeature2, author3, posfeature3, negfeature2, task3, task4, results, task5, task6, task7.

Para realizar la conversión a formato .arff de WEKA, las características se van llenando por cada cita en el orden declarado en la sección de atributos. Si hay varias ocurrencias de un mismo atributo, se debe a que en la cita esta característica se encuentra varias veces en el orden que indica el subíndice de la etiqueta; si por ejemplo tenemos task1, task2,..., task7; quiere decir que en la misma cita aparecieron siete etiquetas con nombre <task> y que se presentaron en el orden correspondiente al subíndice. Los atributos y palabras clave del corpus anotado en formato XML, se guardan en una base de datos, que se procesa para convertirla en .arff, en la que los patrones mantienen la estructura del documento original. Los datos se llenan para los 73 atributos con diferentes instancias por ejemplo:

```
@attribute action2 { 'marked up',selected,created,using,update,'points out'}
```

Quiere decir que se tiene la etiqueta <action> que, por su subíndice, aparece al menos dos veces en el corpus anotado, y, que en la segunda ocurrencia sus instancias son las que se muestran entre llaves.

Tomando como ejemplo una cita cualquiera, los atributos están definidos en un orden específico en la sección inicial del .arff; y en la sección de datos cuando alguno de los atributos no está presente en el orden que fueron definidos se llena con un signo de interrogación. Habrá una línea de datos para cada cita. A continuación se muestra un ejemplo:

```
hed,neg,'Kintsch and Van Dijk (1978)',,' is the most ambitious model','Due to the  
failure','Comprehension-based summarization',?,?,?,?,?,?,'other  
'?,?,?,?,'representation','less knowledge-intensive methods','dominate.  
'?,?,?,?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?  
,?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?,'?',?
```

Para evaluar el rendimiento de los algoritmos usamos las métricas más comunes de evaluación: *Precision*, *Recall* y *F-Measure* llamado también *F1 score*. La *Precision* se define como el porcentaje de ítems que están correctamente clasificados (Ecuación 9), *Recall* es el porcentaje de ítems correctos que son seleccionados, es una medida de cuántas clasificaciones exitosas se realizaron y cuántos valores correctos no se detectaron (Ecuación 10); y, *F-Measure* es una medida que compendia las dos anteriores porque es una combinación de ellas usando su media armónica (Ecuación 11). Un clasificador con un

valor alto de *F-Measure*, tendrá un buen funcionamiento con respecto a las otras dos magnitudes y por lo tanto será un clasificador exitoso.

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos+Falsos\ Positivos} \quad \text{Ecuación 9}$$

$$Recall\ (R) = \frac{Verdaderos\ positivos}{Verdaderos\ positivos+Falsos\ negativos} \quad \text{Ecuación 10}$$

$$F - Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad \text{Ecuación 11}$$

Para evaluar los intervalos de confianza y, por tanto, la significancia estadística, se utilizó la Normal Tipificada $N(0,1)$ que viene determinada por la $\hat{p} \pm Z\hat{\sigma}_{\hat{p}}$

Ecuación 12

$$\hat{p} \pm Z\hat{\sigma}_{\hat{p}} \quad \text{Ecuación 12}$$

Donde \hat{p} es la probabilidad de error que en nuestro caso será 1 menos *F-Measure* y $\hat{\sigma}$ se calcula con la siguiente ecuación:

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad \text{Ecuación 13}$$

Siendo n es el número de muestras de cada clase.

Z es un factor que depende del porcentaje de intervalo de confianza que queramos obtener y viene determinada por la siguiente ecuación:

$$Z = P(X \leq x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad \text{Ecuación 14}$$

Utilizando esta fórmula obtenemos que para un intervalo de confianza del 95% (un x de 0,95), el valor de Z será 1,96 que es el valor que hemos utilizado en todas nuestras medidas. Por último hay que destacar que la Normal Tipificada $N(0,1)$ sólo puede ser utilizada cuando $n\hat{p} \geq 5$ y, además, $n(1 - \hat{p}) \geq 5$.

Un método adicional para evaluar el rendimiento de un clasificador es el área bajo la curva ROC (Fawcett, 2004) que es un gráfico de la razón de los verdaderos positivos (TP rate) o *Recall* versus la razón de los falsos positivos. La razón de los falsos positivos (FP rate) o fallo en la clasificación, es la relación de los falsos positivos para la suma de los falsos positivos con los verdaderos negativos como se ve en la Ecuación 15. En la

evaluación del área bajo la curva ROC, una clasificación aleatoria daría un valor de 0,5 para el área bajo la curva.

$$FP\ rate = \frac{Falsos\ positivos}{Falsos\ positivos + Verdaderos\ negativos}$$

Ecuación 15

Para evaluar el rendimiento del clasificador se usará un set de datos separado al de entrenamiento para evitar que se produzca overfitting, es decir para impedir que la clasificación se sintonice o afine de acuerdo a los datos de prueba. En ese caso, el clasificador solamente funcionaría bien para los datos para los que fue entrenado y tendría malos resultados para un set de pruebas independiente. En WEKA hay una opción de validación que usa un porcentaje de las muestras para entrenamiento y otra porción de los datos es separada para pruebas; esta herramienta es la de “Percentage Split”. Para realizar comparaciones, decidimos presentar los resultados para los experimentos con los clasificadores con un valor de 66% para el corpus de entrenamiento y el restante 34% para los datos de prueba, y, también para 90% de muestras de entrenamiento y 10 de prueba para poder comparar.

El número de citas de cada función que tiene el corpus anotado manualmente se observa en la Tabla 23.

Tabla 23: Citas por función

Etiqueta	Número
use	705
wea	123
ack	782
bas	337
con	108
hed	37

En la Figura 12 se puede ver una representación gráfica de esta tabla.

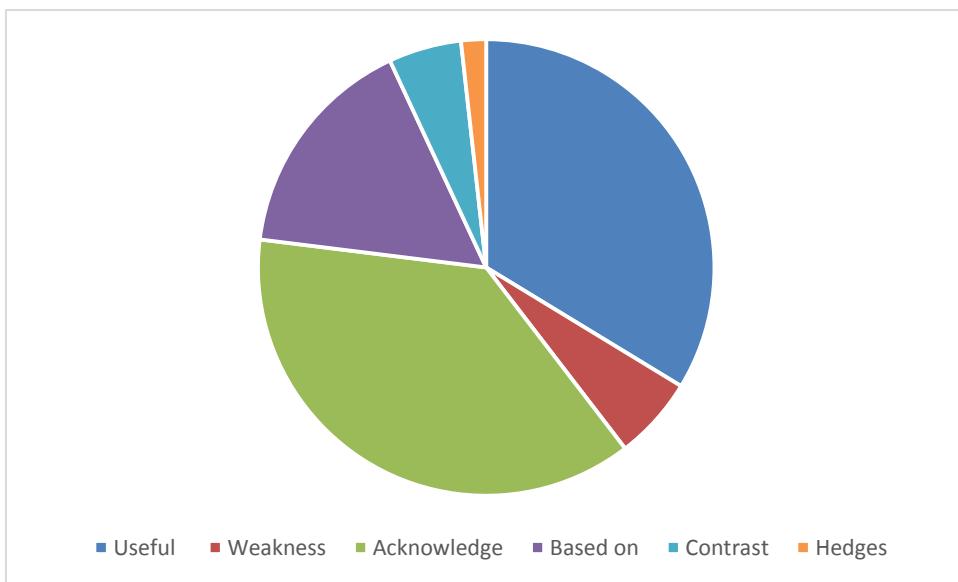


Figura 12: Citas por función

Las clases *Acknowledge* y *Useful* son las categorías más utilizadas y en cierta medida es razonable, pues son del tipo que se utilizan más en la introducción y el estado de la cuestión, donde hay más referencias. Por otra parte, las categorías fundamentalmente negativas como *Weakness* y, sobre todo, *Hedges* son las menos utilizadas. *Contrast* puede ser positiva, negativa o neutral pero la mayoría de las veces es neutral. Esto confirma los resultados de la Tabla 24, en donde la mayoría de las referencias bibliográficas consideradas por los anotadores como neutras ya que las mayorías de las referencias son *Acknowledge* y las que claramente aparecen con menos frecuencia son las negativas. Esto puede indicar que los investigadores en el área del Procesamiento del Lenguaje Natural son reticentes a criticar las obras de otros investigadores o que se suelen citar artículos cuyos resultados o conclusiones no son criticables.

Tabla 24: Citas por polaridad

Etiqueta	Número
Pos	581
Neg	198
Neu	1313

Las distintas proporciones entre el número de citas de las diferentes polaridades se pueden apreciar mejor a través de la gráfica de la Figura 13.

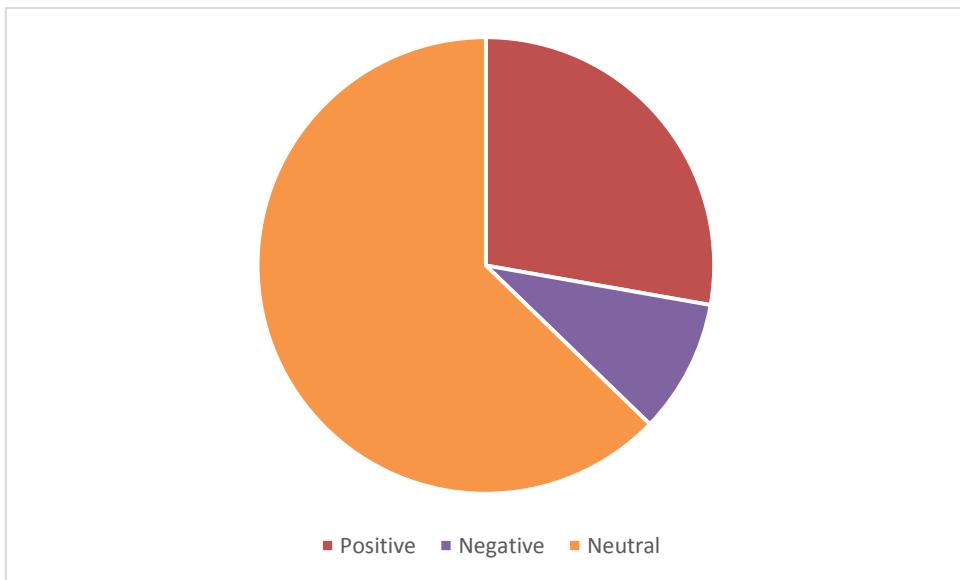


Figura 13: Número de citas por polaridad

En la Tabla 25 se desglosa el número de citas positivas, negativas y neutras por función.

Tabla 25: Número de citas por polaridad y función

	pos	neu	neg
Use	226	479	0
Wea	0	0	123
Ack	62	708	12
Bas	280	57	0
Con	14	69	25
Hed	0	0	37

En esta tabla se observa que la gran mayoría de citas *Acknowledge* son neutras aunque también las hay positivas y en mucha menor medida negativas. Esto tiene mucho sentido si tenemos en cuenta el papel que forma este tipo de citas. Cuando se etiqueta como *Useful*, la proporción de polaridades positivas es mayor aunque siguen habiendo una mayoría de neutras. Como era de esperar, *Based on*, *Supply* es considerado mayoritariamente positivo, pocas veces neutral y nunca se considera negativa y, sin embargo, *Weakness*, *Correct* y *Hedge* reflejan todo lo contrario. Quizás la función más ambigua es la de *Contrast* que tiene un mayor balance entre las distintas polaridades con cierta predisposición a la neutral.

En la Figura 14 se muestran las ocurrencias de cada función junto con la polaridad agregada.

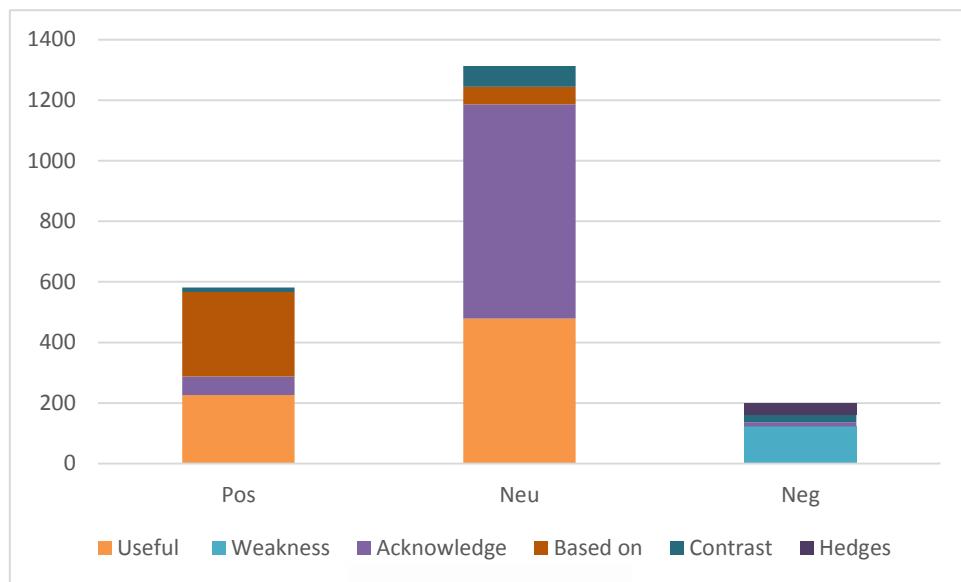


Figura 14: Número de citas por función para cada polaridad

Resultados para la clasificación de función con SVM, 66% de los datos como entrenamiento, 44% set de pruebas

SMO, Kernel used: Linear Kernel: $K(x,y) = \langle x,y \rangle$.

En la Tabla 26 se resumen los resultados para las 711 citas del test de prueba, con 9 no clasificadas para un total de 2092 muestras.

Tabla 26: Resumen de la Evaluación (test Split 66% vs 44%)

Correctly Classified Instances	616	86,6385±0,025
Incorrectly Classified Instances	95	13,3615%±0,025
Kappa statistic	0,8107	
Mean absolute error	0,2265	
Root mean squared error	0,317	
Relative absolute error	95,1902%	
Root relative squared error	92,1599%	
Total Number of Instances	711	
Ignored Class Unknown Instances	9	

Como se aprecia en la tabla, la tasa de acierto es de un $86,6385\% \pm 2,5\%$. Teniendo en cuenta que el sistema debe seleccionar una de las 6 funciones este valor es considerablemente más alto que los alcanzados por cualquier sistema del estado de la cuestión.

En la Tabla 27 se desglosa la tasa de aciertos por tipo de función.

Tabla 27: Precisión detallada por clase

Clase	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
use	0,88	0,089	0,843	0,88	0,861	0,903	$\pm 0,0561$
wea	0,947	0,015	0,783	0,947	0,857	0,993	--
ack	0,909	0,049	0,916	0,909	0,913	0,933	$\pm 0,0418$
bas	0,796	0,032	0,819	0,796	0,808	0,942	$\pm 0,0796$
con	0,61	0,001	0,962	0,61	0,746	0,902	$\pm 0,1706$
hed	0,9	0,003	0,818	0,9	0,857	0,993	--
Weighted Avg.	0,866	0,055	0,87	0,866	0,865	0,926	$\pm 0,0307$

En las tablas los nombres de las funciones y polaridades se abrevian con las tres primeras letras.

Como podemos observar, todas las funciones tienen un *F-Measure* bastante parecido, aproximadamente entre 0,808 y el 0,913, excepto a la función *Contrast* que, debido a su ambigüedad intrínseca que observamos en la Tabla 25, es la más difícil de clasificar porque además tiene pocas muestras. La función que más acertó fue la de *Acknowledge*, con una tasa de acierto de 0,86, seguida por *Useful*. La función *Hedge* también tuvo una tasa de acierto alta, pese a que es una crítica disimulada y se podría pensar a priori que sería una de las más difíciles. Esto es debido a que para realizar un *Hedge* es muy común utilizar ciertas expresiones particulares con un vocabulario que oculta la negatividad o crítica. Estas expresiones han sido eficazmente detectadas por el algoritmo de SVM, gracias a las características semánticas que se anotaron. Por otro lado, los valores para la función *Contrast*, presentan una buena precisión de 0,962, pero un *Recall* más bajo, porque se dejaron de detectar algunas muestras positivas y el número de muestras no es muy alto

Tabla 28: Matriz de confusión

use	wea	ack	bas	con	hed	← clasificado como
131	0	11	4	0	0	use
3	19	0	1	0	0	wea
12	4	154	3	2	0	ack
12	0	13	68	1	0	bas
6	1	4	1	13	0	con
3	1	0	0	0	9	hed

Como se aprecia en la matriz de confusión, los errores cometidos por cada una de las funciones suelen confundirse por otros tipos en proporciones muy parecidas a la de la Tabla 25. Por ejemplo, la función *Hedge* no se anotó nunca como positiva por ningún anotador y el SVM tampoco la ha categorizado con ninguna función intrínsecamente positiva o neutral como *Based on* o *Acknowledge*. Tampoco *Based on* se ha confundido con *Weakness* o *Hedges*.

Resultados para la clasificación de función con SVM, 90% de los datos como entrenamiento, 10% set de pruebas

SMO, Kernel used: Linear Kernel: $K(x,y) = \langle x,y \rangle$.

En este segundo experimento quisimos comprobar si aumentar el número de muestras de entrenamiento a costa de reducir el de test podríamos mejorar los resultados. Por ello en la Tabla 29 se muestran los resultados para las 2092 citas del corpus, con 209 muestras para el conjunto de prueba.

Tabla 29: Resumen de la evaluación (test Split 90% vs 10%)

Correctly Classified Instance	187	89,4737±0,0416
Incorrectly Classified Instance	22	10,5263%±0,0416
Kappa statistics	0,857	
Mean absolute error	0,226	
Root mean square	0,3163	

Relative Absolute error	93,1878%
Root relative squared error	89,847%
Total Number of Instances	209
Ignored Class Unknown Instances	3

Si miramos los resultados por tipo de función de la Tabla 30 podemos apreciar con más detalle el funcionamiento del sistema.

