Aspect-based sentiment analysis of conference review forms with LD-enabled review criteria

Sára Juranková[1], Vojtěch Svátek[1] and Chiara Ghidini[2]

[1]*Prague University of Economics and Business, Czech Republic*
[2]*Fondazione Bruno Kessler, Trento, Italy*

Abstract

Conference (or journal) review forms represent a particular kind of sentiment-bearing documents, which has been to date left largely unexplored. We set up a review corpus including a subset or a whole of anonymized conference reviews from four conferences. Furthermore, by a thorough manual analysis of the review forms and guidelines of eleven conferences we identified a set of generic review criteria, which we used as aspects in aspect-based sentiment analysis (SA) of the review form texts. A sentiment lexicon was created specifically for the domain of conference paper reviews. The first step of the lexicon construction method was the manual definition of a two-level taxonomy. In the second phase, noun phrases and adjectives frequently appearing in the given portions of the review were automatically collected. The results of the lexicon-based SA method were compared with the numerical scores from the reviews. The precision of criterion identification was evaluated at 57.38% and the recall at 53.44%; the sentiment polarity was correct in over 75% of cases. This is an improvement on the result of a similar sentiment analysis carried out in a comparable study.

Keywords

aspect-based sentiment analysis, review form, lexicon, linked data

1. Introduction

Peer-reviewing is a crucial ingredient of sharing the results of scientific research. The reviews assure that the authors receive feedback to their research methods and results, and that the quality of papers eventually published is adequate, thus avoiding the waste of time of the future readers. In the context of this paper we treat scientific reviews as a specific category of *opinion-expressing documents*, which can be subject to NLP-based sentiment analysis. Historically, scientific reviews have been predominantly closed data. They have been hard to access beyond the direct stakeholders, i.e., the reviewers, authors, and the 'mediators' (i.e., various conference chairs, journal editors, graduation committee chairs etc.). This is likely the main reason why the attempts to apply sentiment analysis on scientific review forms are rare in the literature. Actually, we are only aware of a single published study, by Bucur et al. [1], from 2019, only operating on a very small review sample. We however believe that with the increasing pressure on the transparency of the review process, leading to an increase of venues with (mostly anonymized

Workshop on Sentiment Analysis & Linguistic Linked Data (SALLD-1), September 1, 2021, Zaragoza, Spain
sara.jurankova@gmail.com (S. Juranková); svatek@vse.cz (V. Svátek); ghidini@fbk.eu (C. Ghidini)
https://nb.vse.cz/~svatek/welcom_e.htm (V. Svátek)
0000-0002-2256-2982 (V. Svátek); 0000-0003-1563-4965 (C. Ghidini)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

but) open reviewing procedures, the amount of available corpora will rapidly increase. At the same time, there is likely to be demand for sentiment analysis applications in the scientific review domain. In our prior work [2] we demonstrated how the partial numerical scores could be served to a relevant stakeholder such as a meta-reviewer in a pictorial form, through an entertaining metaphor: the individual criteria are represented as components/features of a car (their size/style reflecting the particular score). This representation aims to provide a rapid overview and comparison of different reviews of one paper, on the top of a dense table with several dozens of numerical scores. While such a *visualization method* is currently only possible for those review forms that explicitly capture the partial scores in a numerical form, an adequately reliable NLP-based *polarity detection* method would extend its scope to forms in which only the textual assessment is filled for those criteria, and in combination with a specialized *aspect categorizer*, even to completely unstructured review forms. Aside such use cases, addressing the level of individual papers, we might also consider the exploitation of automatically estimated review criteria scores for *aggregate analyses*, e.g., comparisons of different venues with respect to how much attention is paid to different criteria.

Consequently, we focused on both polarity detection and aspect categorization at the same time. Of the large space of reviewed scientific publications, in our research we narrowed down the scope to *conference* reviews, primarily based on the assumption that their review forms are (at least, in the studied community-specific fragment) more likely to be structured by partial criteria ('aspects') than the review forms of journals.

The implementation described in this paper is publicly available at https://github.com/jurs02/aspect-based-sentiment-analysis-of-conference-submission-reviews. More details on the implementation and experiments are also available in a thesis [3].

While the presented research contains a specific empirical contribution (experiments with lexicon-based detection of both aspect and polarity), we also view it as one of the very first attempts to explore a to-date nearly unknown ground of sentiment analysis application, and thus foresee that even the conceptual thoughts about this domain (not fully matching the actual proof-of-concept experiments) might be found relevant by the community.

2. Structure of conference paper reviews and its linked data representation

Different conferences structure their web-based review forms differently. Generally, the overall evaluation and reviewer's confidence are expressed using a fixed choice. As regards partial evaluation criteria, some forms also have a fixed choice for (at least some of) them, while others merely expect the reviewers to mention them in the form of an unstructured text. In some cases the comments for the partial criteria can be input to dedicated textual fields, too. Yet another alternative is to separate the comments not by the criteria but by their polarity, with a specific textual field dedicated to strengths vs. weaknesses of the paper.

