# Appendices

## A Acceptance Classification Features

Table 7 shows the features used by our acceptance classification model. Figure 2 shows the coefficients of all our features as learned by our best classifier on both datasets.

### A.1 Hyperparameters

This section describes the hyperparameters used in our acceptance classification experiment. Unless stated otherwise, we used the sklearn default hyperparameters. For decision tree and random forest, we used maximum depth=5. For the latter, we also used max_features=1. For MLP, we used $\alpha = 1$. For $k$-nearest neighbors, we used $k = 3$. For logistic regression, we considered both $l1$ and $l2$ penalty.

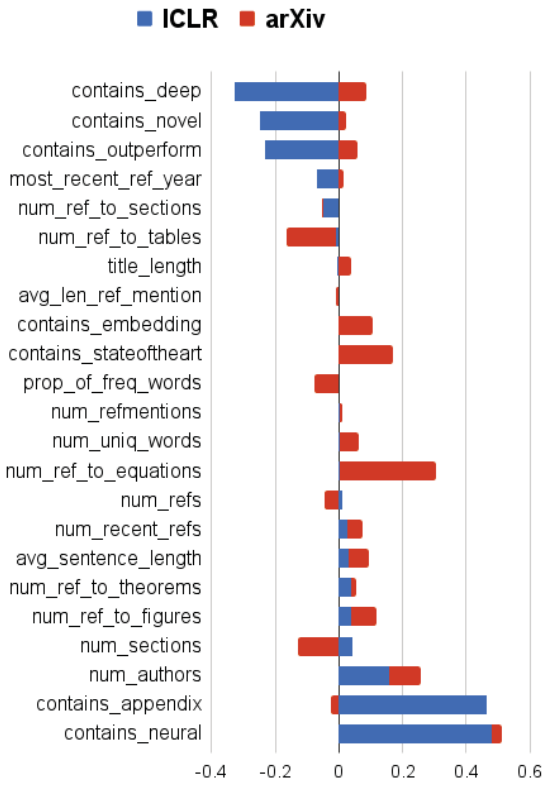


**Figure 2:** Coefficient values for coarse features in the paper acceptance classification, for ICLR and arXiv.

## B Reviewer Instructions

Below is the list of instructions to ACL 2016 reviewers on how to assign aspect scores to reviewed papers.

| | Features | Description | Labels |
|---|---|---|---|
| *coarse* | abstract_contains_X | Whether abstract contains keywords X ⊂ deep, neural, embedding, outperform, outperform, novel, state_of_the_art | boolean |
| | title_length | Length of title | integer |
| | num_authors | Number of authors | integer |
| | most_recent_refs_year | Most recent reference year | 2001-2017 |
| | num_refs | Number of references (*sp*) | integer |
| | num_refmentions | Number of reference mentioned (*sp*) | integer |
| | avg_length_refs_mention | Average length of references mentioned (*sp*) | float |
| | num_recent_refs | Number of recent references since the paper submitted (*sp*) | integer |
| | num_ref_to_X | Number of X ⊂ figures, tables, sections, equations, theorems (*sp*) | integer |
| | num_uniq_words | Number of unique words (*sp*) | integer |
| | num_sections | Number of sections (*sp*) | integer |
| | avg_sentence_length | Average sentence length (*sp*) | float |
| | contains_appendix | Whether contains an appendix or not (*sp*) | boolean |
| | prop_of_freq_words | Proportion of frequent words (*sp*) | float |
| *Lexical* | BOW | Bag-of-words in abstract | integer |
| | BOW+TFIDF | TFIDF weighted BOW in abstract | float |
| | GloVe | Average of GloVe word embeddings in abstract | float |
| | GloVe+TFIDF | TFIDF weighted average of word embeddings in abstract | float |

**Table 7:** List of *coarse* and *lexical* features used for acceptance classification task. *sp* refers features extracted from science-parse.

**APPROPRIATENESS (1-5)**

Does the paper fit in ACL 2016? (Please answer this question in light of the desire to broaden the scope of the research areas represented at ACL.)

5: Certainly.
4: Probably. W
3: Unsure.
2: Probably not.
1: Certainly not.

**CLARITY (1-5)**
For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

5 = Very clear.
4 = Understandable by most readers.
3 = Mostly understandable to me with some effort.
2 = Important questions were hard to resolve even with effort.
1 = Much of the paper is confusing.

**ORIGINALITY (1-5)**
How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.

5 = Surprising: Significant new problem, technique, methodology, or insight -- no prior research has attempted something similar.
4 = Creative: An intriguing problem, technique, or approach that is substantially different from previous research.
3 = Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.
2 = Pedestrian: Obvious, or a minor improvement on familiar techniques.
1 = Significant portions have actually been done before or done better.