Tabla 30: Precisión detallada por clase

Clase	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
use	0,944	0,087	0,848	0,944	0,893	0,934	±0,0719
wea	0,941	0,005	0,941	0,941	0,941	0,997	--
ack	0,881	0,021	0,952	0,881	0,915	0,92	±0,0668
bas	0,838	0,023	0,886	0,838	0,861	0,962	±0,1115
con	0,8	0	1	0,8	0,889	0,9	--
hed	1	0,01	0,5	1	0,667	0,995	--
Weighted Avg.	0,895	0,041	0,903	0,895	0,896	0,938	±0,0414

Los resultados fueron mejores para la proporción de 90% de muestras de entrenamiento vs. 10% de prueba. Esto se debe a que aún con 10% de muestras de prueba, se tiene un tamaño suficiente de más de 200 citas para analizar.

Todas las clasificaciones de funciones mejoran, excepto la *Precision* para *Hedges* porque una muestra de *Weakness* y otra de *Acknowledge* se clasificaron como *Hedges* y como se tiene un limitado número de muestras, estas dos clasificaciones erróneas inciden en el cálculo de ese índice. En la Tabla 31, se puede observar la matriz de confusión.

Tabla 31: Matriz de confusión

use wea ack bas con hed ← clasificado como						
67	0	1	3	0	0	use
0	16	0	0	0	1	wea

6	1	59	0	0	1	ack
5	0	1	31	0	0	bas
1	0	1	1	12	0	con
0	0	0	0	0	2	hed

Aún con una proporción de muestras de entrenamiento vs. Muestras de prueba de 90 - 10%, el corpus es suficientemente grande como para dar buenos resultados. De ahí que parece que para nuestro tamaño del corpus es más conveniente realizar el "Percentage Split" de WEKA, aplicando la proporción 90 - 10%. De todas maneras, el promedio de *F-Measure* para toda la clasificación sigue siendo alto, con un valor de 0,896.

Resultados para la clasificación de polaridad con SVM, 66% de los datos como entrenamiento, 44% set de pruebas

SMO, Kernel used: Linear Kernel: $K(x,y) = \langle x,y \rangle$.

En la Tabla 32 se muestran los resultados de polaridad para las 716 citas del test de prueba para un modelo entrenado en 1372 muestras más 4 instancias no clasificadas.

Con el nuevo modelo de anotación, y gracias a las características extraídas del corpus generado con él, los resultados para la clasificación automática de la polaridad casi alcanzan el 90%, un resultado excelente para este tipo de tareas.

Tabla 32: Resumen de la evaluación (test Split 66% vs 44%)

Correctly Classified Instances	654	91,3408%±0,0206
Incorrectly Classified Instances	62	8,6592%±0,0206
Kappa statistic	0,823	
Mean absolute error	0,2446	
Root mean squared error	0,3103	
Relative absolute error	71,3677%	
Root relative squared error	76,5132%	
Total Number of Instances	716	
Ignored Class Unknown Instances	4	

Si desglosamos los resultados por función como se muestra en la Tabla 33, observamos que para la *Precision* va de 0.842 a 0.943 para neutras; el *Recall* tiene valores entre 0.884 y 0.937. Se tienen valores similares para *Precision* y *Recall* porque se tienen números parecidos de falsos positivos y falsos negativos como se puede ver en la Matriz de Confusión en la Tabla 34.

Tabla 33: Precisión detallada por clase

Clase	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
pos	0,884	0,056	0,842	0,884	0,863	0,921	±0,0501
neg	0,817	0,008	0,907	0,817	0,86	0,934	±0,0878
neu	0,937	0,112	0,943	0,937	0,94	0,906	±0,0214
Weighted Avg.	0,913	0,089	0,914	0,913	0,914	0,912	±0,0205

En la matriz de confusión vemos que las polaridades positivas se han confundido con neutrales pero nunca con negativas. En las citas negativas se confunden con positivas en sólo 5 muestras pero al ser tan bajo su número poco más podemos deducir. La principal causa de confusión en las neutrales ha sido con las positivas y cinco veces menos con las negativas.

Tabla 34: Matriz de confusión

pos	neg	neu	← clasificado como
160	0	21	pos
5	49	6	neg
25	5	445	neu

Los resultados para la clasificación automática de polaridad son también muy satisfactorios con un promedio para *F-Measure* de 0,914. El mejor valor se obtiene para polaridad neutral, aunque también se reconocen muy bien las clases positiva y negativa. El menor valor de *Recall* se presenta para polaridad negativa puesto que hay un mayor porcentaje de muestras con verdaderos positivos que no son detectadas porque se confunden tanto con positivos como con neutrales.

Resultados para la clasificación de polaridad con SVM, 90% de los datos como entrenamiento, 10% set de pruebas

SMO, Kernel used: Linear Kernel: $K(x,y) = \langle x,y \rangle$.

Si intentamos, como hemos hecho con la evaluación de función, aumentar el número de muestras de entrenamiento a costa de las de pruebas los resultados se muestran en la Tabla 35, en la que se han destinado 1880 citas para el conjunto de entrenamiento y 211 muestras para el de prueba, con una sola instancia ignorada

Tabla 35: Resumen de la evaluación (test Split 90% vs 10%)

Correctly Classified Instances	193	91,4692%±0,0377
Incorrectly Classified Instances	18	8,5308%±0,0377
Kappa statistic	0,8372	
Mean absolute error	0,2433	
Root mean squared error	0,3084	
Relative absolute error	70,1673%	
Root relative squared error	74,0724%	
Total Number of Instances	211	
Ignored Class Unknown Instances	1	

Como se aprecia en la Tabla 36, los resultados para todas las clases mejoran en *Precision* y en *Recall*, a los obtenidos con la anterior partición.

Tabla 36: Precisión detallada por clase

Clase	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
pos	0,907	0,057	0,845	0,907	0,875	0,921	±0,0882
neg	0,87	0,011	0,909	0,87	0,889	0,97	--
neu	0,925	0,091	0,947	0,925	0,936	0,92	±0,0414
Weighted Avg	0,915	0,074	0,916	0,915	0,915	0,926	±0,0376

Y si miramos la matriz de confusión de la Tabla 37 podemos ver que los neutrales tienen mejores resultados porque hay mayor número de muestras que permiten que no haya una incidencia mayor debido a las 10 muestras mal clasificadas.

Tabla 37: Matriz de confusión

← clasificado como			
pos	neg	neu	
49	0	5	pos
1	20	2	neg
8	2	124	neu

Los resultados de este experimento se pueden comparar a los obtenidos para la clasificación de función usando Percentage Split de 90% vs. 10%, porque el tamaño de la muestra es suficiente y con un grupo de datos para entrenamiento mayor, se puede explicar que las salidas sean mejores que para la otra proporción. Para función y polaridad con cualquier relación de partición entre instancias de entrenamiento y prueba, usando algoritmo SVM con SMO, los resultados son mejores que el estado de la cuestión.

En experimentos iniciales, cuando teníamos un menor número de citas analizadas en el corpus, se veía que para tamaños pequeños de la muestra de prueba los resultados desmejoraban considerablemente.

Resultados para la clasificación de función con Naïve Bayes, 66% de los datos como entrenamiento, 44% set de pruebas

Para analizar qué tan sensibles eran las características extraídas del corpus con respecto a variaciones del sistema aprendizaje y dado que muchos investigadores utilizan el algoritmo de Naïve Bayes para clasificar sus resultados, decidimos probar este clasificador. Los resultados para las 711 citas del conjunto de prueba, con 9 instancias no clasificadas, para las 2092 citas del corpus, se pueden ver en la Tabla 38.

Tabla 38: Resumen de la evaluación (test Split 66% vs 44%)

Correctly Classified Instances	528	74,2616%±0,0321
Incorrectly Classified Instances	183	25,7384%±0,0321
Kappa statistic		0,6354

Mean absolute error	0,1281
Root mean squared error	0,2581
Relative absolute error	53,8222%
Root relative squared error	75,0193%
Total Number of Instances	711
Ignored Class Unknown Instances	9

Los resultados para el promedio ponderado de *F-Measure* es un 12,37% inferior si lo comparamos con los obtenidos con SVM. En la Tabla 39 podemos apreciar los resultados desglosados por función.

Tabla 39: Precisión detallada por clase

Class	TP	FP	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
use	0,636	0,069	0,832	0,636	0,721	0,824	±0,0556
wea	0,947	0,013	0,8	0,947	0,867	0,996	±0,108
ack	0,852	0,217	0,699	0,852	0,768	0,843	±0,0509
bas	0,769	0,066	0,675	0,769	0,719	0,945	±0,0848
con	0,415	0,006	0,81	0,415	0,548	0,906	±0,1523
hed	0,8	0,001	0,889	0,8	0,842	0,995	--
Weighted Avg.	0,743	0,116	0,757	0,743	0,738	0,866	±0,0323

En la Tabla 40 podemos observar los resultados de la matriz de confusión.

Tabla 40: Matriz de confusión

use	wea	ack	bas	con	hed	← clasificado como
159	0	67	24	0	0	use
0	36	0	0	2	0	wea
20	4	225	13	2	0	ack
10	0	15	83	0	0	bas

2	3	15	3	17	1	con	
0	2	0	0	0	8	hed	

Los valores para la clasificación de función usando el algoritmo Naïve Bayes son más bajos que los conseguidos con SVM entrenado con SMO. En general, veremos que, como se podía anticipar, SVM con SMO tiene resultados más altos que Naïve Bayes.

Resultados para la clasificación de función con Naïve Bayes, 90% de los datos como entrenamiento, 10% set de pruebas

Para seguir con nuestra metodología de experimentación nos propusimos ver si Naïve Bayes también reducía sus resultados cuando ampliábamos el conjunto de entrenamiento a costa del de pruebas. En la Tabla 41 se muestran los resultados para las 2092 citas del corpus con conjunto de entrenamiento y 209 muestras de prueba con 3 instancias no clasificadas.

Con 10% para muestras de prueba, se tiene todavía un buen tamaño para correr el algoritmo y se produjo incluso una mejora en el *F-Measure* ponderado.

En la Tabla 42 podemos observar los valores de *Precision*, *Recall* y *F-Measure* entre otros para cada una de las funciones.

Tabla 41: Resumen de la evaluación (test Split 90% vs 10%)

Correctly Classified Instances	159	76,0766%±0,0578
Incorrectly Classified Instances	50	23,9234%±0,0578
Kappa statistic	0,674	
Mean absolute error	0,1142	
Root mean squared error	0,2472	
Relative absolute error	47,0922%	
Root relative squared error	70,2149%	
Total Number of Instances	209	
Ignored Class Unknown Instances	3	

Al igual que con el SVM son más sensitivas las clases con menor número de muestras, en este caso la *Precision* de *Hedges* se afecta porque solo se tienen dos muestras con esa clasificación y una de ellas se confundió con *Weakness*. Nosotros diferenciamos las clases *Hedges* y *Weakness* porque su reconocimiento tiene sus especificidades, pero en sí mismas las dos clases son afines porque ambas son esencialmente negativas.

Tabla 42: Precisión detallada por clase

Clase	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
use	0,732	0,605	0,852	0,732	0,788	0,872	$\pm 0,0951$
wea	0,941	0,021	0,8	0,941	0,865	0,997	--
ack	0,806	0,19	0,667	0,806	0,73	0,861	$\pm 0,1063$
bas	0,784	0,052	0,763	0,784	0,773	0,927	$\pm 0,135$
con	0,467	0,005	0,875	0,467	0,609	0,901	$\pm 0,2469$
hed	0,5	0	1	0,5	0,667	0,986	--
Weighted Avg.	0,761	0,094	0,761	0,761	0,759	0,892	$\pm 0,058$

La Tabla 43 muestra la matriz de confusión.

Con una muestra de ejemplo más pequeña, el algoritmo puede tener problemas con clases con menos instancias, que se afectan más fácilmente. Como por ejemplo *Contraste* y *Hedges* con *F-Measure* promedio de alrededor de 0,6. No se recomienda usar este algoritmo para muestras de prueba pequeñas, aún si las de entrenamiento son más grandes.

Tabla 43: Matriz de confusión

use wea ack bas con hed ← clasificado como						
52	0	14	5	0	0	use
0	16	0	0	1	0	wea
8	2	54	3	0	0	ack
1	0	7	29	0	0	bas
0	1	6	1	7	0	con

0	1	0	0	0	1	hed
---	---	---	---	---	---	------------

Resultados para la clasificación de polaridad con Naïve Bayes, 66% de los datos como entrenamiento, 44% set de pruebas

Como se observa en la Tabla 44 donde se muestran los resultados para las 716 citas del conjunto de prueba para un modelo entrenado con 1371 muestras con 5 instancias no clasificadas, los resultados vuelven a reducirse un 6,70% si lo comparamos con el SVM.

Tabla 44: Resumen de la evaluación (test Split 66% vs 44%)

Correctly Classified Instances	606	84,6369%±0,0264
Incorrectly Classified Instances	110	15,3631%±0,0264
Kappa statistic	0,6907	
Mean absolute error	0,1385	
Root mean squared error	0,2726	
Relative absolute error	40,4174%	
Root relative squared error	67,2052%	
Total Number of Instances	716	
Ignored Class Unknown Instances	4	

Si consultamos la Tabla 45 que nos muestra los resultados por cada clase no observamos ningún resultado mejor para ninguno de los factores ni para ninguna de las clases de polaridad, pero los resultados están cercanos entre sí entre SVM y Naïve Bayes.

Tabla 45: Precisión detallada por clase

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
pos	0,779	0,103	0,719	0,779	0,748	0,916	±0,0633
neg	0,85	0,012	0,864	0,85	0,857	0,96	±0,0886
neu	0,872	0,195	0,898	0,872	0,885	0,911	±0,0287
Weighted Avg.	0,846	0,156	0,85	0,846	0,848	0,916	±0,0263

Y para la matriz de confusión mostrada en la Tabla 47 tampoco.

Tabla 46: Matriz de confusión

pos	neg	neu	← clasificado como
141	1	39	pos
1	51	8	neg
54	7	414	neu

Igualmente, con Naïve Bayes para clasificación de polaridad, puede verse que el *F-Measure* promedio ponderado es un poco más bajo que con SVM pero aún es muy satisfactorio con un valor de 0,848.

Resultados para la clasificación de polaridad con Naïve Bayes, 90% de los datos como entrenamiento, 10% set de pruebas

Como último experimento y para completar esta serie, se muestra en la Tabla 47 los resultados para las 2092 citas del corpus, con 211 muestras de prueba, 1 instancia no clasificada y el resto como entrenamiento

Los resultados son mejores que para la proporción 66% vs 44%, todas las polaridades tuvieron un mejor *F-Measure* y el promedio ponderado mejoró de 0.848 a 0,892.

Tabla 47: Resumen de la evaluación (test Split 90% vs 10%)

Correctly Classified Instances	188	89,0995%±0,0421
Incorrectly Classified Instances	23	10,9005%±0,0421
Kappa statistic	0,7949	
Mean absolute error	0,1204	
Root mean squared error	0,2483	
Relative absolute error	34,7391%	
Root relative squared error	59,6248%	
Total Number of Instances	211	
Ignored Class Unknown Instances	1	

En la Tabla 48 se muestran los resultados desglosados por clase.

Tabla 48: Precisión detallada por clase

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Intervalo de confianza
pos	0,87	0,07	0,81	0,87	0,839	0,948	$\pm 0,098$
neg	0,913	0,021	0,84	0,913	0,875	0,974	--
neu	0,896	0,104	0,938	0,896	0,916	0,95	$\pm 0,047$
Weighted Avg.	0,891	0,086	0,894	0,891	0,892	0,952	$\pm 0,0419$

Y en la Tabla 49 la matriz de confusión.

Tabla 49: Matriz de confusión

pos neg neu ← clasificado como			
47	1	6	pos
0	21	2	neg
11	3	120	neu

El uso de este algoritmo se recomienda solamente a partir de cierto tamaño de muestras de prueba. Con el tamaño de nuestro corpus, ya se tienen suficientes muestras como para que funcione bien.

7.3 Análisis de resultados

Mogotsi, et al. (2010) especifica que cuando el set de entrenamiento es suficientemente grande la elección de clasificadores no impacta en forma significativa. Para nuestros experimentos, con el número de instancias que tenemos, se nota una diferencia pequeña entre los resultados obtenidos con ambos clasificadores a favor de SVM con SMO porque tiene un mejor y más estable rendimiento que Naïve Bayes. En las Tabla 50 hasta la Tabla 57, y desde la Figura 15 a la Figura 18, se presentan datos para comparar el rendimiento de los algoritmos con distintas proporciones entre el conjunto de entrenamiento y de

pruebas; y, para cotejar los resultados obtenidos en los experimentos realizados con los algoritmos de SVM con SMO y Naïve Bayes.

Tabla 50: Valores de *F-Measure* para función, en relación a los tamaños de la muestras de entrenamiento y pruebas con SVM

Class	F-Measure 66% vs 44%	F-Measure 90% vs 10%
use	0,861±0,0561	0,893±0,0719
wea	0,857	0,941
ack	0,913±0,0418	0,915±0,0668
bas	0,808±0,0796	0,861±0,1115
con	0,746±0,1706	0,889
hed	0,857	0,667
Weighted Avg.		0,865±0,0307 0,896±0,0414

En la Tabla 51 y la Figura 16 se puede observar que, dado un mismo tamaño de corpus, el rendimiento del clasificador es un poco mejor para los experimentos con particiones de datos de entrenamiento y prueba de 90% vs 10%, con una pequeña variación diferente para una clase con minoría de muestras como es *Hedges*.

