36468 Final Project: A Linguistic Analysis of Conference Review

Siqi Zeng
Carnegie Mellon University
Pittsburgh, PA 15213
siqiz@cs.cmu.edu

1 Introduction

Every year, an increasing number of scholars find opportunities to publish their work in journals and conferences, where peer-review is a crucial part of the selection process. In top conferences, reviewers are often senior researchers with sufficient publication records, but there are still many complaints from paper authors that some reviewers do not understand the paper's main topic and give very low ratings. Theoretically, conference program chairs will decide to accept or reject the paper based on reviews and ratings from different reviewers, so it is meaningful and essential to analyze these texts to ensure reviewers do their best to give professional feedback. Although there is a reviewing guideline provided, reviewers' focus might vary even in the same year. Are they analyzing results in papers, expressing their emotions, or just giving random comments? Using NeurIPS, one of the most influential machine learning conferences, as my research object, in this report, I want to resolve the following research questions from a linguistic perspective:

- What linguistic features in reviews are important for ratings, and how make these features impact ratings and decisions?
- Are important linguistic features of reviews the same as the ones affecting the committee's decision?
- Are important features in reviews consistent in a different version of the same academic conference?

Assuming that reviewers are professional enough to make comments, I expect to see words with obvious negative or positive tendencies related to ratings. The same features should contribute to ratings and decisions because the committee is supposed to hear the suggestions from reviewers before deciding to reject or accept the paper. The answer to the third question is unknown because the organizer may adjust reviewing criteria annually, probably because they notice any weakness in the previous reviewing process or a change of conference theme. It is hard to judge the positive or negative effect of the change, but the existence of change may have some implications.

2 Data

NeurIPS 2019 Reproducibility Challenge (**NIPS19**) calls for researchers to verify and reproduce the empirical results in NeurIPS 2019 accepted papers. NeurIPS 2021 Datasets and Benchmark Track (**NIPS21**) is a part of NeurIPS 2021 main conference. Researchers create novel datasets which may drive the progress of a certain field in machine learning. Only reviews of these two tasks in NIPS are currently available to the public on OpenReview[1][2].

[1] https://openreview.net/group?id=NeurIPS.cc/2019/Reproducibility_Challenge

[2] https://openreview.net/group?id=NeurIPS.cc/2021/Track/Datasets_and_Benchmarks/Round2

corpus	number of reviews	number of papers	word count
NIPS19	286	83	99837
NIPS21	269	108	26535

Table 1: Corpus Statistics

Raw review texts are downloaded through web scraping. All texts inside the discussion panel are stored so that contents may be reviewers' comments, authors' feedback, and feedback from others. The last type, feedback from others, is often seen in **NIPS19**. Since **NIPS19** comes from a reproducibility challenge, authors of reproduced papers often provide thoughts, appreciation, or clarification if their papers are selected. Although OpenReview includes authors' comments for rebuttal purposes, these texts are not my research target, so they are manually excluded from the raw dataset. One main difference is **NIPS19** does not allow rebuttal because it is a challenge, while **NIPS21** does. Therefore, the average length of the text in **NIPS21** is greater.

In **NIPS19**, OpenReview gathers a single person's review in a large chunk and occasionally with some additional comments. While in **NIPS21**, OpenReview splits one person's review into many sub-parts: *Summary and Contributions*, *Rating*, *Strengths*, *Confidence*, *Weaknesses*, *Correctness*, *Clarity*, *Relation To Prior Work*, *Documentation*, *Ethics*, *Additional Feedback*, *and Comments*. Except for *Rating* and *Confidence*, everything is concatenated into a whole text file, and only sub-parts with more than 20 words are kept. The 20 word-length filter is set because shorter text might be some evaluative template such as "*Overall reviews are not satisfactory enough for the AC to consider for the journal.*", which is not very helpful to understand reviewers' language in this study. Reviewers' *Ratings* of the paper, ranging from 1 (*strong reject*)-10 (*strong acceptance*), and their *Confidence*, ranging from 1 (*educated guess*)-5 (*absolutely confident*), about their comments are directly extracted from the number in the text.