2.1. Micro-study in the semantic technology domain

In our research we focused on conferences from the semantic technology (ST) field. We analyzed the review forms of nine ST conferences belonging to the most respected in the ST field (in the alphabetic order): ECAI, EKAW, ESWC, FOIS, IJCAI, ISWC, K-CAP, KR and SEMANTiCS, always for the latest edition we could access as author/s or reviewer/s. We semantically clustered the field labels (referring to the reviewer guidelines where in doubts), yielding seven partial *review metrics* that we named as follows (examples of alternative noun phrases from actual form fields names being in parentheses):

- Relevance (appropriateness)
- Novelty (originality; innovation; innovativeness)
- Technical quality (technical soundness and depth; correctness and completeness of the solution; scientific or technical quality; implementation and soundness)
- State of the art (scholarship; references; related work)
- Evaluation (reproducibility and generality of the experimental study)
- Significance (impact)
- Presentation (clarity and quality of writing).

There are also two global metrics, *Confidence* and *Overall score*, present in all forms. The partial criteria converged well despite the varying wording, though some forms missed certain metrics at all.[1] Some conferences (ISWC and SEMANTiCS) were clearly influenced by one another, having the same set of fields. ESWC had two fields that we both ranged under 'Technical quality': "Correctness and completeness of the proposed solution" and "Demonstration and discussion of the properties of the proposed approach". EKAW only had one partial numerical field (relevance), while further information was collected in verbal form through the free-text fields 'Reasons to accept' and 'Reasons to reject'. K-CAP did not have any partial numerical field at all; this may be related to the 'workshop flavor' of this event. Three of the forms also had a field for 'best paper award' and three had field/s for redirecting the paper to another track (mostly, poster/demo).

This analysis revealed that in a particular research domain there may a wide variety of review structuring degrees while the set of review criteria may be very similar across different conferences. This seems to open the way to meaningful application of *aspect-based sentiment analysis*: the more structured forms, with numerical values for particular criteria, can serve as training data, and the built sentiment analysis models can then be applied – through a *mapping* between the review form fields and the unified criteria – to conferences that only have text-only fields. However, the sentiment analysis (polarity detection) also needs to be accompanied with review text classification to individual criteria, as a kind of *aspect classification*.

In the next short paragraph we report on a prototype solution of the criteria mapping problem (already presented as part of a demo at the ISWC'20 conference). A proof-of-concept study in both polarity detection and aspect classification for review forms is then the main subject of this paper, elaborated in the remaining sections.

[1] A complete table is in a previous paper (which presented the study in a briefer form and without relation to sentiment analysis) [2].

2.2. Linked data infrastructure for review result sharing

We prototyped an ontology that supports the publishing of metrics and their relationships to review forms. The ontology is online at http://kizi.vse.cz/pictoreview/ontology/, and contains the classes *ReviewMetrics*, *ReviewForm*, *ReviewFormField* and *F2M_Mapping* (for the field-to-metric mapping), plus the connecting properties. The proposed metrics set (applicable on ST conferences, and probably many other computing field's ones) is at http://kizi.vse.cz/pictoreview/metrics/. Finally, a sample mapping is at http://kizi.vse.cz/pictoreview/map/semantics18/. We also developed a tool allowing the user to create the *mapping* from the custom set of review form fields of a particular event to the proposed set of generic metrics. The mapping can be stored as a JSON structure or as an RDF dataset described by our ontology. More details are in a previous paper [2].

3. Prior research on conference paper review analysis

The research by Bucur et al. [1] focused on sentiment analysis in reviews of scientific publications. In this work a dataset of eleven reviews, each of which was manually annotated by their respective authors, was used. The aspects of the reviews were syntax, style and content. Each reviewer was asked which of these aspects a specific comment in their review focused on, whether the comment was positive or negative, whether an action by the author of the paper was required or just suggested, what was the impact on the overall quality of the paper and whether the author addressed the point raised in the comment. In the eleven reviews there was a total of 421 review comments, most of which (around 44 %) targeted a paragraph or an even smaller part of the paper, almost 30 % were about the paper as a whole and 27 % of comments focused on a section of the paper.

The result of the annotating phase done by the model experts (the authors) shows that most of the review comments are about the content of the paper, with much smaller percentages being comments about the style or the syntax. Also the amount of negative comments far exceeds the number of neutral or positive comments.

They have applied 18 different lexicon-based sentiment analysis tools to compare the results and found that the best performing tool was the SOCAL method [4] with a maximum accuracy of 72.8 %. Most of the methods performed quite poorly according to the authors, however they discovered that the methods with more complex rules performed the best, even if the size of their sentiment lexicon was not large. It is important to note in the context of this work that the sentiment analysis was done separately from the aspects purely focusing on the polarity of the comment and they did not develop or used any tools to automatically determine the aspect of the comment.

Another research project that is focused on the processes of scientific publishing and more importantly peer reviewing of these publications builds off of the last paper, however this study is focused on creating a unified model for representation of publications and their assessments *"as well as the involved processes, actors, and provenance in general"* [5, p. 1] in the format of linked data. Their vision is that to give more context to reviews, by linking them to other data, such as information about the reviewer and about the author of a paper, as well as provide a way to link specific parts of a review to the part of the paper they comment on.