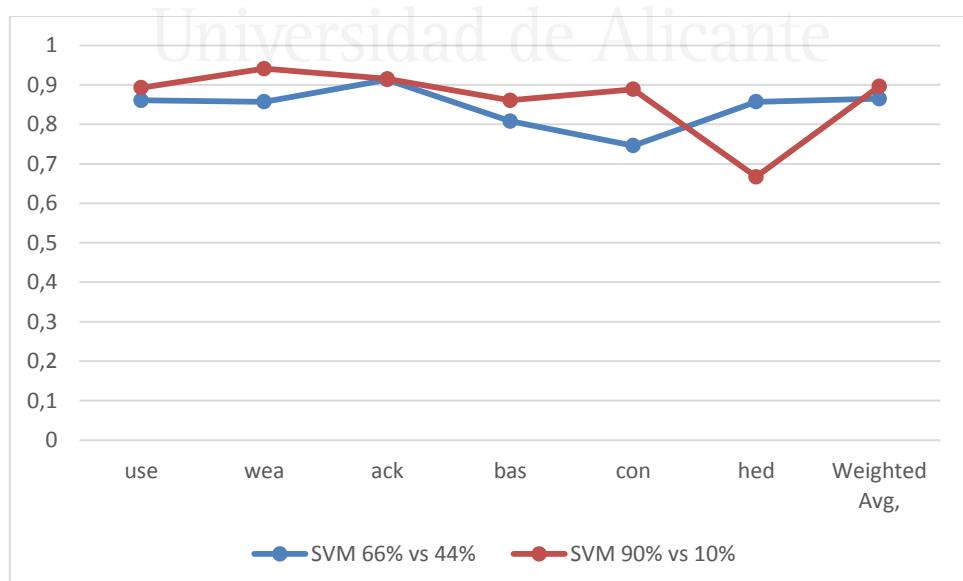


Figura 15: Valores de *F-Measure* para función, en relación a los tamaños de las muestras de entrenamiento y pruebas con SVM

Tabla 51: Valores de *F-Measure* para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con SVM

Class	F-Measure 66% vs 44%	F-Measure 90% vs 10%
pos	0,895±0,0501	0,936±0,0882
neg	0,974±0,0878	1
neu	0,941±0,0214	0,964±0,0414
Weighted Avg.	0,928±0,0205	0,957±0,0376

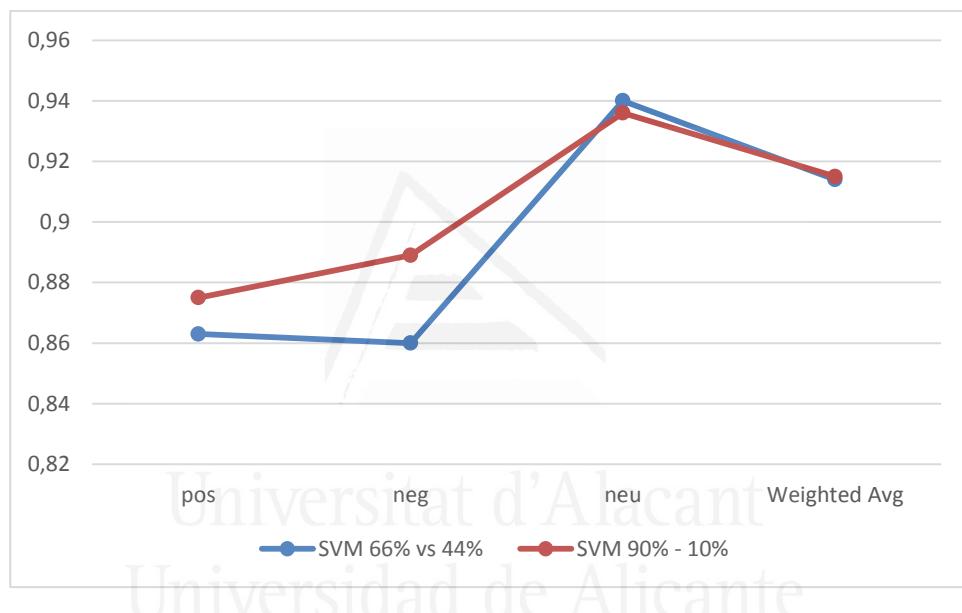


Figura 16: Valores de *F-Measure* para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con SVM

Tabla 52: Valores de *F-Measure* para función, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes

Class	F-Measure 66% vs 44%	F-Measure 90% vs 10%
use	0,721±0,0556	0,788±0,0951
wea	0,867±0,108	0,865
ack	0,768±0,0509	0,73±0,1063
bas	0,719±0,0848	0,773±0,135
con	0,548±0,1523	0,609±0,2469

hed	0,842	0,667
Weighted Avg. 0,738±0,0323 0,759±0,058		

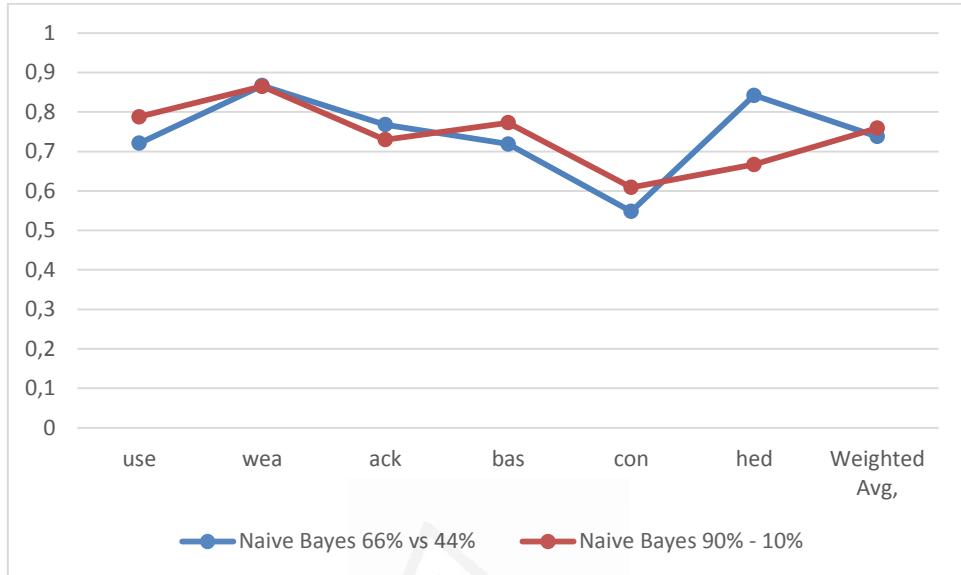


Figura 17: Valores de *F-Measure* para función, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes

Para el algoritmo Naïve Bayes, el tamaño de la muestra de los datos de prueba tiene un efecto que no es significativo, excepto nuevamente para las clases con menor número de muestras como *Hedge*.

Tabla 53: Valores de *F-Measure* para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes

Class	F-Measure	
	66% vs 44%	90% vs 10%
Pos	0,748±0,0633	0,839±0,098
Neg	0,857±0,0886	0,875
Neu	0,885±0,0287	0,916±0,047
Weighted Avg. 0,848±0,0263 0,892±0,0419		

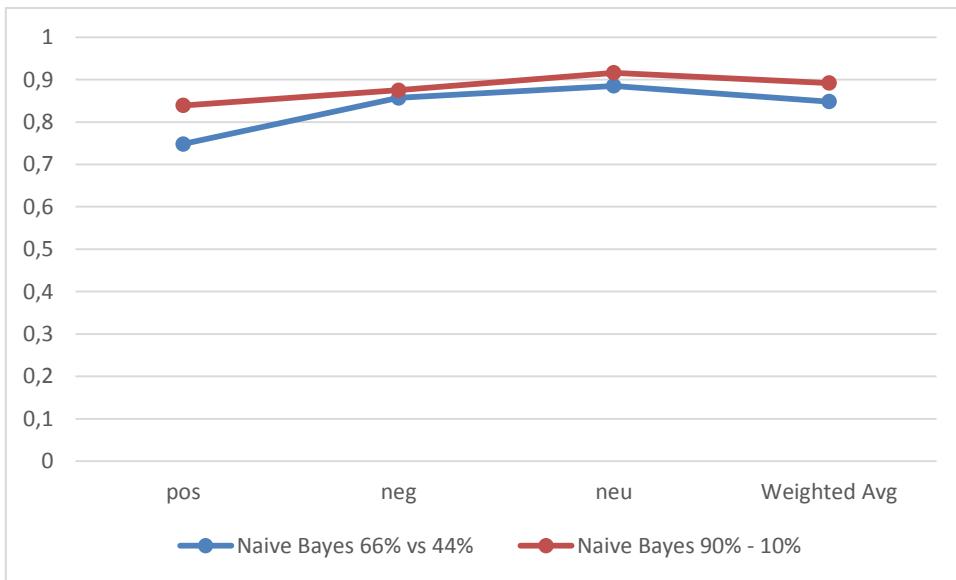


Figura 18: Valores de *F-Measure* para polaridad, en relación a los tamaños de las muestras de entrenamiento y pruebas con Naïve Bayes

En forma consistente se tiene que el rendimiento del clasificador es ligeramente más alto con una proporción entre entrenamiento y prueba de 90% vs 10%.

A continuación comparamos los resultados obtenidos con los dos algoritmos usados para clasificar función y polaridad. SVM con SMO tiene un mejor y más estable rendimiento para cualquier proporción entre el set de datos de entrenamiento y el de prueba. En las Tabla 54 hasta la Tabla 60 y entre la Figura 19 y Figura 22, se puede observar que los valores que se obtienen con SMO son mejores que los que se consiguen con Naïve Bayes tanto para categorización de función como de polaridad. Por otro lado, se consiguen mejores resultados con una proporción del tamaño del set de entrenamiento, con respecto al de prueba de 90% al 10%, excepto para clases con pocas muestras. Un set de prueba pequeño tiende a producir deformaciones en las clases con menor número de muestras porque cualquier error en la clasificación pasa a pesar más estadísticamente.

Tabla 54: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%

Class	SVM F-Measure	Naïve Bayes F-Measure
Use	0,861±0,0561	0,721±0,0556
wea	0,857	0,867±0,108
ack	0,913±0,0418	0,768±0,0509
bas	0,808±0,0796	0,719±0,0848

con	$0,746 \pm 0,1706$	$0,548 \pm 0,1523$
hed	0,857	0,842
Weighted Avg. $0,865 \pm 0,0307$		$0,738 \pm 0,0323$

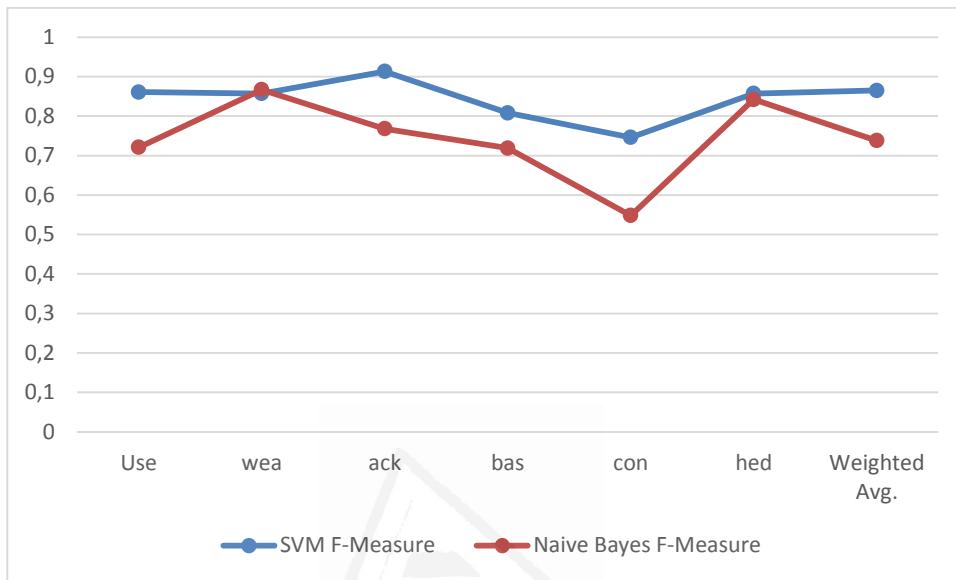


Figura 19: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%

Los resultados para el clasificador que usa el algoritmo SVM son consistentemente iguales o mejores que para Naïve Bayes. Las variaciones para Naïve Bayes son mayores entre las diferentes funciones. Hay puntos de coincidencia para el *F-Measure* de las funciones negativas Weakness y Hedge que, de acuerdo al funcionamiento de Naïve Bayes, puede deberse a que en esos puntos se tengan características independientes de las otras funciones que pueden ser positivas o neutrales. Un comportamiento similar se observa entre SVM y Naïve Bayes para una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%.

Tabla 55: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%

Class	F-Measure con SVM	F-Measure con Naïve Bayes
use	$0,893 \pm 0,0719$	$0,788 \pm 0,0951$
wea	0,941	0,865
ack	$0,915 \pm 0,0668$	$0,73 \pm 0,1063$

bas	0,861±0,1115	0,773±0,135
con	0,889	0,609±0,2469
hed	0,667	0,667
Weighted Avg. 0,896±0,0414		0,759±0,058

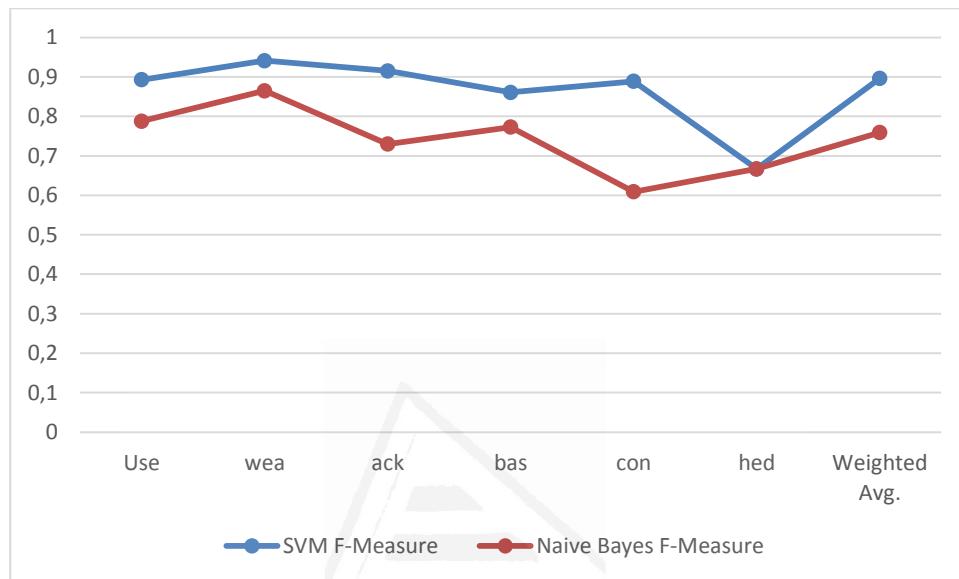


Figura 20: Comparación entre SVM y Naïve Bayes para clasificar función con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%

Tabla 56: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%

Class	F-Measure con SVM	F-Measure con Naïve Bayes
pos	0,863±0,0501	0,748±0,0633
neg	0,86±0,0878	0,857±0,0886
neu	0,94±0,0214	0,885±0,0287
Weighted Avg. 0,914±0,0205		0,848±0,0263

En cuanto a la polaridad los hallazgos son similares, SVM tiene una pequeña ventaja en su rendimiento, excepto para polaridad negativa en la cual los valores convergen. Nuevamente, la razón puede estar en el hecho de que la polaridad negativa tiene características independientes con respecto a las neutrales y positivas, lo que permite un mejor funcionamiento de Naïve Bayes que en ese punto le iguala a SVM (Figura 21). Para

una relación de muestras de 90% a 10%, Naïve Bayes incluso aventaja a SVM en la clasificación de polaridad negativa.

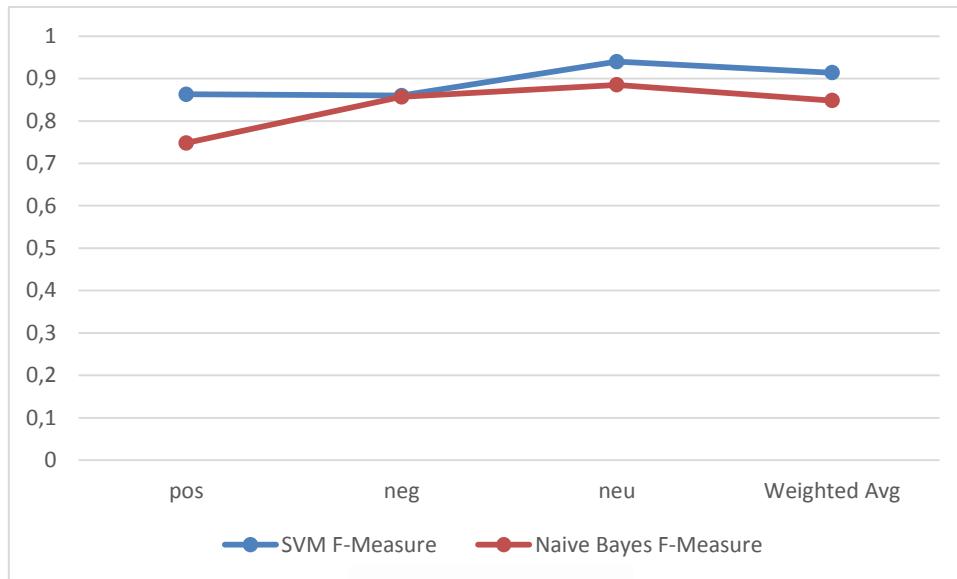


Figura 21: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 66% vs 44%

Tabla 57: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%

Class	F-Measure con SVM	F-Measure con Naïve Bayes
pos	0,875±0,0882	0,839±0,098
neg	0,889	0,875
neu	0,936±0,0414	0,916±0,047
Weighted Avg.	0,915±0,0376	0,892±0,0419

Los resultados son excelentes para el algoritmo SVM en la clasificación de polaridad con una proporción entre muestra de entrenamiento y de prueba del 66% a 44%. Los resultados son un poco más bajos para Naïve Bayes con una mayor diferencia para la función negativa que tiene un menor número de ocurrencias.

Para la clasificación de polaridad se tienen excelentes resultados con los dos algoritmos. Sin embargo, SVM tiene un rendimiento superior sobre Naïve Bayes.

Todos los valores que se obtienen están por encima del estado de la cuestión en cuanto a *Precision*, *Recall* y *F-Measure*. Estos resultados evidencian que las palabras claves, las etiquetas definidas que juntas forman patrones en los que se preserva el orden de

aparición, son cruciales, no sólo como ayuda para el anotador sino también como entrada para los sistemas de aprendizaje automático.