The statistics of **NIPS19** and **NIPS21** corpora after data cleaning are showed in Table 1. LIWC, a commercial linguistic tool, generates the word count. Word count increases by 1 if the tool detects one space interval (1). In this report, each review is considered as a data point. Since authors of the rejected paper in 2021 NIPS can opt-out presenting their paper on OpenReview, most rejected papers are unavailable. Thus, **NIPS21** is a very unbalanced dataset, only containing four reviews from one rejected paper. From the text content of the review, the committee rejected this borderline paper, although all reviewers agreed on acceptance. Nevertheless, in short, these corpora are a great fit for this study because ratings and decisions are from raw data, and the corpora in different years allow me to compare features in a different version of the same conference.

3 Methods

Workflow Summary The workflow can be concluded into several steps: [1] Extract LIWC features. [2] Rank LIWC features based on permutation importance and select the optimal number of features through cross-validation. [3] Use the same features to build a regression model for ratings and a classification model for decisions. [4] Analyze results in step $2, 3$ for two years' data.

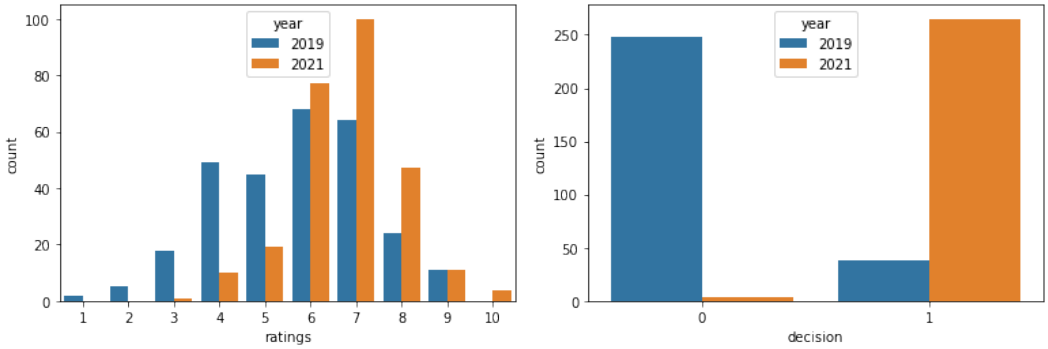
Related Works The method applied in this experiment mainly follows the analytical part by Besselaar et al (2). They used LIWC (`http://liwc.wpengine.com/`), a commercial linguistic analysis engine, to extract the percentage of each word belonging to each predefined linguistic category (1), and then applied linear regression to predict grant application panel scores from the linguistic dimension values.

Linear models is highly interpretable compared with other modern alternatives. Due to the difference in research problems, several adjustments are made:

[1] In Besselaar et al.'s paper, they customized their dictionaries to extract grant-specific features, while details of their dictionary are not publicly available. Therefore, I use the default English LIWC

	ratings19	ratings21	decision19	decision21
min	1.00	3.00	0	0
max	9.00	10.00	1	1
mean	5.71	6.76	0.13	0.98
sd	1.64	1.16	0.34	0.12

Table 2: Statistics of Dependent Variables



(a) Distribution of ratings

(b) Distribution of decision

Figure 1: Distribution of dependent variables in two years

2015 dictionary for feature extraction, and definitions of most features are available in LIWC's manual.

[2] Not all of my dependent variables are continuous variables, so I build linear regression models to predict continuous variables and logistic regression for classification.

[3] Instead of using a naive linear model as in Besselaar's work, I chose the Lasso linear model. Lasso, or L1 regularizer, penalizes unimportant features. Without Lasso, some features such as pronouns will significantly impact the response since its value is high for all samples, but these features cannot help explain the difference between different ratings too much. Besides, we only have around 250 samples, so that the naive linear model will suffer from collinearity and high feature dimension. Lasso can set unimportant features' coefficients as zero, and if two features are collinear, the beta coefficients are distributed between correlated features.

Variables and Feature Selection There are two dependent variables in this report, ratings, and decision (accept/reject). The distribution of decision, helpfulness, and ratings are presented in Table 2. Interestingly, the difference of mean ratings is only 1, not a considerable amount, although most **NIPS19** papers receive a rejection while most **NIPS21** papers are accepted. This may indicate a change in the rating procedure. Figure 1 contains histograms. When building models, rating is treated as a continuous variable, and decision is treated as a discrete binary variable.