**EMPIRICAL SOUNDNESS / CORRECTNESS (1-5)**
First, is the technical approach sound and well-chosen? Second, can one trust the empirical claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?

5 = The approach is very apt, and the claims are convincingly supported.

4 = Generally solid work, although there are some aspects of the approach or evaluation I am not sure about.
3 = Fairly reasonable work. The approach is not bad, and at least the main claims are probably correct, but I am not entirely ready to accept them (based on the material in the paper).
2 = Troublesome. There are some ideas worth salvaging here, but the work should really have been done or evaluated differently.
1 = Fatally flawed.

**THEORETICAL SOUNDNESS / CORRECTNESS (1-5)**
First, is the mathematical approach sound and well-chosen? Second, are the arguments in the paper cogent and well-supported?

5 = The mathematical approach is very apt, and the claims are convincingly supported.
4 = Generally solid work, although there are some aspects of the approach I am not sure about or the argument could be stronger.
3 = Fairly reasonable work. The approach is not bad, and at least the main claims are probably correct, but I am not entirely ready to accept them (based on the material in the paper).
2 = Troublesome. There are some ideas worth salvaging here, but the work should really have been done or argued differently.
1 = Fatally flawed.

**MEANINGFUL COMPARISON (1-5)**
Do the authors make clear where the problems and methods sit with respect to existing literature? Are the references adequate? For empirical papers, are the experimental results meaningfully compared with the best prior approaches?

5 = Precise and complete comparison with related work. Good job given the space constraints.
4 = Mostly solid bibliography and comparison, but there are some references missing.
3 = Bibliography and comparison are somewhat helpful, but it could be hard for a reader to determine exactly how this work relates to previous work.
2 = Only partial awareness and understanding of related work, or a flawed empirical comparison.
1 = Little awareness of related work, or lacks necessary empirical comparison.

**SUBSTANCE (1-5)**
Does this paper have enough substance, or would it benefit from more ideas or results? Note that this question mainly concerns the amount of work; its quality is evaluated in other categories.

5 = Contains more ideas or results than most publications in this conference; goes the extra mile.
4 = Represents an appropriate amount of work for a publication in this conference. (most submissions)

3 = Leaves open one or two natural questions that should have been pursued within the paper.
2 = Work in progress. There are enough good ideas, but perhaps not enough in terms of outcome.
1 = Seems thin. Not enough ideas here for a full-length paper.

**IMPACT OF IDEAS OR RESULTS (1-5)**
How significant is the work described? If the ideas are novel, will they also be useful or inspirational? Does the paper bring any new insights into the nature of the problem?

5 = Will affect the field by altering other people's choice of research topics or basic approach.
4 = Some of the ideas or results will substantially help other people's ongoing research.
3 = Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.
2 = Marginally interesting. May or may not be cited.
1 = Will have no impact on the field.

**IMPACT OF ACCOMPANYING SOFTWARE (1-5)**
If software was submitted or released along with the paper, what is the expected impact of the software package? Will this software be valuable to others? Does it fill an unmet need? Is it at least sufficient to replicate or better understand the research in the paper?

5 = Enabling: The newly released software should affect other people's choice of research or development projects to undertake.
4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
3 = Potentially useful: Someone might find the new software useful for their work.
2 = Documentary: The new software useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
1 = No usable software released.

**IMPACT OF ACCOMPANYING DATASET (1-5)**
If a dataset was submitted or released along with the paper, what is the expected impact of the dataset? Will this dataset be valuable to others in the form in which it is released? Does it fill an unmet need?

5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.
4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
3 = Potentially useful: Someone might find the new datasets useful for their work.

2 = Documentary: The new datasets are useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
1 = No usable datasets submitted.

**RECOMMENDATION (1-5)**
There are many good submissions competing for slots at ACL 2016; how important is it to feature this one? Will people learn a lot by reading this paper or seeing it presented?

In deciding on your ultimate recommendation, please think over all your scores above. But remember that no paper is perfect, and remember that we want a conference full of interesting, diverse, and timely work. If a paper has some weaknesses, but you really got a lot out of it, feel free to fight for it. If a paper is solid but you could live without it, let us know that you're ambivalent. Remember also that the authors have a few weeks to address reviewer comments before the camera-ready deadline.

Should the paper be accepted or rejected?

5 = This paper changed my thinking on this topic and I'd fight to get it accepted;
4 = I learned a lot from this paper and would like to see it accepted.
3 = Borderline: I'm ambivalent about this one.
2 = Leaning against: I'd rather not see it in the conference.
1 = Poor: I'd fight to have it rejected.

REVIEWER CONFIDENCE (1-5)
5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.
4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.
2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
1 = Not my area, or paper is very hard to understand. My evaluation is just an educated guess.