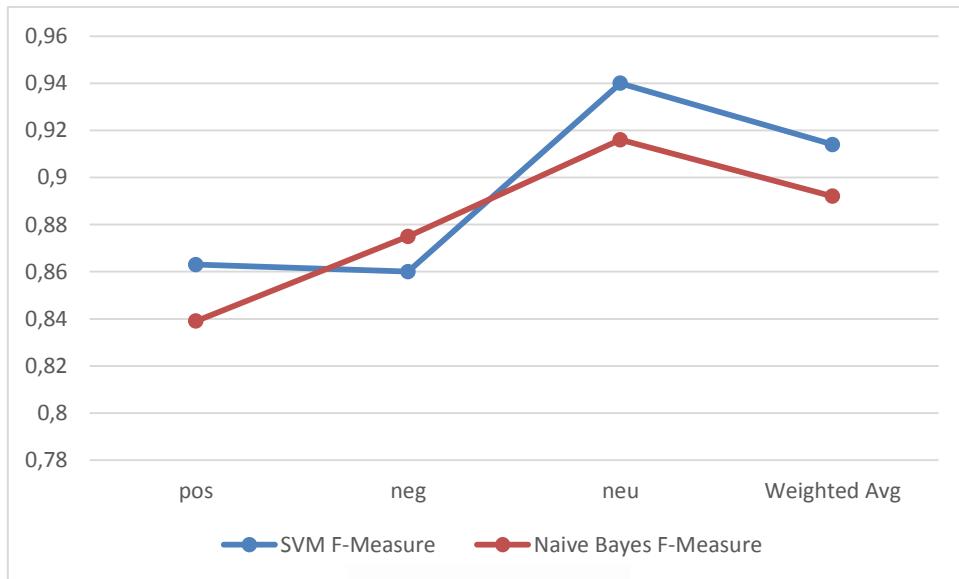


Figura 22: Comparación entre SVM y Naïve Bayes para clasificar polaridad con una relación entre el corpus de entrenamiento y el de prueba de 90% vs 10%

Con estos experimentos, se demuestra que el corpus anotado con las características escogidas puede ser utilizado exitosamente en el análisis de función, polaridad e impacto de citas bibliográficas, lo que sugiere que un corpus como el presentado puede servir como base para futuros sistemas que tengan en cuenta estas características e intenten extraerlas de forma automática para una correcta clasificación.

7.4 Conclusiones del capítulo

En este capítulo se realizaron experimentos para evaluar si los datos anotados en el corpus (etiquetas y palabras clave) sirven como características de entrada a herramientas de aprendizaje automático para realizar clasificación de función y polaridad de citas. Este es un criterio que ratifica la validez del esquema de clasificación y de la elección de las características que se marcaron en el corpus.

El corpus anotado codificado en XML, sirvió como insumo para realizar experimentos con dos algoritmos de aprendizaje automático: SVM entrenado con SMO y Naïve Bayes. Estos algoritmos se escogieron porque, de acuerdo al estudio inicial de la cuestión, presentado en el Capítulo 2, son los que proporcionan los mejores resultados en la

clasificación con ciertas variaciones en el rendimiento que nos permiten realizar comparaciones.

En general para los dos algoritmos, los resultados muestran que las clases que tienen menor cantidad de ocurrencias son las que arrojan rendimientos más bajos en la clasificación.

Se estudia la mejor proporción para la relación entre datos de entrenamiento y prueba para aplicar a nuestro tamaño de corpus. En nuestro caso, los resultados son muy buenos cuando se utiliza la opción de “Percentage Split” de WEKA, aplicando la proporción 66% vs 44%, que corresponde a 933 muestras para entrenamiento y 480 para prueba que alcanza un valor promedio de *F-Measure* de 0,825. En el único caso en que resulta mejor la proporción 90% vs 10%, es decir con 1271 citas para entrenamiento y 142 para prueba, es para clasificación de polaridad con SVM que consigue un valor promedio de *F-Measure* de 0,957.

El algoritmo SVM entrenado con SMO presenta consistentemente mejores resultados que Naïve Bayes por lo que se recomienda su uso para esta aplicación.

Todos los valores que se obtienen están por encima del estado de la cuestión en cuanto a *Precision*, *Recall* y *F-Measure*. Estos resultados evidencian que las palabras claves y las etiquetas proporcionan información que puede ser usada como entrada para los sistemas de aprendizaje automático.

Se demuestra que el corpus anotado con las características escogidas puede ser utilizado exitosamente en el análisis de función, polaridad de citas bibliográficas. Un corpus como el presentado puede servir como base para futuros sistemas que consideren estas características e intenten extraerlas de forma automática.

Con base en estos resultados, proponemos seguir poblando el corpus usando la estructura e información que hemos presentado, para poder conformar un cuerpo de datos sólido que facilite la investigación en el campo del análisis de contexto de citas bibliográficas.

8. Conclusiones y trabajo futuro

We propose to continue populating CONCIT-CORPUS using the structure and given information, to build a solid dataset to be used to facilitate citation analysis research related to function, polarity and impact studies. Analysis of this information will allow improving impact assessment in order to avoid unwanted effects that currently appear with established practices for impact evaluation.

In this work we posed three approaches that use our results to evaluate cited paper impact over citing document. CONCIT-CORPUS is one of the products of our present study and it was designed to serve as support for manual and automated annotation processes. However, it is important to take into account that the techniques for automated semantic annotation need to be enhanced before being successfully applied. At the moment, automated annotation causes too many errors and in best case scenario, it would require strict validation and depuration processes to clean the dataset, so the corpus would hold sufficient quality standards required to be considered useful.

Therefore, as future work emerges the necessity to research in the automated citation corpus annotation field. Mandya (2012) categorize citation function annotation schemes in two categories, first one is for manual classification and second one is for programmed extraction and classification. In the mentioned paper, it is possible to observe that manual classification schemes correspond to medium and high granularity; and, automated classification schemes have low granularity. Corpus annotation using schemes with more than four classes provides valuable information applicable to citation context analysis, but it is a very complex task even for human annotators, this task is very difficult to perform properly when made automatically.

According to the state-of-the-art, introduced in chapter 2, medium and high granularity schemes shown in Table 1, are coded manually by authors; for this kind of schemes the results are poor when an automated annotation is attempted.

Regarding size, corpus delivered in this work is enough for our proposed tasks. In fact, definition of corpus size depends on the applications. Leech (1991) defined that corpus size is not that important per se, optimal extension depends on research questions and practical considerations because datasets that need manual annotation are necessarily small but nevertheless useful for specific usage.

Present work brings new criteria for the development of citation context analysis, with the purpose of applying them in a proposal for a holistic impact evaluation.

Results obtained in this thesis were very satisfactory from various points of view. Our contributions are as follow:

A classification scheme for citation corpus annotation with six grouped functions that combined with three level polarity codification and semantic patterns allow a granularity equal or greater than the one in ontologies as CiTO. Our scheme is simpler, easier to learn and apply for annotators than high granularity schemes as the one mentioned that presents low inter-annotator-agreement (Ciancarini, et al., 2014). Coders can master concepts related with our scheme after few hours of training using a document guide for annotation that is written in detail (Annex 1).

Our proposed annotation methodology requires building of mental models from detection of semantic patterns relevant in the text to define context structure in a standard way. This process produces more informed decisions to classify citation function and polarity more accurately. This novel annotation methodology helps to achieve a very satisfactory rating for inter-annotator agreement, higher than state-of-the-art values. A good inter-annotator-agreement guarantees annotation reproducibility and reliability.

Additionally, patterns and keywords facilitate automated classification because they are input features that provided important information for classification algorithms. With these features we reach very good performance results, higher than current state-of-the-art.

CONCIT-CORPUS was annotated during the development of this work, its size is sufficient for our proposed applications. Also, our goal with this corpus is to provide a tool that enable collaborative work in this field.

Our proposal to enrich current impact assessment methods takes into account function, polarity, citation location, number of occurrences in each paper section, number of sections in which references appear. This information is included in the annotated corpus and allows evaluating closeness between cited documents with citing paper; citation relevance in text as a part of a comprehensive analysis that goes beyond simple citation counting. As future work we suggest to take into account the criteria set forth here to create a holistic impact index that goes beyond simply counting citations.

We suggest to continue annotating papers in a collaborative open environment using our scheme to increase CONCIT-CORPUS. For that, we need online tools and appropriate validation procedures. This is a future work area to develop further citation context analysis.

Another field for future work is the development of automated annotation methods to decrease annotation costs. The condition will be to maintain corpus quality. As we have stated, this is a very complex task considering that even careful manual annotations by experts produce errors. Automated semantic annotation for a citation corpus presents challenges that remain yet unsolved.

CONCIT-CORPUS is made available to citation context analysis researchers in the repository of University of Alicante²⁵.

Universitat d'Alacant
Universidad de Alicante

²⁵ <http://hdl.handle.net/10045/47416>

9. Publicaciones relevantes

Hernández, M., Gómez, J. 2015. Metodología para anotación manual de corpus para clasificación de función y polaridad de citas en bibliografía científica. Simposio Doctoral SEPLN2015. Septiembre 2015.

Hernández-Álvarez, M. & Gómez Soriano, J. (2015). Citation Impact Categorization for Scientific Literature. Submitted to CSE2015.

Hernández-Álvarez, M. & Gómez Soriano, J. (2015). Clasificación automática de función y polaridad de citas bibliográficas usando Análisis de Contexto de Citas. Submitted to *Revista Politécnica*.

Hernández Álvarez, M., & Gómez Soriano, J. (2015). Esquema de anotación para categorización de citas en bibliografía científica. *Procesamiento del Lenguaje Natural*, 54, 45-52.

Hernández, M., Gómez, J. 2014. Survey about Citation Content Analysis: Tasks, Techniques and Resources. Submitted to: *Natural Language Engineering*.

Hernández, M., & Gómez, J. M. (2014). Survey in sentiment, polarity and function analysis of citation. *ACL 2014*, 102. ISBN: 9781634392013.

Hernández, M., & Gómez, J. M. (2014). Sentiment, Polarity and Function Analysis in Bibliometrics: a Review. 2, 43. *NLPCS*. ISBN: 9781501501289.

Hernández, M., & Gómez, J. M. (2014). Análisis de sentimientos aplicado a referencias bibliográficas. *Revista Politécnica*. ISBN: 1390-0129.

Hernández, M., & Gómez, J. M. (2014). Hacia la generación de un corpus para análisis de contenidos de referencias (citas) bibliográficas en literatura científica. Memorias de las V Jornadas JISIC 2014. ISBN: 1390-9266.

Hernández, M., & Gómez, J. M. (2014). An Annotation Scheme for Content Analysis of citations in scientific literature. ACL Anthology. ISBN: 9781941643310.

Hernández, M., Gómez, J. 2013. Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica*. ISSN: 1390-0129.



Universitat d'Alacant
Universidad de Alicante

References

- Abu-Jbara, A., Ezra, J., and Radev, D. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of NAACL-HLT*, Atlanta, GA, pp. 596–606.
- Abu-Jbara, A., and Radev, D. 2012. Reference scope identification in citing sentences. In *12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 80–90. Stroudsburg, PA: Association for Computational Linguistics.
- Angrosh, M. A., Cranfield, S., and Stanger, N. 2013. Conditional random field based sentence context identification: Enhancing citation services for the research community. In *Proceedings of the First Australasian Web Conference*, Vol. 144, pp. 59–68. Adelaide, Australia: Australian Computer Society, Inc.
- Artstein, R., y Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555-596.
- Athar, A. 2011. Sentiment analysis of citations using sentence structure-based features. In *11 Proceedings of the ACL 2011 Student Session*, pp. 81–7. Stroudsburg, PA: Association for Computational Linguistics.
- Athar, A. (2014). Sentiment analysis of scientific citations. Technical Report, University of Cambridge.
- Athar, A., and Teufel, S. 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*, pp. 597–601. Montreal, Canada: Association for Computational Linguistics.
- Biber, D., and Finegan, E. 1994. Intra-textual variation within medical research articles. In N. Oostdjiik and P. DeHaan (eds.), *Corpus-Based Research into Language*, pp. 201-22. Amsterdam: Rodopi.
- Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M. Y., Lee, D., Powley, B., Radev, D. R., and Tan, Y. F. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the 6th*

International Conference on Language Resources and Evaluation Conference (LREC'08), Marrakesh, Morocco, pp. 1755–9.

Blitzer, J., Dredze, M., and Pereira, F. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 440–7.

Boldrini, E., Fernández Martínez, J., Gómez Soriano, J. M., and Martínez Barco, P. 2009. Machine learning techniques for automatic opinion detection in non-traditional textual genres. In *Proceedings of the First Workshop on Opinion Mining and Sentiment Analysis, WOMSA09*, Seville, Spain, pp. 110–9.

Brembs, B., and Munafò, M. 2013. Deep impact: Unintended consequences of journal rank. Digital Libraries; Physics and Society. Available at <http://arxiv.org/abs/1301.3748>.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, 40(4), 284-290.

Chen, M., Xu, Z., Weinberger, K., and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland.

Ciancarini, P., Iorio, A. Di, Nuzzolese, A. G., Peroni, S., and Vitali, F. 2013. Semantic annotation of scholarly documents and citations. In M. Baldoni, C. Baroglio, G. Boella, and R. Micalizio (eds.), *AI*IA 2013: Advances in Artificial Intelligence*, Vol. 8249, pp. 336–47. Berlin: Springer.

Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2014). Evaluating citation functions in CiTO: cognitive issues. In *The Semantic Web: Trends and Challenges* (pp. 580-594). Springer International Publishing.

Davidson, M.J., Dove, L., Weltz, J.: Mental models and usability. Depaul University, Chicago (1999), <http://www.lauradove.info/reports/mental%20models.htm> (last visited April 5, 2015) (retrieved).

Davletov, F., Aydin, A. S., & Cakmak, A. (2014). High Impact Academic Paper Prediction Using Temporal and Topological Features. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 491-498). ACM.

- Dong, C., and Schäfer, U. 2011. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 623–31. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.
- Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis* (Vol. 3). New York: Wiley.
- Fang, F. C., Steen, R. G., and Casadevall, A. 2012. Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America* 109: 17028–33.
- Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, 31, 1-38.
- Fernández, J., Boldrini, E., Gómez, J. M., and Martínez-Barco, P. 2011. Evaluating EmotiBlog robustness for sentiment analysis tasks. In R. Muñoz, A. Montoyo, and E. Métais (eds.), *Natural Language Processing and Information Systems*, pp. 290–4. Heidelberg: Springer-Verlag.
- Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Garfield, E. 1972. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science* 178: 471–9.
- Garzone, M. A. 1997. Automated classification of citations using linguistic semantic grammars. Master's thesis, The University of Western Ontario.
- Garzone, M., and Mercer, R. E. 2000. Towards an automated citation classifier. In *Advances in artificial intelligence* (pp. 337-346). Springer Berlin Heidelberg.
- Green, A., Ashley, K., Litman D., Reed C., and Walker V. 2014. Workshop description, first workshop on argumentation mining at the association for computational linguistics.
- He, Q., Kifer, D., Pei, J., Mitra, P., & Giles, C. L. (2011). Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 755-764). ACM.
- Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science*, 29(6), 308-310.
- Hernández, M., and Gómez, J. M. 2014. Survey in sentiment, polarity and function analysis of citation. In *Proceedings of the First Workshop on Argumentation Mining ACL 2014*, Baltimore, MD, pp. 102–3.

- Hernández Álvarez, M., & Gómez Soriano, J. (2015). Esquema de anotación para categorización de citas en bibliografía científica. *Procesamiento del Lenguaje Natural*, 54, 45-52.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46), 16569-16572.
- Hyland, K. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics* 17: 433-54.
- Hyland, K. 1998. *Hedging in Scientific Research Articles*, Vol. 54. Amsterdam: John Benjamins Publishing.
- Krippendorff, Klaus. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411-433.
- Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2: e124.
- Iorio, A., Di Nuzzolese, A. G., and Peroni, S. 2013. Towards the automatic identification of the nature of citations. In *SePublica*, Montpellier, France, pp. 63-74.
- Jochim, C. 2014. Improving citation polarity classification with product reviews. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, pp. 42-8.
- Jochim, C., and Schütze, H. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of COLING'12* (pp. 1343-58). Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary>
- Kang, I.-S., and Kim, B.-K. 2012. Characteristics of Citation Scopes: A Preliminary Study to Detect Citing Sentences. In *Computer Applications for Database, Education, and Ubiquitous Computing Information Science*, pp. 80-5. Berlin: Springer.
- Kaplan, D., Iida, R., and Tokunaga, T. 2009. Automatic extraction of citation contexts for research paper summarization: a coreference-chain based approach. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pp. 88-95. Suntec, Singapore: Association for Computational Linguistics.
- Kataria, S., Mitra, P., & Bhatia, S. (2010). Utilizing Context in Generative Bayesian Models for Linked Corpus. In *AAAI* (Vol. 10, p. 1).
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, 14(1), 10-25.

- Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 423–30. Stroudsburg, PA: Association for Computational Linguistics.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.
- Leech, G. (1991), The State of the Art in Corpus Linguistics. In: Aijmer, K./Altenberg, B. (eds.), English Corpus Linguistics. London: Longman, 8-29.
- Leech, G. (1993). Corpus annotation schemes. *Literary and linguistic computing*, 8(4), 275-281.
- Li, X., He, Y., Meyers, A., and Grishman, R. 2013. Towards fine-grained citation function classification. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 402–7.
- Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28: 991-1000.
- Livne, A., Gokuladas, V., Teevan, J., Dumais, S. T., & Adar, E. (2014). CiteSight: supporting contextual citation recommendation using differential search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 807-816). ACM.
- MacRoberts, M. H., and MacRoberts, B. R. 1984. The negational reference: or the Art of dissembling. *Social Studies of Science*, 14(1), 91–4.
- Mandya, A. A. (2012). *Enhancing Citation Context based Information Services through Sentence Context Identification* (Doctoral dissertation, University of Otago).
- Marder, E., Kettenmann, H., and Grillner, S. (2010). Impacting our young. *Proceedings of the National Academy of Sciences of the United States of America* 107: 21233.
- McCain, K. W., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, 17(1), 127-163.
- Mei, Q., & Zhai, C. (2008, June). Generating Impact-Based Summaries for Scientific Literature. In *ACL* (Vol. 8, pp. 816-824).
- Mercer, R. E., Di Marco, C., and Kroon, F. W. 2004. The frequency of hedging cues in citation contexts in scientific writing. In *Advances in artificial intelligence*, pp. 75-88. Berlin: Springer Heidelberg.