4. Data collection and pre-processing

The data analyzed in this work are reviews of submissions to semantic technology conferences, namely, editions of a subset of those from listed in Section 2.1: EKAW 2018, ESWC 2018, ESWC 2019, ISWC 2017 and ISWC 2018. The data for each conference were sourced differently, mainly because with the exception of the ESWC 2019 conference, these reviews are not publicly available. All but the ESWC 2019 data were anonymized; the reviewers were previously asked for an explicit consent. Technically, the contact to the authors and the extraction of the reviews were handled by co-authors of this papers, who had served in a senior role (program chair / meta-reviewer) in the respected conferences. Altogether, 247 reviews were collected this way for EKAW 2018, 11 for ISWC 2017, and 20 for ISWC 2018, and 6 for ESWC 2018. The ESWC 2019 data was publicly available through a SPARQL endpoint at https://metadata.2019.eswc-conferences.org/sparql (unfortunately, not functional any longer). Preprocessing for *aspect vocabulary extraction* consisted in word- and sentence-level tokenization and POS tagging, using standard NLTK library tools, and lemmatized using the `WordNetLemmatizer`. No stop words were removed.

Preprocessing for *sentiment vocabulary extraction* was carried out over 1000 review sentences from the ESWC 2019 dataset. Sentences with dual polarity in a single example were manually split in order not to confuse the Naïve Bayes classifier. Each sentence was labeled with either positive or negative sentiment by the author as well as another annotator, independently as to not influence each other. There were eight disagreement cases, which were harmonized during a consensus-reaching session. The resulting dataset consists of 743 negative and 257 positive sentences.

When creating the lexicon, first all contractions are expanded, then the review is tokenized into words and assigned a POS tag. Because only words with a high enough frequency are kept, it was also decided to remove stop words, based on a custom stop word list. The NLTK corpus also includes a dictionary of stop words, however it includes words that were expected to have a noticeable influence on the polarity of a sentence such as *should* which rarely points to a positive sentiment in reviews. The tokens were lemmatized, with the exception of adjectives. In this task especially adjectives needed to be kept in the same form as they were originally written in the review. For example, the distinction between *good* and *better* might be important for the polarity of the adjective as *better* is more likely to be used in a negative comment such as *"it would be better if you…"* while *good* mostly keeps its positive polarity.

Preprocessing for the *evaluation of results* consisted in the annotation of 15 reviews from 3 different conferences (5 from each), namely ESWC 2019, ISWC 2018 and EKAW 2018. The reviews were labeled by two annotators, with criteria and sentiment polarities found in each review comment. Because the labels of a criterion to which the reviewer points to in a comment very often differed across the two sets of annotations (while both annotators mostly agreed about the sentiment), a discussion between the annotators again ensued, to reach a consensus. Out of 136 annotated comments the annotators did not originally agree in 49 cases when it came to the criteria, which is over one third. In 14 out of those 49 cases the annotators did not reach a consensus even after the discussion. For example, regarding the comment *"Then it will be beneficial to provide a justification of the number of entities… used in the experiment."* one annotator argued it should be labeled with the *evaluation* criterion, as the reviewer questions

the small data sample used in the evaluation of the work. The other annotator however insisted on the *presentation* label, based on the fact that the reviewer had asked for a clarification on the sample size without outright criticizing it.

5. Aspect extraction

To extract aspect expressions, a taxonomy-oriented approach is used, where the aspects at the top of the hierarchy represent the criteria and all the terms at the second level represent aspect expressions belonging to the criteria. The chosen set of criteria (aspects) corresponded to those from our study described in Section 2.1, except that Novelty and Significance were combined into one criterion. In order to create a lexicon of terms that represent the chosen set of criteria, to be used for identifying aspect expressions in the reviews, it was decided to use two main approaches. One is the taxonomy extraction and the second one is extraction of frequent words used by the reviewers for different criteria in a text that is already divided by headers into sections for the respective criterion. The manually created seed taxonomy had the criteria (aspects) at the first level and the aspect expressions for each criterion at the second level:

- 'relevance': 'appropriateness', relevance'
- 'novelty & significance': 'originality', 'innovativeness', 'innovation', 'novelty of contribution', 'novelty', 'impact', 'significance'
- 'technical quality': 'scientific quality', 'implementation', 'soundness', 'technical quality'
- 'state of the art': 'scholarship', 'references', 'related work', 'state of the art'
- 'evaluation': 'reproducibility', 'evaluation', 'evaluate', 'evaluating'
- 'presentation': 'clarity', 'quality of writing', 'presentation', 'typo', 'description', 'describe', 'written'

5.1. Crude features extraction

The next step of taxonomy based extraction was to obtain a set of crude features.

The first method of extracting terms that are likely to represent a criterion is to extract frequent *noun phrases* (NP). For that the content of the file (representing a single review) was tokenized and each token was assigned a POS tag. The tuples (token, pos_tag) were then parsed to determine the multi-token sequences which represent nouns and noun phrases. The RegexpParser from the NLTK library was used with a simple custom grammar (over the POS tags) covering both the development of noun phrases from nouns through adjectives and the composition of noun phrases through prepositions. The tree returned by RegexpParser was traversed. For each subtree, labeled as NP, each token of the sequence was lemmatized, it was checked if its length was between 2–20 characters, and if so, the lemmatized tokens were joined into a single string. To obtain the NPs that were frequent enough across all the reviews and may therefore represent the criteria, the support of a NP across the reviews was calculated as $support(w_i) = \frac{N_{w_i}}{N}$, where w_i is a word, N_{w_i} is the number of reviews containing w_i, and N is the total number of reviews. Only if the support was greater than the minimum support, the

criterion expression candidate was kept. Based on experiments, the minimum support threshold of 2 % proved as one returning a reasonable amount of candidate criterion expressions with respect to their manual confirmation.

