On the incidence of part-of-speech on polarity identification of user-generated-contents in Spanish

Rafael E. Banchs Joan Codina

Barcelona Media Innovation Centre,
Av. Diagonal 177, Planta 9, 08018 Barcelona, Spain
{rafael.banchs,joan.codina}@barcelonamedia.org

Abstract. In this work, we study the incidence of different part-of-speech over the quality of polarity estimation for opinionated user-generated-contents in Spanish. A dataset of user opinions in the automotive domain was collected from the web and prepared for opinion mining experimentation. Experimental results demonstrate a clear tendency for nouns and adjectives to play an important role in polarity estimation. While this incidence happens to be statistically significant when a dataset of restricted size is used for training the polarity estimation classifier, the empirical evidence suggests that such an incidence fades out when larger datasets are considered.

Keywords: Opinion Mining, Polarity Estimation, Morpho-Syntactic Analysis, Support Vector Machines.

1 Introduction

The increase of user-generated-contents in the World Wide Web has motivated the study and development of data mining techniques for the automatic analysis of this kind of contents. Such a motivation has been solidly grounded, for the case of textual contents, in both economical and scientific interests, and has generated a new field of study denominated opinion mining and/or sentiment analysis [1], [2], [3], [4], [5], [6]. Most of the research carried up to this date in this specific area has been mainly focused on languages such as English [7], [8], [9], and Chinese [10], [11], among others, but, as far as we know, very little work has been conducted for the case of the Spanish language [12]. In this sense, one of the main motivations of this work has been the collection and annotation of an appropriate dataset for opinion mining experimentation in Spanish. Additionally, this work also intends to study the incidence of the Spanish morphology and syntax on the specific opinion mining task of polarity estimation.

The paper is structured as follows. First, in section 2, a detailed description of the corpus collection and preparation processes is presented along with the main corpus statistics and the morpho-syntactic annotation process carried out. In section 3, the experimental methodology is described and the experimental results are presented and discussed. Finally, in section 4, the most relevant conclusions and guidelines for further research are presented.

2 Dataset Collection and Preparation

In the web, there are many places were opinions are given by users over different products and services; one of these sites is www.ciao.es. This website has a different web for each country/language, and was chosen because of the quantity of opinions and the fact that opinions include both, a global numeric rating and some specific ratings over particular attributes of the subject matter. So, with the intention of constructing an experimental dataset in Spanish for opinion mining purposes, the automotive section of Ciao was crawled.

As a drawback, there are two elements that may condition the users when writing opinions in Ciao's website: first, there is a minimum opinion-length restriction, this can provoke the writing of off-topic text just to reach the minimum size; and, second, the users can get money from their opinions, so some users write opinions over products they do not have or do not know very well. Of course, as in any opinion website, there may be false users that just do marketing, but this is a more general Internet issue not only related to Ciao.

First, section 2.1 describes in detail the crawling conducted. Then, section 2.2 describes the selection and preparation of the final dataset to be used in the experiments. Finally, section 2.3 presents a brief description of the morpho-syntactic annotation procedure that was used.

2.1 Crawling of the Dataset

Once the decision to crawl the opinions within the automotive section of the site www.ciao.es was taken, we started the study of the site and the strategies to follow to extract the opinions. First, a way to find all links to the opinions within the section was needed, and then a strategy to extract the different fields from each opinion.

For the first step, we saw that opinions are attached to car models, so first we needed to list the pages of the car models. By using Linux tools like *wget* and *grep*, a list of pages containing car models was generated. Starting from the car models initial pages (more than 5,000) each linked page was crawled to extract the opinions. This was done using a *php* script that, using the DOM model of the page, extracted the different fields from each model. Then, following the list of opinions for each specific model, the information about the individual opinions was extracted.

All this information was introduced in a relational database. The information available for each model included: the model name, ciao Id, evaluation, description, time in ciao, minimum price, maximum price and URL; and for each opinion: Id, title, textual content, rating, user name, the most positive, the most negative, recommend (if the user finally recommend to buy the product), helpful (if other users find the opinion helpful), date, user Id, time of crawling, and the number of comments.