- Meyers, A. 2013. Contrasting and corroborating citations in journal articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, Hissar, Bulgaria, pp. 460–6.
- Mogotsi, I. C. (2010). Christopher d. manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to information retrieval. *Information Retrieval*, 13(2), 192-195.
- Mullen, T., and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In D. Wu (ed.), *Conference on Empirical Methods in Natural Language Processing*, pp. 412–18. Barcelona, Spain: Association for Computational Linguistics.
- Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 542-550). ACM.
- Nicholson, J. M., and Ioannidis, J. P. A. 2012. Research grants: Conform and be funded. *Nature* 492: 34–6.
- Page, L., Brin, S., Motwani, R., and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab.
- Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04* (p. 271–8). Morristown, NJ: Association for Computational Linguistics.
- Peldszus, A. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining ACL 2014*, Baltimore, MD, pp. 88–97.
- Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods—support vector learning*, 3.
- Prabha, C. G. (1983). Some aspects of citation behavior: A pilot study in business administration. *Journal of the American Society for Information Science*, 34(3), 202-206.
- Prabowo, R., and Thelwall, M. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics* 3: 143–57.
- Qazvinian, V., and Radev, D. R. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on*

Computational Linguistics-Volume 1, pp. 689–96. Stroudsburg, PA: Association for Computational Linguistics.

Qazvinian, V., and Radev, D. R. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 555–64.

Radev, D. R., Muthukrishnan, P., and Qazvinian, V. 2009. The ACL Anthology Network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pp. 54–61. Suntec, Singapore: Association for Computational Linguistics.

Radicchi, F. 2012. In science “there is no bad publicity”: Papers criticized in comments have high scientific impact. *Scientific Reports* 2: 815.

Reyhani Hamedani, M., Kim, S. W., Lee, S. C., & Kim, D. J. (2013, October). On exploiting content and citations together to compute similarity of scientific papers. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 1553-1556). ACM.

Rish, I. (2001, August). An empirical study of the Naïve Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). IBM New York.

Ritchie, A., Teufel, S., and Robertson, S. 2006. How to find better index terms through citations. In *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, Sydney, Australia, pp. 25–32.

Salton, G.; Wong, A.; and Yang, C. S. 1975. A vector space model for automatic indexing. *Commun. ACM* 18(11):613–620.

Sample, I. 2013. Nobel winner declares boycott of top science journals. *The Guardian*. Available at <http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals>

Sayyadi, H., & Getoor, L. (2009). FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In *SDM* (pp. 533-544).

Schreiber, M. 2013. A case study of the arbitrariness of the h-index and the highly-cited-publications indicator. *Journal of Informetrics* 7: 379–87.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34: 1–47.

- Siegel, D., and Baveye, P. 2010. Battling the paper glut. *Science* 329: 1466.
- Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24: 265–9.
- Small, H. 2011. Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics* 87: 373–88.
- Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the medical library association*, 92(3), 364.
- Sugiyama, K., Kumar, T., Kan, M.-Y., and Tripathi, R. C. 2010. Identifying citing sentences in research papers using supervised learning. In *2010 International Conference on Information Retrieval and Knowledge Management (CAMP)*, Shah Alam, Selangor, Malaysia, pp. 67–72.
- Teufel, S. 2000. *Argumentative zoning: Information extraction from scientific text*. Doctoral dissertation, University of Edinburgh.
- Teufel, S., and Moens, M. 1999. Discourse-level argumentation in scientific articles: Human and automatic annotation. In M. Walker (ed.), *Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop*, pp. 84–93. Somerset, NJ: Association for Computational Linguistics.
- Teufel, S., Siddharthan, A., and Tidhar, D. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 103–10. Stroudsburg, PA: Association for Computational Linguistics.
- Teufel, S., Siddharthan, A., and Tidhar, D. 2009. An annotation scheme for citation function. In *Proceedings of t130he 7th SIGdial Workshop on Discourse and Dialogue*, pp. 80–7. Stroudsburg, PA: Association for Computational Linguistics.
- Teufel, S. (2010). The Structure of Scientific Articles: Applications to Citation Indexing and Summarization (Center for the Study of Language and Information-Lecture Notes).
- Tsai, C.-T., Kundu, G., and Roth, D. 2013. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, pp. 1733–1738. New York, New York, USA: ACM Press.
- Van Noorden, R. 2013. Brazilian citation scheme outed. *Nature*, 500(7464), 510–1.

- Verlic, M., Stiglic, G., Kocbek, S., and Kokol, P. 2008. Sentiment in Science - A Case Study of CBMS Contributions in Years 2003 to 2007. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pp. 138–143. IEEE.
- Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., ... Patwardhan, S. 2005. OpinionFinder. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pp. 34–35. Morristown, NJ, USA: Association for Computational Linguistics.
- Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011, October). Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1247-1252). ACM.
- Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. 2008. Why current publication practices may distort science. *PLoS Medicine*, 5(10), e201.
- Zhang, G., Ding, Y., and Milojević, S. 2013. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, 64(7), 1490-1503.
- Zhang, W., Yu, C., and Meng, W. 2007. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*, pp. 831. New York, New York, USA: ACM Press.
- Zhu, X., Turney, P., Lemire, D., and Vellino, A. 2014. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*.
- Ziman, J. M. 1987. *An introduction to science studies: The philosophical and social aspects of science and technology*. Cambridge: Cambridge University Press.

ANEXO 1: Guía de anotación

i. Introducción y Esquema

La guía de anotación es un documento completo e independiente, que sirve para orientar a los etiquetadores del corpus para análisis de referencias bibliográficas.

El corpus tiene como objetivo servir de base para el análisis del propósito y la disposición con que el autor hizo la cita, asociadas con la función de la referencia en el texto y su polaridad positiva, negativa o neutral; y, con el estudio de la evaluación del impacto de esa cita en el artículo, desde el punto de vista de la cercanía de la cita con respecto al desarrollo del trabajo realizado. Las anotaciones del corpus están dirigidas a brindar información que permita realizar esa investigación.

El corpus está conformado por artículos de literatura científica en formato XML, en los que se colocan etiquetas que a su vez pueden tener atributos. La principal etiqueta es “<cite>” en la que se clasifican las citas de acuerdo a su función y polaridad. Por ejemplo: <cite id="1" function="bas" polarity="pos" se refiere a una cita con función “Based on, Supply” y polaridad positiva, tal como están definidas en las El impacto se evalúa en forma automática a partir de elementos en la anotación como la función, la polaridad, la ubicación de la cita en el artículo y el número de veces que es referenciada.

El esquema de clasificación que se muestra en Tabla 58

Tabla 59 y Figura 23: Criterios de clasificación de una cita de acuerdo a su función, polaridad e impacto se aplica en la anotación del corpus, luego de un proceso de pre-anotación que se explica en el punto ii.

En este esquema se juntan categorías de funciones con características similares para disminuir la carga de información que manejan los anotadores y para lograr que las diferencias entre categorías puedan estar más claras y distinguibles. En esta definición las funciones podrán ser separadas combinándolas con la polaridad, tal como se muestra en la Tabla 60, donde se presentan las funciones individuales que resultan de separar los grupos según la polaridad que presenten.

Tabla 58: Esquema de anotación para funciones

Función de la referencia	Descripción
Based on, Supply	El artículo que referencia aplica el trabajo de la cita. El artículo que la referencia se construye a partir del trabajo de la cita (Based on) o el trabajo de la cita es usado como una fuente (Supply).
Useful	El material de la cita (concepto o herramienta) se reconoce como útil y se aplica en algún otro trabajo, no en el propio.
Acknowledge, Corroboration, Debate	La cita se menciona para reconocer algún trabajo previo. El artículo que la referencia puede: simplemente mencionar la cita (Acknowledge); estar de acuerdo con ella (Corroboration); o, discutir, disputar el trabajo de la cita (Debate).
Contrast	La cita se compara con otros trabajos, el resultado es un criterio que puede ser positivo, negativo o neutro.
Weakness, Correct	Se nota un error o debilidad de la cita (Weakness), se corrige un error o debilidad de la cita (Correct).
Hedges	Se usa un lenguaje cuidadoso para ocultar la crítica (Hedges).

Tabla 59: Esquema de anotación para polaridad

Valor	Característica
Negative	No hay relación entre el artículo y la cita, en este caso el autor menciona la referencia con una disposición negativa o crítica.
Perfunctory	Citas triviales, relacionadas solo marginalmente con el artículo que la referencia. Polaridad neutral de la cita.
Significant	Citas importantes para el trabajo que las menciona, que están estrechamente relacionadas con el trabajo que hace la mención. Generalmente están vinculadas a una polaridad positiva hacia la cita.



Figura 23: Criterios de clasificación de una cita de acuerdo a su función, polaridad e impacto

Tabla 60 : Funciones desagrupadas por combinación función-polaridad

Función agrupada	Polaridad	Función desagrupada
Based on, supply	Positive	Based on
	Negative	N/A
Useful	Neutral	Supply
	Positive	Useful
	Neutral	Useful
Acknowledge, Corroboration, Debate	Positive	Corroboration
	Negative	Debate
	Neutral	Acknowledge
Contrast	Positive	Contrast
	Negative	Contrast
	Neutral	Contrast

	Positive	N/A
Weakness, Correct	Negative	Negative
	Neutral	N/A
	Positive	N/A
Hedges	Negative	Negative
	Neutral	N/A

En los puntos iv y v se explicará cómo anotar cada función y polaridad a través de ejemplos y palabras clave asociadas. Las palabras clave son una combinación de hasta cinco palabras que se encuentran relacionadas semánticamente con la función o con la polaridad de la cita. Los anotadores escogerán las palabras clave más características para la clasificación para alimentar en forma significativa al proceso de etiquetado automático.

ii. Procedimiento de anotación

La anotación se realiza usando XML. Para el presente trabajo la anotación se hizo dentro de la herramienta Netbeans por la facilidad de edición y revisión de la validez de la estructura XML. Los artículos han pasado por un pre-procesamiento en el cual se ha definido: título y filiación de los autores, secciones con sus rótulos, párrafos, las citas han sido identificadas y numeradas. La estructura XML ha sido validada.

Ejemplo de una parte de un artículo pre-procesado:

```
<annotatedpaper>
<paper title="Topic-wise, Sentiment-wise, or Otherwise? Identifying the
Hidden Dimension for Unsupervised Text Classification" authors="Sajib
Dasgupta, Vincent Ng" year="2009">
    <section>
        <title>Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden
        Dimension for Unsupervised Text Classification</title>
        Sajib Dasgupta and Vincent Ng Human Language Technology Research
        Institute University of Texas at Dallas Richardson, TX 75083-0688
        {sajib,vince}@hlt.utdallas.edu
    </section>
    <section>
        <title>Abstract</title>
        <paragraph>
            While traditional work on text clustering has largely focused on
            grouping documents by topic, it is conceivable that a user may want to
            cluster documents along other dimensions, such as the author's mood,
            gender, age, or sentiment. Without knowing the user's intention, a
        </paragraph>
    </section>
</paper>
```

clustering algorithm will only group documents along the most prominent dimension, which may not be the one the user desires. To address this problem, we propose a novel way of incorporating user feedback into a clustering algorithm, which allows a user to easily specify the dimension along which she wants the data points to be clustered via inspecting only a small number of words. This distinguishes our method from existing ones, which typically require a large amount of effort on the part of humans in the form of document annotation or interactive construction of the feature space. We demonstrate the viability of our method on several challenging sentiment datasets.

```

</paragraph>
</section>
<section imrad="i">
  <title>1 Introduction</title>
  <paragraph>
    Text clustering is one of the most important applications in Natural Language Processing (NLP). A common approach to this problem consists of (1) computing the similarity between each pair of documents, each of which is typically represented as a bag of words; and (2) using an unsupervised clustering algorithm to partition the documents. The majority of existing work on text clustering has focused on topic-based clustering, where high accuracies can be achieved even for datasets with a large number of classes (e.g., 20 Newsgroups).
  </paragraph>
  <paragraph>
    On the other hand, there has been relatively little work on sentiment-based clustering and the related task of unsupervised polarity classification. Where the goal is to cluster (or classify) a set of documents (e.g., reviews) according to the polarity (e.g., "thumbs up" or "thumbs down") expressed by the author in an unsupervised manner. Despite the large amount of recent work on sentiment analysis and opinion mining, much of it has focused on supervised methods (e.g.,
    <cite id="1" >Pang et al. (2002)</cite>, <cite id="2">Kim and Hovy (2004)</cite>, <cite id="3" >Mullen and Collier (2004)</cite>). One weakness of these existing supervised polarity classification systems is that they are typically domain- and language-specific. Hence, when given a new domain or language, one needs to go through the expensive process of collecting a large amount of annotated data in order to train a high-performance polarity classifier. Some recent attempts have been made to leverage existing sentiment corpora or lexica to automatically create annotated resources for new domains or languages. However, such methods require the existence of either a parallel corpus / machine translation engine for projecting/translating annotations/lexica from a resource-rich language to the target language (<cite id="4">Banea et al, 2008</cite>; <cite id="5">Wan, 2008</cite>), or a domain that is "similar" enough to the target domain (<cite id="6">Blitzer et al, 2007</cite>).</context> When the target domain or language fails to meet this requirement, sentiment-based clustering or unsupervised polarity classification become appealing alternatives. Unfortunately, to our knowledge, these tasks are largely under-investigated in the NLP community. <cite id="7" >Turney's (2002)</cite> work is perhaps one of the most notable examples of unsupervised polarity classification. However, while his system learns the semantic orientation of the phrases in a review in an unsupervised manner, this information is used to predict the polarity of a review heuristically.
  </paragraph>

```

```

    ...
  </section>
  ...
</paper>
</annotatedpaper>
```

Para realizar las anotaciones se recomienda seguir el siguiente procedimiento:

1. Leer todo el párrafo que contiene la referencia.
2. Dentro del párrafo, definir el contexto con el contenido relevante para clasificar cada referencia. Este contexto puede corresponder al párrafo completo o a una parte del mismo. Realizando esta anotación de contexto, en el ejemplo anterior queda:

```

<paragraph>
  On the other hand, there has been relatively little work on sentiment-based
  clustering and the related task of unsupervised polarity classification.
  Where the goal is to cluster (or classify) a set of documents (e.g., reviews)
  according to the polarity (e.g., "thumbs up" or "thumbs down") expressed by
  the author in an unsupervised manner.
  <context>Despite the large amount of recent work on sentiment analysis and
  opinion mining, much of it has focused on supervised methods (e.g., <cite
  id="1">Pang et al. (2002)</cite>, <cite id="2">Kim and Hovy (2004)</cite>,
  <cite id="3">Mullen and Collier (2004)</cite>). One weakness of these
  existing supervised polarity classification systems is that they are
  typically domain- and language-specific.</context>
  Hence, when given a new domain or language, one needs to go through the
  expensive process of collecting a large amount of annotated data in order to
  train a high-performance polarity classifier. Some recent attempts have been
  made to leverage existing sentiment corpora or lexica to automatically create
  annotated resources for new domains or languages.
  <context>However, such methods require the existence of either a parallel
  corpus / machine translation engine for projecting/translating
  annotations/lexica from a resource-rich language to the target language
  (<cite id="4">Banea et al, 2008</cite>; <cite id="5">Wan, 2008</cite>), or a
  domain that is "similar" enough to the target domain (<cite id="6">Blitzer et
  al, 2007</cite>).</context>
  When the target domain or language fails to meet this requirement, sentiment-
  based clustering or unsupervised polarity classification become appealing
  alternatives. Unfortunately, to our knowledge, these tasks are largely under-
  investigated in the NLP community.
  <context> <cite id="7">Turney's (2002)</cite> work is perhaps one of the most
  notable examples of unsupervised polarity classification. However, while his
  system learns the semantic orientation of the phrases in a review in an
  unsupervised manner, this information is used to predict the polarity of a
  review heuristically.</context>
</paragraph>
```

3. Revisar que la sección tenga los atributos que correspondan de acuerdo a la estructura definida en IMRaD. Ejemplo: <section imrad="i"> corresponde a la Introducción de un artículo.

4. Anotar las etiquetas de los patrones que aporten significado a la clasificación. Los patrones contribuyen para que el anotador establezca la estructura de las oraciones en el contexto, aclara los conceptos y le facilita la definición del etiquetado para clasificar la referencia. Este proceso ayuda a formar un modelo mental de esta estructura y se detecta una generalización que ayudará a la clasificación. Continuando con el ejemplo:

<paragraph>

On the other hand, there has been relatively little work on sentiment-based clustering and the related task of unsupervised polarity classification. Where the goal is to cluster (or classify) a set of documents (e.g., reviews) according to the polarity (e.g., "thumbs up" or "thumbs down") expressed by the author in an unsupervised manner.

<context><kw>Despite</kw> the large amount of recent work on sentiment analysis and opinion mining, much of it has focused on supervised methods (e.g., <cite id="1">Pang et al. (2002)</cite>, <cite id="2">Kim and Hovy (2004)</cite>, <cite id="3">Mullen and Collier (2004)</cite>). <kw>One weakness of</kw>these existing supervised polarity classification systems is that they are typically domain- and language-specific. </context> Hence, when given a new domain or language, one needs to go through the expensive process of collecting a large amount of annotated data in order to train a high-performance polarity classifier. Some recent attempts have been made to leverage existing sentiment corpora or lexica to automatically create annotated resources for new domains or languages.

<context> <kw>However</kw>, <method>such methods</method> <kw>require</kw> the existence of either a <data>parallel corpus</data> / machine translation engine for projecting/translating annotations/lexica from a resource-rich language to the target language (<cite id="4">Banea et al, 2008</cite>; <cite id="5">Wan, 2008</cite>), or a domain that is "similar" enough to the target domain (<cite id="6">Blitzer et al, 2007</cite>).</context>

When the target domain or language fails to meet this requirement, sentiment-based clustering or unsupervised polarity classification become appealing alternatives. Unfortunately, to our knowledge, these tasks are largely under-investigated in the NLP community.