Independent variables come from 94 features generated by LIWC. Except for total word count and summary variables (*Analytic, Clout, Authentic, Tone*), raw LIWC feature values are defined by the number of words labeled in an LIWC category over total word count in the file multiplies by 100. All of these features are normalized by dividing the range from their minimum to maximum. 94 is a big number compared to the total number of dependent variables, so feature selection is a necessary step. I calculated permutation feature importance for a Lasso linear regression model fitted on a train set with all 94 features. In a regression model, permutation importance can be defined by decreasing r^2 if I randomly shuffle values for one column feature (3). Then, I use a $10-$fold cross-validation to determine the optimal number of features included in the final model. Meanwhile, α in Lasso is tuned to achieve the highest r^2 on the training dataset.

Coefficients:		
	Estimate	**p-value**
negate	1.9839	0.0002
Clout	7.6270	0.0000
health	4.4412	0.0000
adverb	7.5497	0.0000
leisure	7.5841	0.0000
money	4.8400	0.0000
motion	4.9777	0.0000
reward	8.2725	0.0000
relig	7.0703	0.0000
prep	5.0823	0.0000
power	5.3418	0.0000
sexual	5.1580	0.0000
assent	5.1155	0.0000
ingest	5.6088	0.0000

Coefficients:		
	Estimate	**p-value**
function	8.3858	0.0000
auxverb	7.9169	0.0000
prep	7.5879	0.0000
i	8.3187	0.0000
Clout	7.8424	0.0000
relativ	5.8959	0.0000
article	6.9121	0.0000
space	7.9963	0.0000
differ	7.2404	0.0000
time	10.3937	0.0000
Tone	6.9877	0.0000
sexual	3.3061	0.0000
health	9.7842	0.0000
compare	7.1812	0.0000
Dic	5.4586	0.0000
posemo	7.5914	0.0000

Table 3: Left: Model Summary of **NIPS19-ratings** Right: Model Summary of **NIPS21-ratings**

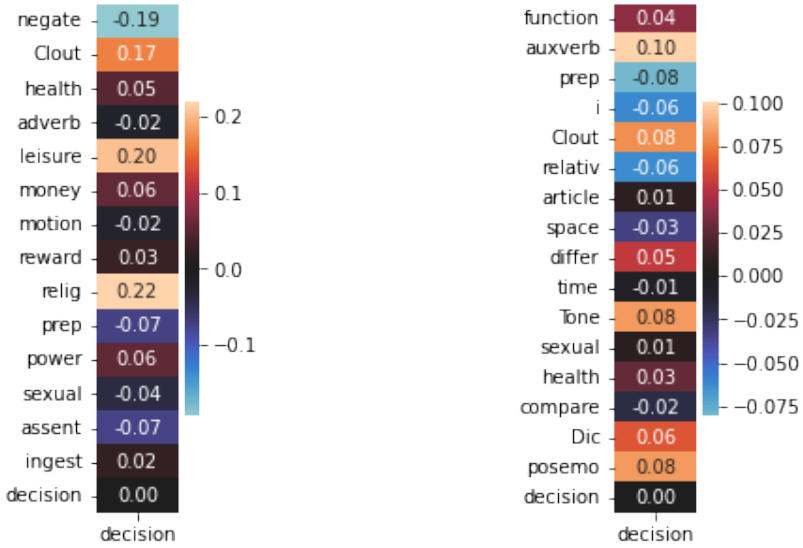
Model Building A Lasso regression model is built on the whole dataset using the selected top k important features after verifying the linearity assumption of the regression model. The coefficients of the model are recorded. I also conducted a significance test for all selected coefficients, where the null hypothesis is that a coefficient is greater than zero. Instead of the classic two-tailed significance test, I choose to use one-tailed because all coefficients are positive. Then, to test how these features perform on decision prediction, an L1 logistic regression model is built. The L1 penalty is used to guarantee consistency with the regression model and prevent the case if some of the selected features are strongly correlated. Accuracy of logistic regression model on train and test set is recorded. Besides, the correlation matrix for all features and decision is made to show the relationship between them. Rather than Pearson's correlation coefficient, each entry is a point biserial correlation coefficient, a special case for Pearson's correlation coefficient, since decision is a categorical variable and other features are continuous variables (4). Point biserial correlation coefficient ranges from -1 to 1, -1 means a strong negative correlation, 0 means no correlation, and 1 means a strong positive correlation.