When going through the training data, it was then discovered that fairly often, the criterion expression found in the reviews takes the form of *adjectives*. It was decided to extract frequent adjectives from the reviews as well, and calculate the support of each adjective across the reviews with the same technique as was used with the NP extraction.

5.2. Extraction by review structure

The data from the 2018 ESWC conference that was obtained had the review text divided into sections where the different sections represented the different criteria. It was decided to leverage this data to extract frequent words from each of these sections across the reviews. The frequent words of each section were included in the new aspect expression taxonomy directly; they did not go through the same process of similarity matching against the manually created taxonomy as the candidates that were chosen purely on their frequency. This is useful because it allows to extract new possible criterion terms that would not match with any of the terms in the taxonomy, which were originally not thought to be included. However, the aspect expression candidates created by this method were still evaluated by the user.

5.3. Validation against the manually created taxonomy and by the user

The similarity of crude features to the aspect expressions in the taxonomy was calculated, and only those with sufficient similarity were retained. The NLTK WordNet tool and its `path_similarity` metric, based on the shortest path that connects the senses in the WordNet hypernym hierarchy, was used. Since the features may contain adjectives (in the case of noun phrases) or be adjectives themselves (in the case of frequent adjectives extraction), which are not a part of the WordNet hierarchy, the adjectives were transformed to their closest related noun. Another issue is measuring similarity with terms consisting of multiple words. Certain multi-token terms are already present in the WordNet thesaurus (such as *state of the art*), and getting their synsets to perform similarity matching is as simple as replacing the spaces between words with underscores. However, some multi-token words cannot be found in WordNet directly. This problem was solved by calculating the maximum similarity between each token of one term to all the tokens of the other term. Of these similarities the maximal one is chosen. The threshold for similarity of a term to the taxonomy was set to 0.3; every term having at least this similarity to any term in the manually created taxonomy became a criterion expression candidate under the same criterion as the term it was most similar to.

The second filtering step consisted in an interactive process of the user validation of the final taxonomy consisted in asking the user 1) whether the given term belongs under the estimated aspect, and if not, under which other aspect (if any) it belongs.

The entire process of aspect expressions extraction and user validation (carried out by the first author of this paper) resulted in 57 new terms in the taxonomy. The new terms added under each criterion are showcased in Table 1 along the original terms from the manually created taxonomy.

Table 1
Result of the aspect expression extraction

criterion	aspect expressions - old	aspect expressions - new
relevance	appropriateness, relevance	important topic, relevant, contribution, topic
novelty	originality, innovativeness, innovation, novelty, novelty of contribution, impact, significance	originality innovativeness, scientific contribution, improvement, novel, idea
technical quality	scientific quality, implementation, soundness, technical quality	running example, scalability, code, design, usability, implementation and soundness, technical detail
state of the art	scholarship, references, related work, state of the ar	reference, related work section, references, benchmark, comparison, previous work, related research
evaluation	reproducibility, evaluation, evaluating, evaluate	evaluation section, coverage, experimentation, score, experimental result, experimental, experimental evaluation, support, empirical evaluation, accuracy, assessment, evaluation result, recall, metric, experiment
presentation	clarity, quality of writing, presentation, typo, written, describe, description	english, scientific paper, notation, text, last sentence, sec, write, figure, introduction, document, explained, current form, reading, writing, first paragraph, intro, readability, format, paragraph

6. Creation of sentiment lexicon

Through some initial experimentation with various existing sentiment lexicons such as the SenticNet sentiment lexicon and the NLTK's SentiWordNet sentiment lexicon it was discovered that these universal sentiment lexicons are not well suited for application on the domain of conference paper reviews.

One issue is that in the specific domain of scientific reviews, some words having neutral polarity in generic lexicons are rather polarized. For example in SenticNet, the word *clarification* has polarity of -0.09, but it is often used in sentences such as *"The section about experimental results needs some clarification"* where the sentiment is clearly negative.

Another possible issue with pre-made sentiment lexicons is that they mostly do not include punctuation. However, punctuation might have great semantic significance in conference paper reviews, as they are often written in plaintext and punctuation is used to compensate for the lack of usual formatting styles such as bullet points.

For that reason, it was decided to create a custom, domain-specific sentiment lexicon. The

NLTK's implementation of the Naïve Bayes classifier was used for this purpose. Of all the lemmatized tokens in the reviews, the number of tokens the classifier needed to process as features was limited to 450. The dataset of labeled review sentences was split to have a testing dataset of 50 example sentences to determine the accuracy of the classifier and the rest of sentences was used for training. The accuracy of the classifier on the testing dataset was 0.78. The classifier allowed us to get a list of features which had the highest contribution to classification. The highest contribution to positive classification was reached (in turn) by the terms 'easy', 'interesting', 'topic', 'sound', 'community', 'idea', 'interest', 'good', 'clearly', 'well', 'bring', 'valuable', 'conference', 'write', 'effort' and 'highly', which had the ratio of usage in positive vs. negative sentences between 36 and 5. For negative classification it was 'but', '?', 'what', 'not', 'me' and 'why', with the analogous ratio (negative vs. positive sentences) between 15 and 5. The interpretation is in most cases intuitive; compared to commonly used polarity lexicons, the role of pronouns, particles and punctuation might be worth mentioning.