<context> <cite id="7">Turney's (2002)</cite> <paper>work</paper> <kw>is perhaps</kw> <posfeature>one of the most notable</posfeature> examples of unsupervised polarity classification. <kw>However</kw>, while his system learns the semantic orientation of the phrases in a review in an unsupervised manner, this information is used to predict the polarity of a review <negfeature>heuristically</negfeature>. </context>

</paragraph>

En el primer contexto, que corresponde a las citas 1, 2 y 3 (Pang et al. (2002), Kim and Hovy (2004), Mullen and Collier (2004)), se tienen dos keywords:

<kw>Despite</kw> <kw>One weakness of</kw>

En el segundo contexto, para las citas 4, 5 y 6 (Banea et al, 2008, Wan, 2008, Blitzer et al, 2007), se obtienen los patrones:

```
<kw>However</kw> <method>such methods</method> <kw>require</kw> <data>parallel  
corpus</data>
```

En el tercer contexto, que corresponde a las cita 7 (Turney's (2002)), se tienen los patrones:

```
<paper>work</paper> <kw>is perhaps</kw><posfeature>one of the most  
notable</posfeature> <kw>However</kw> <negfeature>heuristically</negfeature>
```

5. Si al empezar a anotar dentro de un contexto, se detecta que la función solamente reconoce en forma neutral un trabajo anterior ("Acknowledge, Corroboration, Debate" con polaridad neutral), entonces no se marcan los patrones únicamente para esta combinación particular de función - polaridad: "Acknowledge, Corroboration, Debate" con polaridad neutral, que corresponde a la función desagregada "Acknowledge". La justificación para esto tiene que ver conque experimentalmente se ha comprobado que para ese caso, la información que se debe anotar es diversa y es de poca utilidad para el clasificador automático. Para esta función se debe anotar las palabras clave o etiquetas que sirven para diferenciar la polaridad.
6. Las palabras clave que se marquen deben estar relacionadas directamente con la decisión para la clasificación. Ejemplo: <kw>is used by</kw> podría estar relacionada con la función Useful; <kw>satisfactory outcome</kw> estaría relacionada con polaridad positiva; <kw>prior work</kw> podría ser parte de una comparación y estar en un contexto que clasifica a una cita como "Contrast".
7. Usar la información de los patrones para descubrir la función y polaridad de la referencia. Realizar la clasificación de función y polaridad y colocarlas como atributos de la clase "cite". Se escriben las tres primeras letras de la clasificación para función y polaridad. Ejemplo: "<cite id="22" function="bas" polarity="pos">REFERENCIA_A_UNA_CITA</cite>"

En el ejemplo, abajo, se tienen en el primer contexto las palabras clave: <kw>Despite</kw> <kw>One weakness of</kw> que corresponden a presentar una debilidad general de la orientación de los trabajos, pero en un marco de reconocimiento de lo que se ha hecho anteriormente. Por lo tanto la función se etiqueta como parte del grupo "Acknowledge, Corroboration, Debate"; con polaridad negativa que corresponde a la función desagregada Debate. Esta anotación es la misma para todo el contexto.

El etiquetado resultante es:

```
<context>Despite the large amount of recent work on sentiment analysis and  
opinion mining, much of it has focused on supervised methods (e.g., <cite id="1"
```

```
function="ack" polarity="neg">Pang et al. (2002)</cite>, <cite id="2" function="ack" polarity="neg">Kim and Hovy (2004)</cite>, <cite id="3" function="ack" polarity="neg">Mullen and Collier (2004)</cite>). One weakness of these existing supervised polarity classification systems is that they are typically domain- and language-specific. </context>
```

En el segundo contexto se tienen los patrones: <kw>However</kw> <method>such methods</method> <kw>require</kw> <data>parallel corpus</data>, que expresan también una debilidad general de la tendencia de los trabajos realizados. Por lo tanto igual que en el caso anterior; la función se etiqueta como parte del grupo “Acknowledge, Corroboration, Debate”; con polaridad negativa que corresponde a la función desagregada Debate. Esta anotación es la misma para todo el contexto.

El contexto etiquetado queda como se muestra a continuación:

```
<context><kw>However</kw>, <method>such methods</method> <kw>require</kw> the existence of either a <data>parallel corpus</data> / machine translation engine for projecting/translating annotations/lexica from a resource-rich language to the target language (<cite id="4" function="ack" polarity="neu">Banea et al, 2008</cite>; <cite id="5" function="ack" polarity="neu">Wan, 2008</cite>), or a domain that is "similar" enough to the target domain (<cite id="6" function="ack" polarity="neu">Blitzer et al, 2007</cite>).</context>
```

En el tercer contexto, se tienen los patrones: <paper>work</paper> <kw>is perhaps</kw><posfeature>one of the most notable</posfeature> <kw>However</kw> <negfeature>heuristically</negfeature>; que corresponden a un Hedge porque empiezan estableciendo una característica positiva, para luego decir una característica negativa que le da un vuelco de crítica encubierta al contexto. La clasificación de la cita en el contexto es de Hedge, que siempre tiene polaridad negativa. La anotación se realiza de la siguiente forma:

```
<context> <cite id="7" function="hed" polarity="neg">Turney's (2002)</cite> <paper>work</paper> <kw>is perhaps</kw> <posfeature>one of the most notable</posfeature> examples of unsupervised polarity classification. <kw>However, </kw> while his system learns the semantic orientation of the phrases in a review in an unsupervised manner, this information is used to predict the polarity of a review <negfeature>heuristically</negfeature>. </context>
```

8. Para asegurar que el etiquetado está correctamente realizado, se recomienda que se revise lo anotado, antes de seguir adelante.
9. Validar el XML. Si se usa NetBeans, usar la herramienta Check File XML dentro de la opción Run.
10. No se etiquetan pies de figuras o tablas debido a que se están usando archivos de datos que no las muestran correctamente.

iii. Etiquetas para la anotación y palabras clave de ejemplo

Para realizar las anotaciones se usan las siguientes etiquetas semánticas en XML para reconocer partes importantes en la estructura del texto, partes que servirán para formarse un modelo mental que aclare la función y polaridad de la cita. XML es un lenguaje de marcado para describir estos datos de modo que se puedan realizar procesos inteligentes sobre ellos.

Esta lista, que se expone en la Tabla 61, ha sido depurada durante el proceso de anotación y las etiquetas que se han mantenido han demostrado su utilidad para representar las más variadas instancias en el texto de los artículos científicos.

Tabla 61: Etiquetas

Conceptos	Etiquetas XML	Etiquetas
Trabajo citado	<cited>	CITE
Autor que cita	<author>	AUTHOR
Teoría	<theory>	THEORY
Acción	<action>	ACTION
Método	<method>	METHOD
Datos	<data>	DATA
Contribución	<contribution>>	CONTRIBUTION
Aplicación	<application>	APPLICATION
Herramienta	<tool>	TOOL
Concepto	<concept>>	CONCEPT
Tarea	<task>	TASK
Resultado	<result>	RESULT
Persona(s)	<person>	PERSON

Experimento	<experiment>	EXPERIMENT
Campo del conocimiento	<field>	FIELD
Artículo	<paper>	PAPER
Característica	<feature>	FEATURE
Característica positiva	<posfeature>	POSFEATURE
Característica negativa	<negfeature>	NEGFEATURE

Por ejemplo, las etiquetas <posfeature> y <negfeature> tienen por objeto reconocer construcciones especiales como Hedges, pero pueden ser intercambiables con palabras clave. Las palabras clave en forma de n-gramas sirven para reconocer las funciones a las que se encuentran asociadas y le sirven al anotador para guiarle en la clasificación y tendrán una utilidad en el proceso de etiquetado automático del corpus. Estas palabras se muestran en las Tabla 62 y Tabla 63.

Tabla 62: Ejemplos de etiquetas y palabras clave asociadas a funciones agrupadas

Función agrupada	Ejemplos de etiquetas semánticas asociadas	Ejemplos de palabras clave asociadas
Based on, Supply	<author>we</author>	adopt
	<author>our work</author>	follow
	<author>name_of_our_tool</author>	evaluate
	<posfeature>have been shown to be effective</posfeature>	use
	<posfeature>have achieved good>	are implemented
Useful	<task>rating scale problem</task>	elected to use
	<tool>k-way classifiers</tool>	by using
	<method>Support Vector Machines>	is described in
	<method>approach</method>	proposed a solution
	<method>"Good Grief"	may be addressed by
	algorithm</method>	by taking
	<concept>ME models</concept>	some example applications include
Contrast	<tool>zone classifier</tool>	compare to
	<paper>prior work</paper>	like

	<p><author>this work</author></p> <p><result>NB performance</result></p> <p><negfeature>is problematic for</negfeature></p> <p><posfeature>has a particular relevance</posfeature></p> <p><tool>PageRank</tool></p>	<p>with the exception of most current techniques while compares favourably to is similar to is different from is equivalent to rather than whilst</p>
Acknowledge, Corroboration, Debate	<p><concept>models of biological reality</concept></p> <p><experiment>other experiments</experiment></p>	<p>for example existing studies for instance such as described in previous work they explained other work efforts proposed some research is observed in</p>
Weakness, Correct	<p><negfeature>there is little work</negfeature></p> <p><concept>collective system</concept></p> <p><negfeature>to omit</negfeature></p>	<p>however but nevertheless there is no comprehensive are not able to they do not attempt to was not evaluated an issue with problematic erroneous although</p>
Hedges	<p><posfeature>one of the most notable</posfeature></p> <p><negfeature>insufficient data</negfeature></p> <p><posfeature>has been greatly explored</posfeature></p> <p><negfeature>any satisfactory</p>	<p>to our knowledge from our point of view receive negative feedback it is not clear that suggest that however</p>

solution</negfeature>	it is unclear
<posfeature>powerful statistical approach</posfeature>	as far as we understand
<negfeature>leave problems without resolution</negfeature>	

Tabla 63: Ejemplos de palabras clave asociadas a polaridad

Polaridad	Ejemplos de palabras clave asociadas
Positiva	high accuracy
	in concordance
	agreement on
	there is increased interest
	the earliest work
	easier to operationalise
	has been very effective
Negativa	best solution
	however
	argues
	disputes
	room for improvement
	main problem
	are not part of
Neutral	convey opposing perspectives
	do not take into account
	it is observed
	have been proposed
	studies show
	indicates
	describes
	that appeared
	uses
	for more details see

En el siguiente punto, como información adicional para los anotadores, se presentan ejemplos de etiquetas y palabras clave y su correspondencia con algunas funciones y polaridad.

iv. Ejemplos de anotación para cada función

En esta sección pasaremos a ilustrar el proceso de anotación para cada una de las funciones. Con este objeto, se muestran las etiquetas que evidencian la estructura de la referencia y nos facilitan la clasificación de función y polaridad. Se pide a los anotadores que marquen únicamente las características relevantes para la función y polaridad que reconocen.

En los ejemplos se muestran directamente los contextos de las citas y los formatos de anotación se presentan en XML, donde el atributo “#” de la cita <cite id="#" ...> es el número que identifica a la referencia analizada.

a. Based on, Supply

Función desagrupada “Based on”: El artículo que referencia se construye sobre conceptos, herramientas, métodos y/o datos de la cita.

Función desagrupada “Supply”: Conceptos, herramientas, métodos y/o datos de la cita son usados en el artículo que referencia.

Las dos funciones se agrupan en “Based on, Supply” porque ambas tienen que ver con artículos que **son usados en el artículo que las menciona**, generalmente en la anotación se va a encontrar la etiqueta <author>. Si la polaridad es positiva la función se trata de “Based on”, si la polaridad es neutral, hemos constatado que probablemente se trata de “Supply”.

Para etiquetar esta función se debe detectar una etiqueta <author>; si luego de analizar el contexto, no se va a clasificar a la cita como “Based on, Supply”, entonces es mejor no etiquetar <author>, a menos que sea para una función “Contrast”.

Ejemplos de aplicación:

This feature set is based on (Dong and Schäfer, 2011) [CEPF]. It includes a list of cue words (cuesk), then the frequency features popularity, density, avgDensity, and the syntactic feature POS-patternk.

Patrón: <data>This feature set</data><kw>is based on</kw><cite id="#" function="bas" polarity="pos">Dong and Schäfer, 2011</cite>

Observación: El autor usa el contenido de la cita como base de su trabajo.

Clasificación: De acuerdo a lo que se anotó como atributos de la cita, en este caso la función es “Based on, Supply”, la polaridad es “Positive” porque la disposición del autor del artículo que hace la mención debe ser favorable puesto que, de acuerdo a lo que dicen las palabras clave, su trabajo se basó en algún contenido de la cita. Siendo “Based on, Supply” con polaridad positiva, entonces la función desagregada es “Based on”.

We trained the Stanford MaxEnt classifier (Manning and Klein, 2003) for each of the four facets in a 5-fold cross validation setup with default settings except that we set the regularization parameter $\alpha = 10$ based on previous experiments.

Patrón: <author>we</author><action>trained</action><tool>Stanford MaxEnt classifier</tool><cite id="#" function="bas" polarity="neu"> Manning and Klein, 2003</cite>

Observación: El autor usa la herramienta de la cita para entrenar sus datos.

Clasificación: La función agrupada es “Based on, Supply”, la polaridad es “Neutral”; es decir la función desagregada es “Supply”.

In order to find the marginal probabilities of x_i s in a MRF is best to use Belief Propagation (Yedidia et al.).

Patrón: <author>we</author><kw>can use</kw><method> Belief Propagation</method><cite id="#" function="bas" polarity="neu"></cite>

Observación: El autor usa el método.

Clasificación: La función es “Based on, Supply”, la polaridad es “Neutral”; es decir que la función desagregada es “Supply” porque no indica una disposición ni positiva ni negativa hacia la referencia.

For this application, we have good results using the ACE 2005 corpus (Walker et al., 2006), which consists of 599 documents coming from broadcast conversation, broadcast news, conversational telephone speech, newswire, weblog and usenet newsgroups.

Patrón: <author>we</author><kw>have good results using</kw><data>ACE 2005 corpus</data><cite id="#" function="bas" polarity="pos">Walker et al., 2005</cite>

Observación: El autor usa los datos de la cita y muestra una disposición favorable porque menciona que tuvieron buenos resultados.

Clasificación: La función es “Based on, Supply”, la polaridad es “Positive”; es decir “Based on”.

In this study, we adopt the widely-used SVM classifier (Joachims, 2002).

Patrón: <author>we</author> <kw>adopt</kw> <posfeature>widely-used</posfeature><method>SVM classifier</method><cite id="#" function="bas" polarity="pos"> Joachims, 2002</cite>

Observación: El autor adopta el método planteado en la cita y dice algo positivo sobre él: “widely used”.

Clasificación: La función es “Based on, Supply”, la polaridad es “Positive”, es decir la función desagregada es “Based on”.

We used SimFinder (Hatzivassiloglou et al., 2001), a state-of-the-art system for measuring sentence similarity based on shared words, phrases, and WordNet synsets.

Patrón: <author>we</author> <kw>used</kw> <tool>SimFinder</tool> <posfeature>state-of-the-art</posfeature> <cite id="#" function="bas" polarity="pos">Hatzivassiloglou et al., 2001</cite>

Observación: El autor usa una herramienta de la cita y la califica positivamente.

Clasificación: La función es “Based on, Supply”, la polaridad es “Positive; por lo tanto la función desagregada es “Based on”.

b. Useful

Useful: Concepto, método, herramienta y/o dato se aplica en algún otro trabajo que se referencia; no se aplica en el propio trabajo que lo menciona.

Se enfatiza la usabilidad del material por lo que la polaridad es generalmente positiva, aunque también puede ser neutral cuando únicamente se habla de algún trabajo en el que se usa la cita y no se emite un juicio claro sobre él. El autor no usa el material; se refiere en forma indefinida a los beneficiarios del mismo, ya sea sin mencionarlos o refiriéndose a ellos como “usuarios”, o algo similar, o usando voz pasiva para omitir el sujeto. Es importante que en la definición de esta función, en la estructura se encuentre etiquetas como <task>, <method> o <tool> que por sus características denotan que se trata de un contenido que puede ser usado; o palabras clave que demuestran utilidad como por ejemplo <kw>in order to</kw> o <action>applied</action>

La función “Useful” no tiene polaridad “Negative”.

Ejemplos de aplicación:

In scientific texts, knowing the type of information that a zone represents (e.g., background knowledge, hypothesis, experimental observation, conclusion, etc.) allows for automatic isolation of new knowledge claims (Sandor and de Waard, 2012).

Patrón: <kw>allows for</kw><task>automatic isolation of</task><cite id="#" function="use" polarity="pos">Sandor and de Waard, 2012</cite>

Observación: El autor enfatiza la usabilidad del concepto, pero no define a los beneficiarios de la contribución. Esta función generalmente está asociada a un método, una tarea o una herramienta.

Clasificación: La función es “Useful”, la polaridad es “Positive”.

U-Compare (Kano et al., 2011) is a UIMA-based workflow construction platform that provides a graphical user interface (GUI) via which users can rapidly create NLP pipelines using a drag-and-drop mechanism.

Patrón: <tool>U-Compare</tool> <cite id="#" function="use" polarity="pos">Kano et al., 2011</cite> <kw>provides</kw> <tool>graphical user interface</tool><kw>rapidly</kw>

Observación: La herramienta mencionada en la cita es útil y tiene una característica positiva.

Clasificación: La función es “Useful”, la polaridad es “Positive”.

Graph-based ranking algorithms such as Kleinberg's HITS (Kleinberg, 1999) or Google's PageRank (Brin and Page, 1998) have been successfully applied in citation network analysis and ranking of webpages.

Patrón: <method>Graph-based ranking algorithms</method><kw>have been successfully applied in</kw><task>citation network analysis</task>

Observación: Dos métodos han sido usados en forma exitosa por alguien más (no el autor).

Clasificación: La función es “Useful”, la polaridad es “Positive”.

In recent years, graph-based ranking algorithms have been successfully used for document summarization (Mihalcea and Tarau, 2004, 2005; ErKan and Radev, 2004) and keyword extraction (Mihalcea and Tarau, 2004). Such algorithms make use of "voting" or "recommendations" between sentences (or words) to extract sentences (or keywords) .

Patrón: <method>graph-based ranking algorithms</method> <kw>have been successfully used for</kw> <task>document summarization</task>

Conclusión: Un método ha sido usado exitosamente para realizar tareas, no especifica el autor que lo usó.

Clasificación: La función es “Useful”, la polaridad es “Positive”.

(Re) rankers have been successfully applied to numerous NLP tasks, such as parse selection (Osborne and Baldridge, 2004; Toutanova et al., 2004), parse reranking (Collins and Duffy, 2002; Charniak and Johnson, 2005), question-answering (Ravichandran et al., 2003) .

Patrón: <method>(Re) rankers</method> <kw>have been successfully applied to</kw> <task>NLP tasks</task> <cite id="#" function="use" polarity="pos"> Osborne and Baldridge, 2004</cite> <cite id="#" function="use" polarity="pos"> Toutanova et al., 2003</cite> <cite id="#" function="use" polarity="pos"> Collins and Duffy, 2002</cite> <cite id="#" function="use" polarity="pos"> Charniak and Johnson, 2005</cite> <cite id="#" function="use" polarity="pos"> Ravichandran et al., 2003</cite>

Observación: Una herramienta ha sido exitosamente aplicada para realizar varias tareas por diferentes autores (no el autor que cita). El contexto y la clasificación son compartidas por varias citas.