Relation to Research Questions Feature selection resolves the third research question. If there is a huge difference between years among selected features, then the answer to the third question is very likely to be a no. After feature selection, I only keep features beneficial to prediction, reflecting the significance of these features. Significance can be validated again through hypothesis tests. Since I am building linear models, coefficients can be easily interpreted to show the relationship between committee ratings and linguistic features. A correlation heatmap can show the relationship between decision and all other features. Thus, the first question is also answered. The accuracy of classification answers the second question. If accuracy is high, then it implies a good transferability between two problems.

4 Results

4.1 Selected Features

Selected features for **NIPS19** and **NIPS21** are listed in Table 3, and their distributions is showed in 3. All features are greater than zero with strong statistical significance. All selected independent variables' definitions can be found in Table 5. Recall the range of ratings is from 1 to 9 in **NIPS19** and 3 to 10 in **NIPS21**. Although I use Lasso linear regression, coefficients can be interpreted similarly to the approach in naive linear regression. For example, 1.9839 means the expected difference of ratings if we happen to increase $negate$ by one.



(a) **NIPS19-decision** (b) **NIPS21-decision**

Figure 2: Transposed last row of correlation matrix in two years

Task	ACC%
NIPS19-train	87.94
NIPS19-test	75.86
NIPS21-train	98.76
NIPS21-test	96.30

Table 4: Decision Classification Accuracy

4.2 Model Evaluation

Train r^2 and test r^2 are 0.3788 and 0.1258 in **NIPS19**, 0.1225 and 0.0303 in **NIPS21**. Models are tuned to achieve the best training accuracy, and overfitting happens in the test set. I focus more on inference between features and response, so the models in Table 3 are built on train plus test set. r^2 for models in Table 3 are 0.3603, 0.1183 for **NIPS19** and **NIPS21**. r^2 means how much variance of ratings can be explained by features. Low r^2 and overfitting imply that this is not a good prediction model but does not affect the inferential relations, supported by small p-values. Limitations of a low r^2 model will be discussed in Section 5.3.

To see how selected features work in a decision classification model, I made two correlation matrices. The diagonals are set to zero so that the color of the heatmap can be distributed evenly. Some highly correlated features are discovered in this matrix. As an extreme example, correlation between *posemo* and *Tone* is as high as 0.77. Nevertheless, I use L1 penalties in all models to alleviate this problem. The transposed last row of the heatmap shows how each feature contributes to decision, included in Figure 2.

The accuracy of the logistic regression model for decision classification is shown in Table 4. Models have a good performance using the selected data.

5 Discussion

5.1 NIPS19 vs. NIPS21

5.1.1 Ratings

Selected features of **NIPS19** contain 3 functional (in a grammatical sense) features: *negate*, *adverb*, *prep*, 1 LIWC summary feature: *Clout*, and 10 descriptive features: *health*, *leisure*, *money*, *motion*, *reward*, *relig*, *power*, *sexual*, *assent*, *ingest*. In **NIPS21**, there are 5 functional features: *function*, *auxverb*, *prep*, *i*, *article*, 3 LIWC summary features: *Clout*, *Tone*, *Dic*, and 8 descriptive features: *relativ*, *article*, *space*, *differ*, *time*, *sexual*, *health*, *compare*, *posemo*. The distribution between categories is close, but the specific LIWC dimension varies within the same category.

From observation, increasing any feature in Table 3 by one will directly give the value of ratings. Thus, I split the features by corpus mean. If the coefficient is lower than, for example, 5.71 in **NIPS19**, then this feature is considered as a low ratings indicator. 5 high ratings indicators are *Clout*, *adverb*, *leisure*, *reward*, *relig* in **NIPS19**. 13 high ratings indicators are 13 *function*, *auxverb*, *prep*, *i*, *Clout*, *article*, *space*, *differ*, *time*, *Tone*, *health*, *compare*, *posemo*. More high ratings indicators may be caused by more accepted papers in **NIPS21**.

Functional features In **NIPS19**, *negate* indicates the reviewer does not agree with the method or conclusion in a paper and thus is naturally related to low ratings. As I will claim in Descriptive and Summary features sections, if a reviewer strongly recommends a paper, he or she will praise it at one's best. This is how *adverb* like *very*, *really* are used in the context. For feature *prep*, it has the strongest positive relationship with *Analytic*, a summary LIWC variable for a text file. In **NIPS19**, more *prep* often relates to more detailed analysis to the paper, which leads to a low score in the evaluation process in 2019.