Each of the most informative tokens was given a value of -1 or +1 (where -1 corresponds to negative sentiment and +1 corresponds to positive sentiment). The top 100 tokens were eventually chosen to be included to the sentiment lexicon (setting the ratio threshold to 2.4 to 1). Then the results were compared with the SenticNet sentiment lexicon, to see what is the level of agreement between the two lexicons and it was discovered that in 13 cases, the polarity of the sentiment of words found in both lexicons differed and in 31 cases a word from my lexicon was not found in SenticNet. Surprisingly not all the words that were not found in SenticNet were not found due to the aforementioned lack of punctuation in SenticNet or because these words could truly be considered neutral. SenticNet was missing some words which are considered fairly meaningful in sentiment analysis such as *rather* or *should*.

40 positive and 66 negative words compiled manually during the process of labeling the training dataset were also included. The resulting sentiment lexicon contains 186 sentiment words out of which 88 have positive polarity of +1 and 98 have a negative polarity of -1.

7. Aspect-based sentiment analysis

The high-level categories of sentiment analysis methods are the lexicon-based and the machine-learning-based ones. Most machine learning methods require high amounts of labeled training data, which makes it difficult to transfer a trained model to another domain. Although there are domain adaptation methods, they are primarily focused on sentiment analysis at the document level rather than at the aspect level [6]. Also, for a novel domain like the conference review one, it is desirable to allow for hands-on exploration of the data. Therefore we opted for a lexicon-based approach.

Our sentiment analysis method applied in our project is a variation of the holistic lexicon-based approach [7], which expects an existing list of aspect expressions as well as a sentiment lexicon, and addresses aspect-based sentiment analysis. The basic algorithm of the holistic approach finds all words or phrases describing features in a sentence as well as opinion (or sentiment) words. Then, for each feature in the sentence, its sentiment score is calculated using the polarity of the opinion words and their distance in the sentence from the feature expression

using the following function:

$$\text{score}(f) = \frac{\sum_{w_i : w_i \in s \wedge w_i \in V} w_i.SO}{dis(w_i, f)} \tag{1}$$

where w_i is an opinion word, V is the set of all opinion words, s is the sentence that contains the feature f, $dis(w_i, f)$ is the distance between feature f and opinion word w_i in the sentence s, and $w_i.SO$ is the semantic orientation of the word w_i. The algorithm is also extended to deal with negation (by negating the polarity of a sentiment word which follows after a negation word), but-clauses (by first trying to determine the sentiment of an opinion word within the but-clause using the basic algorithm and if the sentiment score is zero it assigns the negation of the clause before *but*). Then it has three rules for dealing with context-dependent opinion words based on their co-occurrence with context-independent words and the polarity of the neighboring sentences. The details of the adapted algorithms can be found in the underlying thesis [3].

Polarity detection using aspect expressions and sentiment words First, of all, the sentences with more than 5 aspect expressions are discarded, since for them it would be too hard to evaluate which opinion words belong to which aspect. It is especially an issue with the numerical evaluations at the beginning of reviews, which are sometimes not structured by newlines or any other separator.

Given a sentence s_i that contains a set of aspect expressions, the polarity score is calculated for each expression. Given a set of sentiment words in the sentence, their collective influence is calculated based on the sentiment score given by the sentimentr algorithm which is divided by the distance of the sentiment word from the aspect expression. These scores are then aggregated by a sum function for each aspect expression. An expression is assigned the sentiment polarity using the `orientation_to_interval` function. Its default functionality is to return -1 or +1 if the orientation is positive or negative, and 0 otherwise.

Adjectives as aspect expressions When for some aspect expressions the orientation is unknown after aggregating polarities of nearby opinion words, they are added to s_i's list of non-polarized expressions. These are next evaluated using the adjective rule, where if the aspect expression is an adjective, it may be used as an opinion word itself. This is the case with sentences such as *"This paper is highly relevant to the conference"*, where the adjective *relevant* points to the relevance criterion, but also expresses a positive polarity on said criterion. In cases when the aspect expression is an adjective, its polarity is determined using a re-implemented and slightly modified version of the sentimentr algorithm [8] as if it was an opinion word, to cover cases such as negation. Sentimentr uses four kinds of modifiers: negators, amplifiers, de-amplifiers and adversative conjunctions.

Intra sentence rules If the aspect expression orientation is still unknown after the use of the adjective rule it is evaluated using the intra sentence rules, which rely on the fact that a sentence only expresses one polarity unless it includes an adversative conjunction. For each so far non-polarized aspect expression the closest opinion word is found as well as all the words

between the aspect and the opinion word. If there is an adversative conjunction in between the opinion word and the aspect expression, that likely means the sentiment polarity was inverted by the conjunction and therefore the aspect expression should be given the opposite polarity of the opinion word. This should help in sentences such as *"The evaluation shows great results but the dataset was small."* in which *dataset* might be an aspect expression pointing to the evaluation criterion but *small* might not be in the sentiment lexicon. For a human, it is easy to point to small as a negative word here based on the context, but that is not always the case, for example in a phrase *small error* it would be positive. The closest identified opinion word in this sentence would be *great*, with a positive polarity, but given the fact that there is *but* in between *great* and *dataset*, the polarity assigned would be negated. If no adversative conjunction is found, the aspect expression is given the polarity of the closest opinion word without any changes, following the *one sentence – one polarity* idea.