The crawling was performed in February 2009, and contains 25,330 comments of 2,326 car models. The remaining car models did not have any opinion at the moment of the crawling. The ratings are highly biased towards positive polarity: 54% of the opinions give a maximum rating of 5, 32% a rating of 4, 8% of 3, 4% of 2 and only 2% of 1. The average number of words per comment is 250.

2.2 Preparing an Experimental Dataset for Opinion Mining

As already described in the previous section, a total of 25,330 comments related to the automotive domain were collected during the crawling phase. However, this raw collection of comments did not constitute an appropriate dataset for opinion mining experimentation basically because of the following two reasons:

- The collected data contained a lot of corrupted characters due to encoding problems. Most of these corrupted characters were directly related to specific Spanish characters, and others were related to some specific html encoding sequences. Usually, these characters appear when the users edit the opinion in an external editor and then perform a cut & paste into the web page.

- The collected dataset was severely biased towards the positive-polarity. From the total amount of 25,330 comments comprising the collection, there were 21,702 comments with ratings equal or greater than 4. This constitutes the 86% of the total amount of collected comments.

In order to extract an appropriate experimental dataset from the collected data, two preprocessing steps were implemented. First, the entire collected corpus was semi-automatically post-edited for correcting all corrupted characters that were detected. A total amount of 174 Spanish- and 59 html-related different encoding problems were corrected.

Table 1 presents some examples of the most common types of corrupted Spanish characters and html sequences that had to be corrected in the crawled corpus of comments.

Table 1. Examples of the most common types of corrupted characters and html sequences that were corrected.

Spanish Characters		Html Sequences	
corrupted	corrected	corrupted	corrected
Ã¡	á	&#149;	*
Ã©	é	&#324;	ñ
Ã-	í	&#733;	‘
Ã³	ó	&#8212;	“
Ãº	ú	&#9484;	*
Ã±	ñ	&#9668;	–

Second, a subset of comments was selected from the post-edited collection such that the balance between positive and negative polarities was guaranteed. The resulting subset contained a total amount of 3,060 comments. All the 1,530 comments with ratings below 3 were included as negative-polarity data samples, and a random selection of the same size was extracted from those comments with ratings equal to 5. This subset was the one actually used as an experimental dataset.

Table 2 presents the distribution of comments per rating for both, the originally crawled collection and the balanced dataset that was finally selected for experimental purposes.

Table 2. Distribution of comments per rating for both, the crawled collection and the selected experimental dataset.

Ratings	0	1	2	3	4	5
Collected Comments	2	561	967	2098	7951	13751
Selected Comments	2	561	967	0	0	1530

2.3 Morpho-syntactic Annotation of the Dataset

First, all comments constituting the selected experimental dataset were segmented into both, lexical units and sentences, by using the segmentation approach based on maximum entropy algorithms that are available under the OpenNLP toolkit [13].

In second place, part-of-speech and lemma annotation was carried out for each token identified during the segmentation phase. For part-of-speech annotation, the standard variant of the PAROLE tag set most frequently used for Spanish was used [14], and for lemma assignment, the most probable one was selected in each case. The annotation algorithm used for the annotation was the TreeTagger [15], which is based on decision threes. The specific model used for Spanish annotation was trained with the LEXESP corpus [16] and a Spanish dictionary of lemmas and categories of approximately 970K entries.

3 Experimental Work

The main objective of the experimental work described here is to explore the incidence, if any, of the Spanish morphology and syntax on the task of estimating the polarity of user-generated-contents. For this purpose, two groups of experiments were designed and conducted.

First, section 3.1 presents the most relevant details of the methodological framework; and, second, section 3.2 presents and discusses the experimental results.

3.1 Methodological Framework

For all the performed experimental work, a very standard methodological framework was considered. It can be summarized as follows:

- For the task of polarity identification, a supervised approach was considered. In this sense, binary classification engines were implemented by using Support Vector Machines [17].

- For feature space model representation of the data, the popularly known vector space model was used [18], for which the standard normalization and TF-IDF weighting schemes were used [19]. Stopword removal was not performed before computing the vector models in any of the experiments.