Clasificación: La función es “Useful”, la polaridad es “Positive”.

Their annotation guidelines follow those of the OntoNotes project (Hovy et al., 2006)

Patrón: <method>annotation guidelines</method><kw>follow</kw><cite id="#" function="use" polarity="neu"> Hovy et al., 2006</cite>

Observación: La cita tiene aplicaciones.

Clasificación: La función es “Useful”, la polaridad es “Neutral”.

Kin and Webber (2006) investigate a special aspect, citation sentences where a pronoun such as “they” refers to a previous citation. The study is performed on astronomy journal and a maximum-entropy classifier is trained.

Patrón: <cite id="#" function="use" polarity="neu">Kim and Webber (2006)</cite> <tool>classifier</tool> <action>is trained</action>.

Observación: Se reconoce que el estudio que realiza la cita es útil para entrenar datos en una revista de astronomía usando el clasificador con el algoritmo de máxima-entropía. Se reconoce que es “Useful” porque está asociado a las etiquetas <tool> y <action>.

Clasificación: La función es “Useful”, la polaridad es “Neutral”.

Texto: CST is an expanded rhetorical structure analysis based on RST Mann and Thompson, 1988, and attempts to describe relations between two or more sentences from both single and multiple document sets.

Patrón: <tool>CST</tool> is an expanded rhetorical structure analysis <kw>based on</kw> <method>RST</method> (<cite id="4" function="use" polarity="pos">Mann and Thompson, 1988</cite>), <kw>and attempts to</kw> <task>describe relations between two or more sentences</task> from both single and multiple document sets.

Observación: A pesar de que se tiene el <kw>is based on</kw>, la función se clasifica como “Useful” porque no se tiene <autor> como beneficiario de la herramienta.

Clasificación: La función es “Useful”, la polaridad es “Neutral”.

c. Acknowledge, Corroboration, Debate

Acknowledge: La cita se menciona para reconocer algún trabajo o concepto previo; la polaridad es neutral.

Corroboration: El artículo que referencia reconoce y tiene una disposición positiva hacia la cita; la polaridad es positiva.

Debate: La cita es reconocida y el autor del artículo que la referencia tiene una disposición negativa hacia ella; la polaridad es negativa.

Las tres funciones se agrupan porque se relacionan con un **reconocimiento de trabajos anteriores**. Se diferencian de “Useful” porque no se enfatiza la usabilidad del material de la cita.

Ejemplos de aplicación:

This non-local expression of sentiment has been observed in other genres as well (Wilson et al., 2009; Polanyi and Zaenen, 2006).

Patrón: <kw>has been observed</kw><cite id="#" function="ack" polarity="neu"> Wilson et al., 2009</cite><cite id="#" function="ack" polarity="neu"> Polanyi and Zaenen, 2006</cite>

Observación: Se reconoce la existencia del concepto planteado en la cita. El contexto y la clasificación son compartidos por dos citas.

Clasificación: La función es “Acknowledge, Corroboration, Debate”, la polaridad es “Neutral”; que corresponde a la función desagregada “Acknowledge”.

It has previously been shown that considering several functional discourse annotation schemes in parallel can be beneficial (Liakata et al., 2012b), since each scheme offers a different perspective.

Patrón: <kw>can be beneficial</kw> <cite id="#" function="ack" polarity="pos">Liakata et al., 2012b</cite>

Observación: Se reconoce que el contenido de la cita puede ser beneficioso; corrobora ese concepto.

Clasificación: La función es “Acknowledge, Corroboration, Debate”, la polaridad es “Positive”; es decir corresponde a la función desagregada “Corroboration”.

According to A. Esuli and F. Sebastiani, 2006, opinion mining consists both in searching for the opinions or sentiments expressed in a document.

Patrón: <kw>according</kw><cite id="#" function="ack" polarity="neu"> A. Esuli and F. Sebastiani, 2006</cite>

Observación: Reconoce el concepto emitido en la cita.

Clasificación: La función es “Acknowledge, Corroboration, Debate”, la polaridad es “Neutral”; es decir que la función desagregada es “Acknowledge”.

Work on applying machine learning techniques for automatic citation classification is currently underway (Teufel et al., 2006); the agreement of one annotator and the system is currently K=0.57, leaving plenty of room for improvement.

Patrón: <cite id="39" function="ack" polarity="neg">Teufel et al., 2006</cite><kw>room for improvement</kw>

Observación: Se reconoce un trabajo pero se debate con respecto a él, porque se presenta una característica negativa.

Clasificación: La función es “Acknowledge, Corroboration, Debate”; la polaridad es “Negative” por lo tanto la función desagregada es “Debate”.

d. Contrast

Contrast: Se hace una comparación entre dos citas implícitas o explícitas, en la que se emite un criterio positivo, neutral o negativo sobre la cita. La comparación puede ser con el trabajo del autor o entre dos citas fuera del trabajo que hace la referencia. Se deben etiquetar todas las palabras que expresen comparación y las que sirvan para detectar la polaridad; por ejemplo las etiquetas <kw>compare</kw>, <kw>better than</kw>, <kw>In contrast</kw>, <kw>similar tan</kw>; o las que se refieren a otro trabajo con el que se compara la cita: <kw>prior work</kw>, <autor>our results</kw>; o indicadores de polaridad: <kw>positive</kw>, <kw>room for improvement</kw>.

Ejemplos de aplicación:

Firstly, unlike annotation platforms such as brat (Stenetorp et al., 2012), U-Compare allows analysis components to be integrated into workflows in a straightforward and user-interactive manner.

Patrón: <kw>unlike</kw> <tool>brat</brat> <cite id="#" function="con" polarity="pos"> Stenetorp et al., 2012</cite> <tool>U-Compare</tool> <kw>allows</kw> <posfeature>straightforward and user-interactive manner</posfeature>

Observación: Se comparan herramientas y la herramienta de la cita tiene características mejores.

Clasificación: La función es “Contrast”, la polaridad es “Positive”.

The models of (T.Mullen and N.Collier, 2004), (V.Stoyanov et al.,2005) and (H.Yu and V.Hatzivassiloglou, 2003) were considered more specific than those of level 2 because they stressed the importance of target and source for opinion mining.

Patrón: <cite id="#" function="con" polarity="pos">T.Mullen and N.Collier, 2004</cite> <cite id="#" function="con" polarity="pos"> V.Stoyanov et al.,2005</cite> <cite id="#" function="con" polarity="pos"> H.Yu and V.Hatzivassiloglou, 2003</cite> <kw>were considered more specific than</kw> <posfeature>stressed the importance</posfeature>

Observación: Se comparan modelos contenidos en varias citas con otros y el resultado es positivo para las citas en la referencia.

Clasificación: La función es “Contrast”, la polaridad es “Positive”.

In the biomedical domain, the Biomedical Discourse Relation Bank (BioDRB) (Prasad et al., 2011) annotates a similar set of relation types, whilst BioCause focusses on causality (Mihaila et al., 2013).

Patrón: <tool>BioDRB</tool> <cite id="#" function="con" polarity="neu"> Prasad et al., 2011</cite> <kw>a similar set</kw> <kw>whilst</kw> <tool>BioCause</tool> <cite id="#" function="con" polarity="neu"> Mihaila et al., 2013</cite>

Observación: Se compara en forma neutral dos herramientas y sus diferentes enfoques con la respectiva cita.

Clasificación: La función es “Contrast”, la polaridad es “Neutral”.

The BioScope corpus (Vincze et al., 2008) annotates the scopes of negative and speculative keywords, whilst Morante and Daelemans (2009) have trained a system to undertake this task.

Patrón: <cite id="#" function="con" polarity="neu"> Vincze et al., 2008</cite> <kw>whilst</kw> <cite id="#" function="con" polarity="neu"> Morante and Daelemans (2009)</cite>

Observación: Se comparan dos citas, la primera anota un corpus, y la segunda entrena un sistema para realizar la misma tarea.

Clasificación: La función es “Contrast”, la polaridad es “Neutral”.

Our result is a 23% F-score increase on the Computational Linguistics conference papers marked up by Teufel (1999) .

Patrón: <author>our result</author> <kw>increase</kw> <result>23%F-score</result> <cite id="#" function="con" polarity="neg">Teufel (1999)</cite>

Observación: Se comparan los resultados propios con los de la cita y se detecta que los de los propios son mejores, por lo que la polaridad para la cita es negativa.

Clasificación: La función es “Contrast”, la polaridad es “Negative”.

Also, when compared our approach to argumentative zoning and more specifically its extension for chemistry papers, AZ-II Teufel et al., 2009, it was shown to provide a greater level of detail in terms of categories denoting objectives, methods and outcomes whereas AZ-II focusses on the attribution of knowledge claims and the relation with previous work.

Patrón: <kw>compared</kw><author>our approach to</author>><cite id="#" function="con" polarity="neg">Teufel et al., 2009</cite><posfeature>greater level of</posfeature>

Observación: Se comparan los resultados propios con los de otro autor y los resultados son negativos para el otro autor.

Clasificación: la función es “Contrast”, la polaridad es “Negative”.

e. Weakness, Correct

Weakness: Se nota un error o debilidad del trabajo citado.

Correct: Se corrige un error o debilidad del trabajo citado.

Estas funciones se agrupan porque tienen que ver con un error o debilidad de la cita que se menciona directa o veladamente. Se diferencia entre Weakness y Correct usando el contexto de la cita. Siempre son negativas.

Ejemplos de aplicación:

However, GATE (Cunningham et al., 2002) implements a limited workflow management mechanism that does not support the execution of parallel or nested workflows.

Patrón: <kw>however</kw> <cite id="#" function="wea" polarity="neg">Cunningham et al., 2002</cite> <negfeature>limited</negfeature>

Observación: La herramienta de la cita tiene una debilidad que se hace notar.

Clasificación: La función es “Weakness, Correct”, la polaridad es “Negative”.
Corresponde a la función desagregada “Weakness”.

For implicit citation extraction, Kaplan et al. (2009) explore co-reference chains for citation extraction using a combination of co-reference resolution techniques. However, their corpus consists of only 94 sentences of citations to four papers, which is likely to be too small to be representative.

Patrón: <cite id="#" function="wea" polarity="neg">Kaplan et al. (2009)</cite> <kw>However</kw> <negfeature>too small to be representative</negfeature>

Observación: Se expone una debilidad de la la cita.

Clasificación: La función es “Weakness, Correct”, la polaridad es “Negative”.
Corresponde a la función desagregada “Weakness”.

From this, we can see that the n-grams (unigrams and bi-grams) have by far the largest impact - and neither of these feature types was directly implemented by Teufel and Moens (2002).

Patrón: <negfeature>neither</negfeature><cite id="#" function="wea" polarity="neg">Teufel and Moens (2002)</cite>

Observación: Se evidencia una debilidad de la cita.

Clasificación: La función es “Weakness, Correct”, la polaridad es “Negative”.
Corresponde a la función desagregada “Weakness”.

Qazvinian & Radev (2010) also make use of citation sentences in other scientific papers to summarize the contributions of a paper. The drawback of citation summaries is that a paper must be already cited, so this type of summary will not be useful to a paper reviewer. Also, citations of articles will have been influenced by other citations rather than the paper itself.

Patrón: <cite id="#" function="wea" polarity="neg"> Qazvinian & Radev (2010)</cite><kw>drawback of</kw><kw>will not be useful</kw>

Observación: Se nota una debilidad de un método usado por la cita.

Clasificación: La función es “Weakness, Correct”, la polaridad es “Negative”.
Corresponde a la función desagregada “Weakness”.

Our work improves the agreement between annotators presented by Teufel (2006) with K = 0.59, which is an unsatisfactory outcome.

Patrón: <author>Our work</author> <kw>improves</kw> <cite id="#" function="wea" polarity="neg">Teufel (2006)</cite> <negfeature>unsatisfactory outcome</negfeature>

Observación: Se hace notar una debilidad del trabajo citado y se presenta que el autor ha mejorado ese resultado.

Clasificación: La función es “Weakness, Correct”, la polaridad es “Negative”; corresponde a “Correct” porque en el contexto se nota que la debilidad presentada ya se ha superado.

f. Hedges

Para esta categorización de función, se usa un lenguaje cuidadoso para disimular una crítica. Se puede clasificar como “Hedges” cuando se tiene primero una característica positiva, pero luego se presenta una característica negativa de la misma cita. La característica favorable solamente se la menciona para suavizar la segunda afirmación que contiene una crítica. Otra forma de “Hedges” se puede encontrar cuando se tienen frases como: “it is unclear”, “as far as we know” y otras con las que se pretende expresar desconocimiento sobre el tema para eludir la expresión directa de un juicio negativo.

It is not clear whether unsupervised topic modelling such as (Chen et al., 2009) can be applied to scientific articles (over 100 sentences long), which by nature include repetition of topics.

Patrón: <kw>It is not clear</kw> <cite id="#" function="hed" polarity="neg">Chen et al., 2009</cite> <kw>can be applied to</kw>

Observación: Es un “hedge” orientado al lector en el que se admite falta de conocimiento para evitar criticar directamente al contenido de la cita.

Clasificación: La función es “Hedges”; la polaridad es “Negative”.

Sentence extraction, e.g. Brandow et al. (1995) and Kupiec et al. (1995), selects a small number of abstract worthy sentences from a larger text. We do not know if the resulting sentences form a collection of excerpt sentences that capture the essence of the text.

Patrón: <cite id="#" function="hed" polarity="neg">Brandow et al. (1995)</cite> <cite id="#" function="hed" polarity="neg">Kupiec et al. (1995)</cite> <kw>do not know</kw>

Observación: Es un hedge en el que se usa la expresión “do not know” para expresar falta de confianza en los resultados de la cita.

Clasificación: La función que se marca es “Hedges” y la polaridad es “Negative”.

The first experiments in Argumentative Zoning used Naïve Bayes (NB) classifiers Kupiec et al., 1995; Teufel, 1999) which assume conditional independence of the features. However, this assumption is rarely true for the kinds of rich feature representations we want to use for most NLP tasks.

Patrón: <cite id="#" function="hed" polarity="neg"> Kupiec et al., 1995</cite> <cite id="#" function="hed" polarity="neg"> Teufel, 1999</cite> <kw>However</kw> <kw>is rarely true</kw>

Observación: Es un hedge porque no se dice categóricamente que lo que se asume en la cita no es cierto, sino que se manifiesta que es “rarely true”.

Clasificación: La función que se marca es “Hedges” y la polaridad es “Negative”.

Pang et al. (2002) did not compare the result of using and not using the negation context effect, so it is not clear how much it improved their result. In our task, it is clear that the MOREL'XESS feature has a significant effect on the performance, especially for the frequency features.

Patrón: <cite id="#" function="hed" polarity="neg">Pang et al. (2002)</cite> <kw>it is not clear</kw>

Observación: Es un hedge orientado al lector porque suaviza su crítica expresando falta de conocimiento respecto a una característica negativa de la cita.

Clasificación: La función que se marca es “Hedges” y la polaridad es “Negative”.

The only recent work on citation sentiment detection using a relatively large corpus is by Athar (2011). However, this work does not handle citation context.

Patrón: <posfeature>relatively large corpus</posfeature> <cite id="#" function="hed" polarity="neg">Athar (2011) (2002)</cite> <kw>however</kw> <negfeature>does not handle</negfeature>

Observación: Para la función “Hedge” es un patrón repetitivo la presencia de una característica positiva seguida de una negativa, que generalmente se expresan con palabras clave como la que se presentan en este ejemplo. La característica positiva se expresa con las palabras clave “using a relatively large” y la negativa “does not handle”.

Conclusión: Es un hedge orientado al lector porque suaviza su crítica diciendo primero una característica negativa.

Clasificación: La función que se marca es “Hedge” y la polaridad es “Negative”.



ANEXO 2: Archivos para el cálculo del acuerdo entre anotadores

En la Tabla 64 se presenta el contenido del archivo para el cálculo del acuerdo entre anotadores cuando se realiza el paso de pre-anotación y por lo tanto se tiene una mejoría evidente en la comparación de las anotaciones realizadas por las tres personas participantes.