In **NIPS21**, *function* includes all functional features, and all of them are high ratings indicators, which is different from 2019 corpus. This is caused by a longer average length of text in **NIPS21**, and functional words emerge very commonly in any paragraphs. Also, most papers in **NIPS21** are accepted, and the average rating is higher. Given the low r^2 of the regression model and high frequency of functional words, the explanatory power of any single functional feature in the 2021 corpus is limited. However, a longer average length means a longer discussion thread in a rebuttal process, meaning that authors are actively tweaking their papers to satisfy reviewers. Function words are a sign of high ratings for **NIPS21**, as OpenReview only records the final score from reviewers. This process can be seen in one comment that

> *Despite some initial disagreements around terminology and claimed contributions, the reviewers and the authors engaged in a very constructive review process, resulting in a significant improvement of the submission.*

The difference between behavior of *prep* and *function*, superset of *prep*, is very likely to be caused by a different review process. In **NIPS19**, authors cannot add an explanation on reviewers' comments, so a reviewer only has one chance for explanation. He or she will analyze the paper in detail and tell every reason why this paper is not the best if a low rating is given. However, in **NIPS21**, everyone has more freedom in communication, so reviewers do not feel obligated to put their analysis to the paper in their comments.

Descriptive features Examples can help explain how descriptive features affect ratings in the context. For instance,

> *In the NeurIPS paper, the authors perform experiments on the Wine, Cancer, Face, and MNIST datasets. In the reproducibility report the authors state that they perform experiments on Abalone, Iris, Reddit, Wine, and Cancer.*

Cancer is the only word in *health* category and all *health* labels in **NIPS19** are from *Cancer* because authors implemented the same paper in that years' main conference. Another example in **NIPS21** is

> *Overall, this is a very good paper that will significantly contribute to promote researches toward high-quality medical image diagnosis in an explainable manner.*

The *health* labels come from *medical* and *diagnosis*. Other *health* labels are from new medical image benchmarks. In general, for features of this type, the feature mean is usually low compared to summary variables because these types of words are unique to the paper's topic. Other features of the same type are *leisure, money, motion, relig, power, sexual, ingest, relativ, space, time*. Their influence on ratings varies in the same corpus, so that these coefficients estimates may not be well generalized in the new dataset. This can be observed from lower r^2 in the test dataset in comparison to the training dataset. But the amount of estimates can reflect trending topics favored by the reviewer group in that year. In the context of a machine learning conference, these topics can be translated as application fields. For instance, since the conference paper *Abalone, Iris, Reddit, Wine, and Cancer* is difficult to reproduce, *health* becomes a low ratings indicator. However, medical imaging is a relatively unexplored field where scholars can design innovative benchmarks. These features may guide future scholars who want to submit papers. Authors may want to keep working on last year's hot topics and avoid those that reviewers dislike. Alternatively, they can work on a tough topic to make a great contribution to the field, depending on the conference subject. Although LIWC can not discover some theoretical and abstract topics, I can see a general pattern that reviewers prefer to add a paper summary and restate what they read. This is often a good start, showing reviewers' understanding of the paper concisely. If reviewers misunderstand, authors can quickly react and explain to them in the rebuttal without reading several long paragraphs.

Furthermore, there are several features without concrete meaning. Compared to previous groups of features more like a noun, these features usually represent some action. These LIWC features are *compare, reward, posemo, assent, differ*. *compare* is a high ranking indicator because in LIWC dictionary, provided examples are *greater, best*. Reviewers use *compare*-labeled words in that this paper is greater than some previous work, which explains that it is a high-ratings indicator. *reward* and *posemo* are positively related to ratings because these are groups of complimentary words, often showing that reviewers agree with authors' main idea, therefore, giving a high rating. Influence of *assent* and *differ* are counter-intuitive at a first glance, but these can be resolved by inspection into sentences. While *reward* and *posemo* are a sign of high ratings, direct agreement, or *assent* is not. *assent* are often seen in a concession context, such as

> *However due to space constraints and relative impact, AC is unable to recommend this paper for the journal. Although, the AC acknowledge and agree with the reviewers on the contribution of the reproducibility effort and hope the readers and original authors gain useful insights from the findings.*

It is reasonable because agreement is not enough for a paper to stand out as the best candidate. In a top conference, the competition is intense. New and efficient methods, not widely acceptable conclusions, will receive higher ratings. Raters make their best effort to choose the best fit for the conference. *differ* is classified as a cognitive process, differentiation, in the LIWC manual. It usually occurs in disagreement:

> *I know the authors mentioned that they downsampled the GB1 data(which may in itself be a problem), but is it possible to download all the raw data so that other groups may try different methods of accounting for the class imbalance? It wasn't entirely clear to me if this was possible from what I could currently see.*

Yet this is often a sign of high ratings. Unlike *assent* where reviewers are not interested in communicating with authors, *differ* feature words give authors enough space for making future improvements because they often occur in an interrogative sentence. Reviewers expect feedback from authors.