- For measuring the performance quality of the binary classifiers, the accuracy metric was considered, which is defined as 100% minus the classification error rate [20].

- For tackling the statistical variability of the data, a multiple fold cross-validation scheme was implemented [21]. For each experimental result reported in this work, twenty independent realizations have been carried out by randomly selecting both, the train and the test, datasets.

Two pairs of experiments were conducted. In the first group of experiments, the incidence of morphology was evaluated. To accomplish this, the accuracy of a reference classifier, which was trained by considering all lemmas in the vocabulary for constructing the vector space models, was compared with a second classifier, which was trained by considering full forms instead of lemmas for constructing the vector space models.

Two experimental datasets were considered for training the two classifiers: the balanced dataset of selected comments that was already described in section 2.2, and a smaller subsample of it. The idea behind this paired set of experiment was to evaluate if the size of the considered experimental dataset has any incidence on the observed results.

Table 3 provides the basic statistics for both, the large and the small, datasets considered in the experiments.

Table 3. Basic statistics for both considered datasets: total amount of comments and token realizations, size of full form and lemma vocabularies, and the average length of comments measured in number of tokens. (K stands for thousands)

Dataset	Polarity	Comments	Tokens	Full Forms	Lemmas	Length
Large	Positive	1,530	448.8 K	23.4 K	15.8 K	293.38
	Negative	1,530	423.4 K	22.4 K	15.1 K	276.75
	Both	3,060	872.3 K	34.6 K	23.3 K	285.07
Small	Positive	241	81.5 K	8.9 K	6.2 K	338.45
	Negative	232	72.9 K	7.1 K	5.0 K	314.38
	Both	473	154.5 K	12.5 K	8.5 K	326.65

In the second group of experiments, the incidence of each specific part-of-speech category was evaluated. To accomplish this, the accuracy of the same reference classifier considered in the first pair of experiments was compared to nine different classifiers for which all the lemmas corresponding to one specific part-of-speech category were removed before constructing the vector space models. Table 4 presents the categories along with their corresponding part-of-speech tags that were removed for each of the nine aforementioned classifiers.

In this second group of experiments, the two experimental datasets described in table 3 were also considered for training the classifiers. In all experiments conducted with the small dataset, balanced train and test sets of 200 comments were always considered. Otherwise, in all experiments conducted with the large dataset, balanced train and test sets of 1000 and 200 comments, respectively, were always considered.

Table 4. Part-of-speech categories and their corresponding part-of-speech tags that were systematically removed for training the classifiers.

POS Category	Identifier	Corresponding POS Tags
Adjectives	ADJ	AQ, AC
Conjunctions	CON	CC, CS
Determiners	DET	DA, DD, DE, DI, DN, DP, DT
Punctuation	PUN	Fa, Fc, Fd, Fe, Fg, Fh, Fi, Fp, Fs, Fx, Fz
Nouns	NOU	NC, NP
Pronouns	PRO	PO, PD, PI, PN, PP, PR, PT, PX
Adverbs	ADV	RG, RN
Prepositions	PRE	SP
Verbs	VER	VA, VM, VS

3.2 Experimental Results and Discussion

The results corresponding to the experiments designed to evaluate the incidence of morphology on polarity identification are summarized in tables 5 and 6. While table 5 presents overall classification accuracy values, table 6 reports precision and recall for both positive and negative classes. The reported figures correspond to mean metric values over the 20 independent simulations conducted in each case. Along with the mean values, standard deviations are reported within parenthesis.

Table 5. Mean accuracy values and associated standard deviations, in parenthesis, from experiments evaluating the incidence of morphology over polarity identification.

Dataset	Reference (lemmas)	Contrastive (full forms)
Small	75.96 (1.52)	77.47 (1.90)
Large	72.25 (2.61)	73.85 (2.51)

Table 6. Mean precision and recall for both, positive and negative, classes from experiments evaluating the incidence of morphology over polarity identification.