Tabla 64: Resultados de la anotación para cálculo del acuerdo entre anotadores

Id cita	Función			Polaridad		
	Anotador 1	Anotador 2	Anotador 3	Anotador 1	Anotador 2	Anotador 3
1	USE	USE	USE	POS	POS	POS
2	WEA	WEA	WEA	NEG	NEG	NEG
3	USE	USE	USE	POS	POS	POS
4	WEA	WEA	WEA	NEG	NEG	NEG
5	WEA	WEA	WEA	NEG	NEG	NEG
6	USE	USE	USE	POS	POS	POS
7	BAS	BAS	BAS	POS	POS	POS
8	USE	USE	USE	POS	POS	POS
9	WEA	WEA	WEA	NEG	NEG	NEG
10	USE	USE	USE	POS	POS	POS
11	USE	USE	USE	POS	POS	POS
12	USE	USE	BAS	POS	POS	POS
13	USE	USE	USE	POS	POS	POS
14	USE	USE	USE	POS	POS	POS
15	ACK	ACK	USE	POS	POS	POS

16	ACK	ACK	USE	POS	POS	POS
17	BAS	BAS	ACK	POS	POS	POS
18	ACK	ACK	BAS	POS	POS	POS
19	USE	BAS	ACK	POS	POS	POS
20	ACK	ACK	ACK	POS	POS	POS
21	ACK	ACK	ACK	POS	POS	POS
22	ACK	ACK	ACK	POS	POS	POS
23	ACK	ACK	ACK	POS	POS	POS
24	ACK	ACK	ACK	POS	POS	POS
25	ACK	ACK	ACK	POS	POS	POS
26	ACK	ACK	ACK	POS	POS	POS
27	ACK	ACK	ACK	POS	POS	POS
28	ACK	ACK	ACK	POS	POS	POS
29	USE	USE	USE	POS	POS	POS
30	ACK	ACK	ACK	POS	POS	POS
31	ACK	ACK	ACK	POS	POS	POS
32	WEA	WEA	WEA	NEG	NEG	NEG
33	WEA	WEA	WEA	NEG	NEG	NEG
34	ACK	ACK	ACK	POS	POS	POS
35	ACK	ACK	ACK	POS	POS	POS
36	WEA	WEA	WEA	NEG	NEG	NEG
37	ACK	ACK	ACK	POS	POS	POS
38	ACK	ACK	ACK	POS	POS	POS
39	ACK	ACK	ACK	POS	POS	POS
40	ACK	ACK	ACK	POS	POS	POS
41	ACK	ACK	ACK	POS	POS	POS
42	ACK	ACK	ACK	POS	POS	POS

43	ACK	ACK	ACK	POS	POS	POS
44	ACK	ACK	ACK	POS	POS	POS
45	ACK	ACK	ACK	POS	POS	POS
46	ACK	ACK	BAS	POS	POS	POS
47	ACK	ACK	ACK	POS	POS	POS
48	BAS	BAS	BAS	POS	POS	POS
49	BAS	BAS	BAS	POS	POS	POS
50	BAS	USE	USE	POS	POS	POS
51	ACK	ACK	ACK	POS	POS	POS
52	ACK	ACK	ACK	POS	POS	POS
53	ACK	ACK	ACK	POS	POS	POS
54	ACK	ACK	ACK	POS	POS	POS
55	ACK	ACK	ACK	POS	POS	POS
56	ACK	ACK	ACK	POS	POS	POS
57	ACK	ACK	ACK	POS	POS	POS
58	ACK	ACK	ACK	POS	POS	POS
59	USE	USE	USE	POS	POS	POS
60	USE	USE	USE	POS	POS	POS
61	USE	USE	USE	POS	POS	POS
62	USE	USE	USE	POS	POS	POS
63	BAS	BAS	BAS	POS	POS	POS
64	ACK	ACK	ACK	POS	POS	POS
65	BAS	BAS	BAS	POS	POS	POS
66	BAS	BAS	BAS	POS	POS	POS
67	BAS	BAS	BAS	POS	POS	POS
68	BAS	BAS	BAS	POS	POS	POS
69	BAS	BAS	BAS	POS	POS	POS

70	BAS	BAS	BAS	POS	POS	POS
71	ACK	ACK	ACK	POS	POS	POS
72	ACK	ACK	ACK	POS	POS	POS
73	ACK	ACK	ACK	POS	POS	POS
74	ACK	BAS	BAS	POS	POS	POS
75	BAS	BAS	BAS	POS	POS	POS
76	BAS	BAS	BAS	POS	POS	POS
77	BAS	BAS	BAS	POS	POS	POS
78	ACK	ACK	ACK	POS	POS	POS
79	ACK	ACK	ACK	POS	POS	POS
80	ACK	ACK	ACK	POS	POS	POS
81	BAS	BAS	BAS	POS	POS	POS
82	BAS	BAS	BAS	POS	POS	POS
83	ACK	ACK	ACK	POS	POS	POS
84	ACK	BAS	BAS	POS	POS	POS
85	ACK	ACK	ACK	POS	POS	POS
86	ACK	WEA	WEA	POS	NEG	POS
87	ACK	WEA	WEA	POS	NEG	POS
88	WEA	ACK	ACK	NEG	POS	POS
89	ACK	ACK	ACK	POS	POS	POS
90	ACK	ACK	ACK	POS	POS	POS
91	ACK	ACK	ACK	POS	POS	POS
92	ACK	ACK	ACK	POS	POS	POS
93	ACK	ACK	ACK	POS	POS	POS
94	ACK	ACK	ACK	POS	POS	POS
95	WEA	WEA	WEA	NEG	NEG	NEG
96	WEA	WEA	WEA	NEG	NEG	NEG

97	WEA	WEA	WEA	NEG	NEG	NEG
98	WEA	WEA	WEA	NEG	NEG	NEG
99	WEA	WEA	WEA	NEG	NEG	NEG
100	ACK	ACK	ACK	POS	POS	POS
101	ACK	ACK	ACK	POS	POS	POS



Universitat d'Alacant
Universidad de Alicante

ANEXO 3: Análisis de Factores para la valoración del Impacto con resultados de encuesta a autores

Tabla 65: Impacto calculado con nuestro método vs. Impacto según autores

Cita	Función	Polaridad	Sección(es)	Número de secciones en que aparece	Impacto calculado con nuestro método	Impacto según autores
Van Deemter and Kibble 2000	Ack	Neu	Introducción	1	Perfunctory	Perfunctory
Ng 2010	Ack	Neu	Introducción	1	Perfunctory	Perfunctory
Mitkov 1999	Ack	Neu	Introducción	1	Perfunctory	Perfunctory
Grishman and Sundheim 1996	Use	Neu	Introducción	1	Perfunctory	Perfunctory
Doddington et al. 2004	Use,Bas	Neu,Neu	Introduction, Method	2	Significant	Significant
Pradhan et al. 2011	Use	Neu	Introduction	1	Perfunctory	Perfunctory
Nguyen et al. 2011	Use	Neu	Introducción	1	Perfunctory	Perfunctory
Watson et al. 2003	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Gaizauskas and Humphreys 2000	Use	Neu	Introduction	1	Perfunctory	Perfunctory

Miwa et al. 2012	Use	Pos	Introduction	1	Perfunctory	Perfunctory
Yang et al. 2004	Con	Neg	Method	1	Negative	Negative
Chen et al. 2008	Con	Pos	Method	1	Perfunctory	Perfunctory
Gasperin 2009	Wea	Neg	Method	1	Negative	Perfunctory
Cohen et al. 2010	Con	Pos	Method	1	Perfunctory	Perfunctory
Hovy et al. 2006	Use	Neu	Method	1	Perfunctory	Perfunctory
Kim and Webber 2006	Use	Neu	Method	1	Perfunctory	Perfunctory
Kaplan et al. 2009	Wea	Neg	Method	1	Negative	Perfunctory
Bird et al. 2008	Bas	Pos	Method	1	Significant	Significant
Schäfer et al. 2012	Bas	Pos	Method	1	Significant	Significant
Müller and Strube 2006	Bas	Pos	Method	1	Significant	Significant
Klein and Manning 2003	Bas	Pos	Method	1	Significant	Significant
LDC 2004	Bas	Neu	Method	1	Significant	Significant
Drozdzynski et al. 2004	Bas	Neu	Method	1	Significant	Perfunctory
Haghghi and Klein 2009	Use	Neu	Discussion	1	Perfunctory	Perfunctory
Hirschman et al. 1997	Ack, Bas	Neu	Introduction, Results	2	Significant	Significant
Vilain et al. 1995	Bas	Neu	Results	1	Significant	Significant
Luo 2005	Wea	Neg	Results	1	Negative	Perfunctory
Bengtson and Roth 2008	Use	Neu	Discussion	1	Perfunctory	Perfunctory

Rizzolo and Roth 2010	Use	Neu	Discussion	1	Perfunctory	Perfunctory
Stoyanov et al. 2010	Use	Neu	Discussion	1	Perfunctory	Perfunctory
Raghunathan et al. 2010	Use	Neu	Discussion	1	Perfunctory	Perfunctory
Lee et al. 2011	Use	Neu	Discussion	1	Perfunctory	Perfunctory
Taboada2011	Ack	Neu	Method	1	Perfunctory	Perfunctory
Medhat2014	Ack	Neu	Method	1	Perfunctory	Perfunctory
Pang2008	Ack	Neu	Method	1	Perfunctory	Perfunctory
Tan2009	Ack	Neu	Method	1	Perfunctory	Perfunctory
Kim 2004	Ack	Neu	Method	1	Perfunctory	Perfunctory
Cruz 2013	Use	Pos	Method	1	Perfunctory	Perfunctory
Medhat 2014	Ack	Neu	Method	1	Perfunctory	Perfunctory
Aranberri 2013	Ack	Neu	Method	1	Perfunctory	Significant
Guthrie 2006	Bas	Pos	Method	1	Significant	Significant
Fernandez 2014	Ack	Pos	Method	1	Significant	Significant
Saif 2012	Bas	Pos	Method	1	Significant	Significant
Sebastiani 2002	Bas	Pos	Method	1	Significant	Significant
Mohammad 2013	Bas	Pos	Method	1	Significant	Significant
Chang 2011	Use	Neu	Method	1	Perfunctory	Perfunctory
Nakov 2013	Use	Neu	Results	1	Perfunctory	Perfunctory
Rosenthal 2014	Use	Neu	Results	1	Perfunctory	Perfunctory
Villena 2013	Use	Neu	Results	1	Perfunctory	Perfunctory
Villena 2013B	Use	Neu	Results	1	Perfunctory	Significant

Brown et al 1990	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Melamed 1998	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Davis and Dunning 1995	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Landauer and Littman 1990	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Gale and Church 1991	Ack, Use	Neu, Neu	Introduction, Results	2	Significant	Significant
Melamed 1997	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Koehn 2005	Use	Pos	Introduction	1	Perfunctory	Significant
Gale and Church 1993	Use	Neu	Introduction, Method	2	Significant	Significant
Varga et al. 2005	Ack	Neg	Introduction	1	Perfunctory	Perfunctory
Schmid 1994	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Ding Zhang Chambers Song Wang and Zhai 2014	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Arstein y Poesio 2008	Ack	Neu	Introduction, Method	2	Significant	Significant
Hernandez and Gomez 2014	Bas	Pos	Method	1	Significant	Significant
Zock 2012	Use	Neu	Method	1	Perfunctory	Perfunctory
Salton et al. 1975	Bas, Bas	Pos, Pos	Method, Method	1	Significant	Significant
Guthrie Allison Liu Guthrie and Wilks 2006	Ack	Neu	Method	1	Perfunctory	Perfunctory
Brown et al. 1983	Hed	Neg	Method	1	Negative	Negative

Krippendorff 2004	Ack	Neu	Method	1	Perfunctory	Perfunctory
Landis and Koch 1977	Ack	Neu	Method	1	Perfunctory	Perfunctory
Picard 1997	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Calvo 2013	Ack, Ack,Use, u,Ack	Neu,Neu,Ne u,Neu,Neu	Method,Results, Results,Results, Discussion	3	Significant	Significant
Strapparava 2008	Ack,Use, Use,Use	Neu,Neu,Ne u,Neu	Introduction,Re sults,Results,Re sults,	3	Significant	Significant
Rodriguez 2012	Use	Neu	Introduction	1	Perfunctory	Perfunctory
Desmet 2013	Use	Neu	Introduction	1	Perfunctory	Perfunctory
Vaassen 2014	Use	Neu	Introduction	1	Perfunctory	Perfunctory
Cowie 2003	Ack	Neu	Method	1	Perfunctory	Perfunctory
Francisco 2013	Ack	Neu	Method	1	Perfunctory	Perfunctory
Ekman 1999	Ack,use	Neu,Neu	Method,Results	2	Significant	Significant
Plutchik 1980	Use	Neu	Method	1	Perfunctory	Perfunctory
Kim 2011	Ack,Ack, Ack,Ack	Neu,Neu,Ne u,Neu	Method, Method,Discussi on,Discussion	2	Significant	Significant
Russell 1980	Use	Neu	Method	1	Perfunctory	Perfunctory
Mehrabian 1996	Use	Neu	Method	1	Perfunctory	Perfunctory
Calvo 2013	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Balahur 2011	Use	Neu	Results	1	Perfunctory	Perfunctory
Sykora 2013	Use	Neu	Results	1	Perfunctory	Perfunctory
Deerwester 1999	Use	Pos	Results	1	Perfunctory	Perfunctory
Gill 2008	Use	Neu	Results	1	Perfunctory	Perfunctory

Wang 2013	Use	Neu	Results	1	Perfunctory	Perfunctory
Kovahi 1998	Ack	Neu	Results	1	Perfunctory	Perfunctory
Bishop 2006	Ack	Neu	Results	1	Perfunctory	Perfunctory
Mohri 2012	Ack,Use	Neu,Neu	Results,Results	1	Significant	Significant
Hasan 2014	Ack,Use	Neu,Neu	Results,Results	1	Significant	Significant
Wang 2012	Ack	Neu	Results	1	Perfunctory	Perfunctory
Roberts 2012	Ack,Ack	Neu,Neu	Results,Results	1	Significant	Significant
Suttles 2013	Ack,Use, Ack	Neu,Neu, u	Results,Results, Discussion	2	Significant	Significant
Hasan 2014a	Ack	Neu	Results	1	Perfunctory	Perfunctory
Alm 2005	Ack	Neu	Results	1	Perfunctory	Perfunctory
Roth 1999	Use	Neu	Results	1	Perfunctory	Perfunctory
Roberts 2012	Ack,Ack	Neu,Neu	Results,Results	1	Significant	Significant
Suttles 2013	Ack,Use, Ack	Neu,Neu, u	Results,Results, Discussion	2	Significant	Significant
Mintz 2009.	Use	Neu	Results	1	Perfunctory	Perfunctory
Hasan 2014	Ack,Use	Neu,Neu	Results,Results	1	Significant	Perfunctory
Strapparava 2004	Use	Neu	Results	1	Perfunctory	Perfunctory
Agrawal 2012	Use	Neu	Results	1	Perfunctory	Perfunctory
Bradley 1999	Use	Neu	Results	1	Perfunctory	Perfunctory
Suttles 2013	Ack	Neu	Discussion	1	Perfunctory	Perfunctory
WHO 2014	Ack,Ack, Ack	Neu,Neu, u	Introduction,Int roduction,Intro duction	1	Perfunctory	Significant
WHO 2012	Ack	Neu	Introduction	1	Perfunctory	Significant
Holmes et al. 2007	Ack	Neu	Introduction	1	Perfunctory	Perfunctory

Wasserman et al. 2004	Ack,Ack	Neu,Neu	Introduction,Method	2	Significant	Significant
Berk and Dodd 2006	Ack	Neu	Introduction	1	Perfunctory	Perfunctory
Garcia-Rabagó et al. 2010	Ack	Neu	Method	1	Perfunctory	Perfunctory
Moreno Gea and Blanco Sanchez 2012	Ack	Neu	Method	1	Perfunctory	Perfunctory
Sarno 2008	Ack	Neu	Method	1	Perfunctory	Perfunctory
Mingote et al. 2004	Ack	Neu	Method	1	Perfunctory	Perfunctory
Owen et al. 2012	Ack	Pos	Method	1	Significant	Significant
Isometsa 2001	Ack	Neu	Method	1	Perfunctory	Perfunctory
Cantor 2000	Ack	Neu	Method	1	Perfunctory	Perfunctory
Rudestam 1971	Ack	Neu	Method	1	Perfunctory	Perfunctory
Ruder et al. 2011	Ack	Neu	Method	1	Perfunctory	Perfunctory
Dunlop et al. 2011	Ack,Ack	Neu,Neu	Method,Method	1	Significant	Perfunctory
Lenhart et al. 2010	Ack	Neu	Method	1	Perfunctory	Perfunctory
Ellison et al. 2007	Ack	Neu	Method	1	Perfunctory	Perfunctory
Mandrusiak et al. 2006	Ack	Neu	Method	1	Perfunctory	Perfunctory
Kuny and Stassen 1993	Ack,Ack	Neu,Neu	Method,Method	1	Significant	Perfunctory
Sobin and Sackeim 1997	Ack	Neu	Method	1	Perfunctory	Perfunctory
Bachorowski and Owren	Ack	Neu	Method	1	Perfunctory	Perfunctory

1995

Sobin and Alpert 1999	Ack	Neu	Method	1	Perfunctory	Perfunctory
Scherer 2003	Ack	Neu	Method	1	Perfunctory	Perfunctory
Goudbeek and Scherer 2010	Ack	Neu	Method	1	Perfunctory	Perfunctory
Katikala-pudi et al. 2012	Ack	Neu	Method	1	Perfunctory	Perfunctory
Moreno et al. 2011	Ack	Neu	Method	1	Perfunctory	Perfunctory
Choudhury et al. 2012	Ack	Neu	Method	1	Perfunctory	Perfunctory
Park et al. 2013	Ack	Neu	Method	1	Perfunctory	Perfunctory
Quercia et al. 2012	Ack	Neu	Method	1	Perfunctory	Perfunctory
Connor et al. 2010	Ack	Neu	Method	1	Perfunctory	Perfunctory
Liakata et al. 2012	Use	Neu	Method	1	Perfunctory	Perfunctory
Navigli 2009	Use	Neu	Results	1	Perfunctory	Perfunctory
Salton and McGill 1986	Use	Neu	Results	1	Perfunctory	Perfunctory
Cowie and Lehnert 1996	Use	Neu	Results	1	Perfunctory	Perfunctory
Sebastiani 2002	Use	Neu	Results	1	Perfunctory	Perfunctory
Pang and Lee 2008	Use	Neu	Results	1	Perfunctory	Perfunctory
Iakovidis and Smailis 2012	Use	Neu	Results	1	Perfunctory	Perfunctory
Vest 2012	Use	Neu	Results	1	Perfunctory	Perfunctory
Dietz et al. 2011	Use	Neu	Results	1	Perfunctory	Perfunctory

Wang and Paul 2011	Use	Neu	Results	1	Perfunctory	Perfunctory
Shneidman and Farberow 1956	Ack,Use	Neu,Neu	Results,Results	1	Significant	Perfunctory
Osgood and Walker 1959	Ack,Use	Neu,Neu	Results,Results	1	Significant	Perfunctory
Gleser et al. 1961	Ack,Use	Neu,Neu	Results,Results	1	Perfunctory	Perfunctory
Edelman and Renshaw 1982	Ack	Neu	Results	1	Perfunctory	Perfunctory
Shapero 2011	Ack	Neu	Results	1	Perfunctory	Perfunctory
Pennebaker and Chung 2011	Ack	Neu	Results	1	Perfunctory	Perfunctory
Pestian et al. 2010	Use,Ack	Neu	Results,Results	1	Significant	Significant
Janssen et al. 2013	Ack	Neu	Results	1	Perfunctory	Perfunctory
Howes et al. 2014	Use	Neu	Results	1	Perfunctory	Perfunctory
Blei et al. 2003	Use	Neu	Results	1	Perfunctory	Perfunctory
Jurafsky and Martin 2008	Use	Neu	Results	1	Perfunctory	Perfunctory
Mitchell 1997	Use	Neu	Results	1	Perfunctory	Perfunctory
Huang et al. 2007	Ack	Neu	Results	1	Perfunctory	Perfunctory
Pestian et al. 2012	Ack	Neu	Results	1	Perfunctory	Perfunctory
Schwartz et al. 2014	Use	Pos	Results	1	Perfunctory	Perfunctory