Sentences with neutral sentiment If the orientation of an aspect expression is still 0 after the application of all rules, it is finally evaluated as neutral. These aspect expression do not influence the numerical scores of a review, the aspects with neutral orientation are however still shown in the algorithm's output at a sentence level.

8. Evaluation of results

To assess the algorithm quality, the numerical scores of each criterion estimated by the algorithm were first compared to the scores given to these criteria by the reviewers. Then, the algorithm's output was evaluated in more detail, by establishing the precision and recall of aspect identification and the accuracy of the sentiment analysis, using an annotated set of reviews. Finally, the results of the evaluation were discussed and suggestions are made for future improvements.

8.1. Evaluation using reviews with numerical scores

The reviews from ISWC 2018 contain numerical scores for a wide range of criteria as was mentioned in section 2. It was decided to compare the numerical scores output by the sentiment analysis algorithm with the ground-truth scores taken from the reviews. Because the ISWC set of criteria is more detailed that the set of criteria the algorithm works with, it was necessary to create a mapping between them which you can see in Table 2.

The numerical output was evaluated using the mean absolute error function (MAE), which measures the absolute average distance between the real data Y and the predicted data $\bar{Y}$ as $MAE = \frac{1}{n} \times \sum_{i=1}^{n} |Y_i - \bar{Y}_i|$. The MAE was calculated separately for each criterion to see if the algorithm performs better or worse for some of them. The default range of [1;5] for scores outputted by the algorithm was matched to the [-2;2] range of the ISWC reviews. Because the algorithm outputs "n/a" instead of a number for criteria for which no sentiment value was found, the number of times the "n/a" value occurs for each criterion was also calculated.

The results of the numerical evaluation carried over the 20 ISWC 2018 reviews can be seen in Table 3. It is clear that the algorithm often struggles with finding any criterion score, especially

Table 2
Mapping between the chosen set of criteria and ISWC 2018 criteria

Algorithm's criteria	ISWC 2018 criteria
relevance	appropriateness
novelty	originality/innovativeness
	impact of ideas and results
technical quality	implementation and soundness
state of the art	related work
evaluation	evaluation
presentation	clarity and quality of writing

when it comes to relevance, novelty and technical quality. Even when it does give a numerical score, it is often fairly off. This could have several explanations. Firstly it is possible that when reviewers have the option of expressing their opinion numerically, they sometimes do not feel the need to also give a more elaborate explanation. The second explanation is that the algorithm simply does not perform well when it comes to discovering aspect expressions and/or sentiment words. This is studied more closely in the next section, where the results are evaluated on the sentence level.

Another issue might be the way in which the numerical scores are estimated – each time an aspect expression is discovered and assigned a polarity of +1 or -1 the value is added to the respective criterion score. Finally the scores are averaged by the number of times a value was added and normalized to a given score range. Therefore, if for a criterion only one aspect expression is found with a given polarity the final score will always be an extreme in the score range, but that is a much rarer occurrence in the scores given by human reviewers. To check if this might be the issue it was tested by changing the normalization of polarity to four different values, where a criterion is added a score of +1 if the orientation of an aspect expression is higher than 0.5, a score of +0.5 if the orientation is higher that 0 and analogously the -0.5 and -1 scores for negative orientations. The results after that change can be also seen in Table 3. It is apparent that this leads to better results and so a more fine-grained approach to polarity is necessary.

Table 3
Results of the numerical evaluation of ISWC 2018 reviews

Criterion	MAE	MAE (more granular)	Number of missing values
relevance	1.44	1.33	11
novelty	1.68	0.93	0
technical quality	1.64	1.09	9
state of the art	1.38	1.06	4
evaluation	1.35	0.82	3
presentation	0.94	0.69	4

8.2. Evaluation using annotated review comments

This section presents the outcome of the evaluation first for the criterion identification and then for the sentiment analysis.

Evaluation of the criterion identification The annotated dataset was compared to the output of the sentiment analysis algorithm. Each annotated comment for which the appropriate criterion was found by the algorithm was classified as *true positive* (TP), each comment which was labeled by a criterion but the algorithm did not discover it was labeled by *false negative* (FN) and each time the algorithm output a criterion incorrectly it was classified as *false positive* (FP).

Some of the reviews contain numerical scores for the extracted criteria, however while the algorithm always correctly identifies the respective criteria, it is unable to assign a sentiment polarity based on the score value. Because each conference might use a different scale for these scores and the algorithm should work independently of any knowledge about the specific source of reviews, it was decided against using the scores for the sentiment analysis and instead only focus on the textual data. Therefore the parts of reviews with numerical scores were ignored for both the annotation and evaluation phase.

The results of the comparison between annotated criteria and outputted criteria can be seen in Table 4. This yielded a precision of 57.38 % and a recall of 53.44 %.