Dataset	Class	Reference (lemmas)		Contrastive (full forms)	
		Precision	Recall	Precision	Recall
Small	Positive	74.72 (3.52)	77.90 (4.32)	76.81 (3.59)	78.70 (3.67)
	Negative	77.00 (2.83)	73.30 (5.46)	78.19 (2.69)	76.00 (4.91)
Large	Positive	71.78 (3.60)	73.20 (5.74)	72.29 (3.99)	75.55 (4.08)
	Negative	72.82 (4.49)	71.15 (4.61)	74.38 (4.02)	70.90 (5.18)

As observed from table 5, the mean accuracy values obtained for those classifiers trained by using vector models constructed with full forms are higher than the values obtained when using lemmas. Although these differences between the accuracies are within the range of the estimated confidence intervals, and therefore they cannot be considered significant from the statistical point of view, the observed trend happens to

be consistent for the two considered experimental settings: with large and small datasets. In this sense, these results suggest the possibility that the richness of Spanish morphology could be contributing in a positive manner to the polarity identification task considered here.

Notice also that, according to results presented in table 6, precision and recall exhibit a very similar trend for the two classes under consideration. However, and similarly to what was observed for accuracy, the observed tendency is not statistically significant according to the confidence intervals that have been estimated. Another interesting observation that can be drawn from table 6 is the fact that, for all reported cases, a higher recall is always achieved for the class "positive" while the highest precision is always achieved for the class "negative".

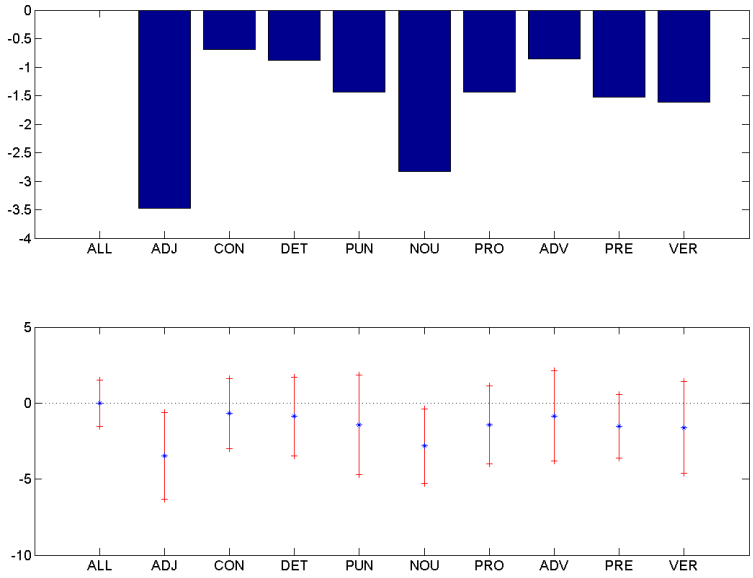

Fig. 1. Results obtained from experiments evaluating the incidence of syntax over polarity identification when considering the small dataset. Upper panel: differences between the mean accuracy values of the reference classifier and the nine contrastive classifiers (which remove one part-of-speech category at a time). Lower panel: estimated confidence intervals.

The results corresponding to the experiments designed to evaluate the incidence of syntax on polarity identification are graphically depicted in figures 1 and 2. Figure 1 presents the results corresponding to those experiments considering the small dataset and figure 2 presents the results corresponding to those experiments considering the large dataset. In both cases, the bar plots in the upper panel represent the difference of mean accuracies between each of the nine classifiers described in table 4 and the reference ones (referred to as ALL in the figures), which results are provided in table 5. The lower panels present estimates of confidence intervals for each result based on the standard deviation of the corresponding simulations.

As seen from figure 1, and according to the confidence intervals estimated from the standard deviation of each set of experiments, adjectives and nouns are the only categories exhibiting a statistically significant incidence on classification accuracy. For the case of adjectives, the task performance is reduced from a mean accuracy of 75.96% in the reference system down to 72.50% when all adjectives are removed from the vector space model representations. This represents an accuracy reduction of 3.46%. Similarly, it is observed that when nouns are not taken into account for vector space model construction, the accuracy drops down to 73.14%, which represents a reduction of 2.82%. These results confirm what the intuition would tell about the most significant syntactic categories for expressing polarity in opinionated contents.

As seen from figure 2, differently from what was observed for the case of the experiments considering the small dataset, no significant incidence on performance is observed for any contrastive system, according to the confidence intervals estimated from the standard deviation of each set of experiments. However, the results provide a clear evidence for a similar trend of adjectives and nouns being the most influential syntactic categories in the problem of polarity identification.

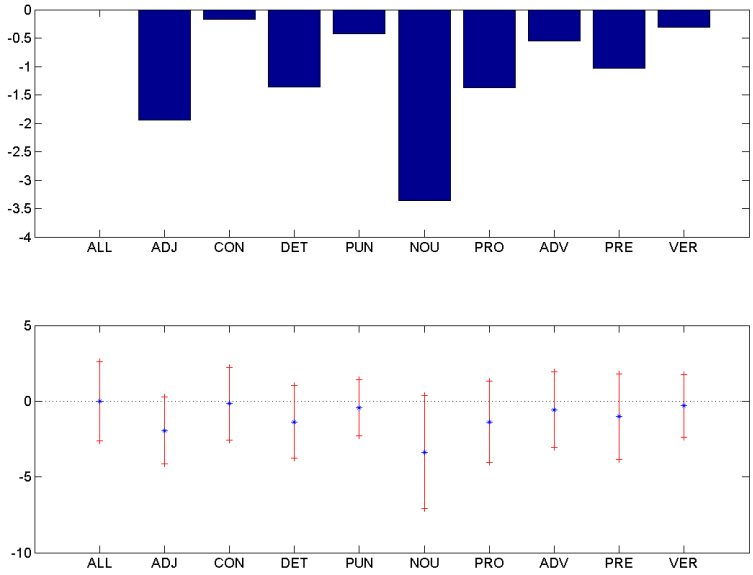

Fig. 2. Results obtained from experiments evaluating the incidence of syntax over polarity identification when considering the large dataset. Upper panel: differences between the mean accuracy values of the reference classifier and the nine contrastive classifiers (which remove one part-of-speech category at a time). Lower panel: estimated confidence intervals.

Moreover, a very interesting additional observation can be drawn from figure 2. Notice how, in this case, the trend between adjectives and nouns has been reverted with respect to the results depicted in figure1. Indeed, when the large dataset was used, while suppressing adjectives from the models produced a reduction of 1.95% in

accuracy (from 72.25% in the reference system to 70.30% in the contrastive one); suppressing nouns drops the accuracy down to 68.88%, which represents a reduction of 3.37% with respect to the reference system. This reverted predominance between adjectives and nouns could be explained by the absolute differences in vocabulary for both syntactic categories between the large and the small datasets. Such differences are presented in table 7.

Table 7. Lemma vocabularies for the two most influential syntactic categories: adjectives and nouns; and their absolute differences between the small and large datasets.

Dataset	Adjectives	Nouns
Small	1,749	4,067
Large	4,743	11,157
Absolute increment	2,994	7,090

As seen from the table, the absolute increment in vocabulary size for nouns between the small and large datasets is more than twice the corresponding absolute increment for adjectives. This could explain the apparent higher impact of nouns on polarity identification for the case of the experiments conducted with the larger dataset.

4 Conclusions and Future Work

In this work, we studied the incidence of morphology and syntax over the quality of polarity estimation for opinionated user-generated-contents in Spanish. A large dataset of user opinions in the automotive domain was collected from the web and two balanced datasets were prepared for opinion mining experimentation.

Although it was consistently observed (for experiments carried out with the two datasets of different sizes) that the Spanish morphology could be contributing in a positive manner to the polarity identification task, the observed differences were not statistically significant according to the confidence intervals that were estimated.

On the other hand, experimental results dealing with the possible incidence of syntax demonstrated a clear tendency for nouns and adjectives to play an important role in polarity estimation. However, while these observations happened to be statistically significant when the small dataset of 473 comments was used, that was not the case when the large dataset of 3,060 comments was considered.

Among the main limitations of the performed experiments and the obtained results we can mention the fact that, although those comments with ratings in the center of the scale where not considered for constructing the experimental dataset, most of the comments collected are generally composed of a mixture of positive and negative opinions up to some degree. Another important limitation of this work is the evident difficulty for obtaining high quality morpho-syntactic annotations for user-generated-contents.