Table 4
Evaluation of the criterion identification

conference	TP	FP	FN
ESWC 2019	17	15	14
EKAW 2018	31	17	34
ISWC 2018	22	20	13
Total	70	52	61

As regards the contribution of each criterion to the result (detailed data and charts are in the thesis [3]), the algorithm visibly struggled the most with discovering the *technical quality* criterion, where the false negatives amount for 61 % of the assigned evaluation labels, 25 % were false positive and the algorithm succeeded in correctly finding the correct criterion in just 14 % of cases. The algorithm also did not fare particularly well for the *state of the art* criterion, where although the relative amount of false negatives is fairly small, the number of false positives is at 62 %, meaning that a considerable amount of sentences get labeled with this criterion wrongly.

In terms of discovering the correct criteria, the *presentation* criterion amounts for 37 % of all true positives and the *evaluation* criterion for 31 %, however in 38 % of all cases that *evaluation* was assigned any of the evaluation labels, 31 % were false positives, which is a fairly high amount even though the relative amount of true positives is 52 %.

Generally a considerable amount of false positives come from the fact that when a reviewer talks about some issue that falls under the *presentation* criterion, they refer to some section of the paper by the section name or topic, which often falls under an aspect expression belonging to some other criterion. Consider the sentence *"The section numbers appear off, too–the introduction says that evaluations are performed in Section 6, for instance, but it's actually in Section 5."*, which

for a human reader obviously mentions a problem with the quality of writing, however the algorithm picks up on the *evaluations* keyword and incorrectly labels the sentence with the *evaluation* criterion.

Another issue that leads to a significant number of false positives is that at the beginning of most reviews, there is a summary of the paper, stating its objectives. These summaries often contain different aspect expressions, as there is an overlap between the criteria names or expressions and the general vocabulary of the field. For example, the sentence *"To achieve this, an inference rule is translated into a SPARQL CONSTRUCT query that is evaluated against the schema of the RDF dataset."* only mentions how the solution proposed in the reviewed paper works, it is not the reviewers comment on the quality of the actual evaluation but it gets recognized by the algorithm as such. Similarly the sentence *"In this paper authors investigate how state-of-the art language technologies LT (tools, algorithms and resources) can be ported to the historical ecology domain."* gets labeled with the *state of the art* criterion, even though the sentence does not refer to how the reviewer feels about the research of the domain done by the authors (which would indeed fall under *state of the art*).

The false negatives mostly come from the fact that the comments rarely included expressions immediately leading to a certain criterion. For instance, for the *technical quality* aspect it would require a significantly more complex domain specific lexicon of technical terms to lessen the amount of false negative in sentences such as *"Even if a rule is determined as potentially applicable after running the query derived from the rule on the data schema, the rule cannot be executed until relevant instances are entered into the dataset."*.

To explain the high amount of missing numerical values discovered in the previous section, the output of ISWC 2018 data was inspected and compared with the annotated dataset. It was discovered that out of the 7 "n/a" criteria values outputted for the 5 annotated reviews 4 criteria were also missing in the annotations (in other words according to the annotators there was no comment related to these criteria). All unknown values from the examined dataset belonged either under the *relevance* criterion (4 unknown values and 3 reviews with no *relevance* annotation) or the *technical quality* criterion (3 unknown values and 1 review with no *technical quality* annotation). As was already explained, the algorithm does not perform well in classifying comments regarding *technical quality* which is likely the cause of the missing values. On the other hand *relevance* has a significantly lower amount of false negatives, so it is more likely that reviewers simply do not feel the need to expand on their numerical score when this criterion is concerned.

Sentiment analysis evaluation The comments with aspects that were correctly classified (those with the TP label) were also evaluated based on whether they were labeled with the correct sentiment. In the annotated dataset there was only one comment where the annotators did not reach an agreement on the appropriate sentiment – *"Though, this approach solves a relevant problem there are several concerns:"*. This was due to that comment containing the previously mentioned dual polarity. This comment could therefore be taken as either positive or negative, and was evaluated accordingly. Over 75.7 % of comments with correctly identified criterion were also correctly classified by sentiment while 24.3 % were not.

Relative to the number of TP comments regarding a certain criterion the algorithm performs

best for the sentiment analysis of the *state of the art* criterion with an accuracy of 100 %, however since only 5 comments belonging to this criterion were correctly identified, the number might not be objectively accurate. Interestingly the algorithm does well at determining the polarity of the *presentation* criterion as the comments tend to contain similar sentiment expression such as *well-written* or *clear* so due to their frequency across reviews they were picked up during the creation of the sentiment lexicon.

8.3. Discussion of results

The comparison between the numerical output of the system with the numerical scores given in the reviews resulted in a mean average error of 0.99 on a scale from -2 to 2, meaning the algorithm was usually nearly one point off. This was deemed a significant error margin and in order to get more insight into the accuracy of the criterion identification and the sentiment analysis a more detailed assessment ensued on a sentence level which estimated the precision of the criterion identification at 57.38 % and the recall at 53.44 %. As such, the error rate is quite high, however even the annotators had a substantial level of disagreement, initially diverging in their criterion labeling in over a third of the comments. This suggests that reconstructing the intended meaning of the review comments is a particularly difficult task even for human annotators.