According to this, we intend to concentrate our future efforts in overcoming the main technical limitations described above. First, we will work on improving the quality of the constructed dataset so it can provide a more trustful representation of the two polarity categories in the Spanish language at both, the overall comment and the individual sentence, levels. Additionally, we will invest some important efforts in adapting existent resources, or developing new ones if necessary, for improving the quality of morpho-syntactic annotation of user-generated-contents in Spanish.

Finally, special attention will be paid to evaluating different strategies for dealing with negations, comparisons and conditional clauses, as a possible way for improving the state-of-the-art in the polarity estimation task. Additionally, we will also work on the related problems of subjective/objective nature identification and information extraction for opinion summarization.

Acknowledgments

This work has been partially funded by the Spanish Department of Education and Science through the "Ramon y Cajal" fellowship program. The authors also want to thank Dr. Carlos Rodríguez Penagos for providing the morpho-syntactic annotation tools and scripts, as well as Barcelona Media Innovation Centre for its support and permission to publish this research.

References

1. B. Pang, L. Lee, S. Vaithyanathan, 2002, "Thumbs up? Sentiment classification using machine learning techniques", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
2. B. Pang, L. Lee, 2008, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1-2.
3. R. Tong, 2001, "An operational system for detecting and tracking opinions in on-line discussions", in *Working Notes of the SIGIR Workshop on Operational Text Classification*.
4. P. Turney, 2002, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.
5. J. Wieb, E. Riloff, 2005, "Creating subjective and objective sentence classifiers from unannotated texts", *Computational Linguistics and Intelligent Text Processing*, pp.486-497.
6. T. Wilson, J. Wiebe, P. Hoffmann, 2009, "Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis", *Computational Linguistics*, Vol. 35, No. 3.
7. M. Hu, B. Liu, 2004, "Mining and summarizing customer reviews", in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
8. A. M. Popescu, O. Etzioni, 2005, "Extracting product features and opinions from reviews", in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*.
9. D. Rao, D. Ravichandran, 2009, "Semi-supervised polarity lexicon induction", in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

10. W. Du, S. Tan, 2009, “An iterative reinforcement approach for fine-grained opinion mining”, in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
11. Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, Z. Su, 2008, “Hidden sentiment association in Chinese Web opinion mining”, in *Proceedings of the International World Wide Web Conference*.
12. A. Balahur, E. Boldrini, A. Montoyo, P. Martínez-Barco, 2009, “Opinion and generic question answering system: a performance analysis”, in *Proceedings of the ACL-IJCNLP*, pp.157-160.
13. J. Baldridge, T. Morton, G. Bierner, 2002, “The OpenNLP maximum entropy package”, *Technical Report*, available at http://maxent.sourceforge.net/, Sourceforge.
14. M. A. Martí, M. Taulé, 1998, “Documentación sobre el analizador morfológico MACO”, *Technical Report*, available at http://nlp.uned.es/~anselmo/catalogo_rile.html#MACO, Universitat de Barcelona.
15. H. Schmid, 1994, “Probabilistic part-of-speech tagging using decision trees”, in *Proceedings of the International Conference on New Methods in Language Processing*.
16. J. Carmona, S. Cervell, L. Márquez, M. A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, J. Turmo, 1998, “An environment for morphosyntactic processing of unrestricted Spanish text”, in *Proceedings of the First International Conference on Language Resources and Evaluation*.
17. T. Joachims, 1998, “Text categorization with support vector machines: learning with many relevant features”, in *Proceedings of the European Conference on Machine Learning*.
18. G. Salton, A. Wong, C. S. Yang, 1975, “A vector space model for automatic indexing”, *Communications of the ACM*, Vol. 8, No. 11.
19. G. Salton, C. Buckley, 1988, “Term-weighting approaches in automatic text retrieval”, *Information Processing and Management*, Vol. 24, No. 5.
20. D. Olson, D. Dursun, 2008, *Advanced Data Mining Techniques*, Springer.
21. R. Kohavi, 1995, “A study of cross-validation and bootstrap for accuracy estimation and model selection”, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.