The accuracy of the sentiment analysis was calculated using a set of comments with correctly identified criterion determined by human reviewers. The system detected the accurate sentiment in more than 75 % of executed cases. Notably, this is an improvement on the result of a similar sentiment analysis carried out in a comparable study with focus on peer reviews of scientific publications. This analysis algorithm only managed to reach an accuracy level of nearly 73 % even though the sentiment analysis performed in this study was not aspect based [1], which makes determining the sentiment easier. Consider the sentence *"While it is a fairly relevant topic, there were too many typos and the results were not at all evaluated against any of the existing state-of-the-art systems, so I do not consider the work mature enough for acceptance."*. The overall sentiment of the sentence is negative, as it presents more reasons for rejecting the paper rather than accepting it and for that same reason it would not be difficult for most dictionary-based sentiment analysis methods to correctly classify the sentence as negative. However, if we classify the sentiment on an aspect level, it is necessary to also find out which sentiment expressions of the sentence belong to which aspect, which in this case means it needs to recognize that even though the sentence contains more negative expressions, relevance is judged positively.

The algorithm is capable of being substantially improved quite easily when given the availability of a significantly larger dataset. This could help put together both a better aspect expression dictionary and a vast sentiment lexicon. The creation of which heavily relies on the frequency of terms across reviews. Additionally, should a larger dataset be available, it would be possible to perform a more detailed exploration of the specificities of language used in similar data, creating a more complex algorithm using the discovered knowledge.

Another possible way to improve the results would be to put together specialized rules for the different criteria. For example, as comments about the *presentation* criterion often lead to false positives matches on other criteria, it would be possible to create a rule where criteria expression found in a sentence where there is also a match on *presentation* would be discarded. Another

example of a criterion-specific rule would be to study the structure of a sentence in a more complex manner by using syntactic analysis, to for instance discover that in a sentence such as *"The rule cannot be executed until relevant instances are entered into the dataset."* the adjective *relevant* is connected to *instances*, which is not a term related to the work itself and therefore the sentence should not be considered as a comment on the *relevance* criterion. However, these rules would require to limit the algorithm to a specific set of criteria, which goes against the original idea to create an algorithm that could be reused on a number of different domains.

The sentiment analysis as well as criterion identification could also be enhanced by creating a sentiment lexicon specific to each criterion. Firstly, there are many sentiment expressions that are context-dependent: For example, *small* would most likely considered a positive word in relation to an evaluation error but a negative one when commenting on the contribution of the work. This problem could be solved by having a dedicated sentiment lexicon for each criterion, and thus providing the necessary context. Secondly, some sentiment expressions have a strong connection to certain criteria. For example, *clear* is almost exclusively used in relation with the *presentation* criterion. This could be leveraged to aid in aspect identification, as a sentiment expression which is strongly related to a particular criterion could be used as an indication of its presence, if no criterion expression is found.

9. Conclusions

The aim of the research was to explore the possibility of extracting sentiment from conference paper reviews. The evaluation of sentiment analysis accuracy shows an improvement on the results of similar existing research and so with indicated improvements the system is going to be a valuable tool for helping to facilitate the meta-reviewing process. It can also help with the unification of criteria scores across different conferences and reviewers using the numerical scores outputted by the system.

All the algorithms created in this work – the aspect expression identification, sentiment lexicon compiler and aspect-based sentiment analysis – are implemented in a way which would require slight adjustments for application on a number of different domains of conferences. The main change that would be required is to manually create a new taxonomy of aspect expressions serving as a base for identification of other aspect expressions, helping to define the set of the domain-specific criteria. Allowing for minor adjustments, the practical use of the system in other domains is clear, providing the appropriate datasets to retrain the algorithm for the new application.

Acknowledgments

TThe research has been supported by the IGA VŠE project no. 56/2021 and by the Nexus Linguarum COST Action (no. CA18209).

References

[1] C. Bucur, T. Kuhn, D. Ceolin, Peer reviewing revisited: Assessing research with interlinked semantic comments, in: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019, ACM, 2019, pp. 179–187. URL: https://doi.org/10.1145/3360901.3364434. doi:10.1145/3360901.3364434.

[2] V. Svátek, S. Juranková, R. Šalda, P. Strossa, Z. Vondra, Creating and exploiting the mappings from conference review forms to a generic set of review criteria, in: Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), 2020. URL: http://ceur-ws.org/Vol-2721/paper567.pdf, online, accessed 2-November-2020.

[3] S. Juranková, Aspect-based sentiment analysis of conference review forms, Master's thesis, Prague University of Economics and Business, Prague, Czech Republic, 2021. URL: https://insis.vse.cz/zp/index.pl?podrobnosti_zp=70448;.

[4] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2011) 267–307. doi:10.1162/COLI_a_00049.

[5] C.-I. Bucur, T. Kuhn, D. Ceolin, A unified nanopublication model for effective and user-friendly access to the elements of scientific publishing, 2020. arXiv:2006.06348.

[6] B. Liu, Sentiment analysis: mining opinions, sentiments, and emotions, Cambridge University Press, 2015.

[7] X. Ding, B. Liu, P. S. Yu, A holistic lexicon-based approach to opinion mining, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM '08, ACM, New York, NY, USA, 2008, pp. 231–240. URL: http://doi.acm.org/10.1145/1341531.1341561. doi:10.1145/1341531.1341561.

[8] T. Rinker, V. Spinu, sentimentr, 2016. URL: https://github.com/trinker/sentimentr. doi:10.5281/zenodo.222103.