Summary features Three types of summary features are *Tone, Dic, Clout.* Their mathematical expression is hidden from the LIWC manual. *Tone* behave the same way as *posemo* during interpretation, although there are some built-in algorithms in LIWC to calculate *Tone.* So its relation with ratings matches with *posemo*'s relation with ratings. *Dic* is a low rating indicator because the more usage of common words in LIWC dictionary, the less the review is related to the topic of the paper. In a review of an academic paper, many technical words should not occur in the dictionary, such as *convolutional*, *overfitting* in machine learning. The deviation from academic discussion shows a wrong direction of the reviewer's understanding and thus associates with low ratings. *Clout* is a strongly positive factor for ratings in both corpora, which shows the confidence of reviewers. Several sources (5; 6; 7) claim that a higher *Clout* score reflects a sense of authority. From the model, higher *Clout* relates to higher ratings, so in general, reviewers avoid giving a confident low score although they are in a double-blind reviewing process. Reviewers' prior belief is that they may not understand authors' work thoroughly, while the paper is beneficial to academics. In contrast, if they discover any great fit for acceptance, they will not hesitate to recommend and utilize their authoritativeness of implicitly in their wordings. More recommendations may increase the committee's work, but most good papers can not be ignored as long as more than one person reviews them. When there are some controversies, committee members will not give an absolute rejection. They may consider more before the final decision if reviewers have low confidence in a low rating. To sum up, I can see a part of a reviewing process rigorous for reviewers and tolerant for paper authors.

5.1.2 Decision

Although logistic regression models show high classification accuracy using the same features as linear regression models, many relationships are inconsistent with Ratings.

Comparing Figure 2a with high and low ratings indicators using the definition in Ratings section, the most positive relationship matches its correspondence. *Clout*, a high ratings indicator, is positively related to the decision. However, some low rating indicators behave in the other direction, such as *power, ingest.* Although these correlations are not strong, the fact should draw attention. The correspondent between Figure 2b and 2021's rating coefficients is messier. In particular, although *function* includes all functional words, its subsets may behave in opposite directions. However, recall in **NIPS21-ratings**, all functional words are all high rating indicators. Combining observations of both decision models, I think the program chair or committee, who made the final decision, value the same dimension as reviewers, but they may weigh these features differently.

5.2 Conclusion

From this study, features critical for ratings and decision are detected through fitting a regression model. The same group of linguistic features can be transferred from ratings to explain decision, but how these features affect ratings is different from decisions. One highlighting consistent feature is *Clout*, which is always positively related to dependent variables. A certain extent of consistency is reflected within ratings and decision, and also within **NIPS19** and **NIPS21**. There are changes in important features between the two corpora. Reviewers adjust their wordings and contents in review under different settings, even in the same conference.

To sum up, since all similarities, differences, and outliers are interpretable, I claim that reviewers in NeurIPS follow a good standard. Their comments and ratings are well-established.

5.3 Limitation and Future Research

Although the number of words in corpora is large, the sample size is relatively small since it is defined by the number of reviews. The conclusion from this paper is hard to generalize, and an unbalanced

dataset worsens this problem. This indirectly leads to a model with a small r^2, reducing the model's explanatory power. The effect size of each LIWC feature requires more validation. In a previous study for the grant decision-making process (2), Besselaar et al. claim this is caused by considerable controversies from reviewers, and we can observe high variance within **NIPS19** and **NIPS21** dataset. A highly unbalanced dataset undermines the persuasiveness of the classification model because a majority vote can already present good performance values. All problems can be solved if there are some larger publicly available data related to this topic, and the conclusion can be improved by further causal analysis.

The definition of feature values causes another major weakness. LIWC only evaluates a single word, which may mask some hidden relationship under the context. Using n-grams or collocates analysis will help uncover some patterns, but LIWC dimensions are no longer a good fit. Systematic annotation of the psychological dimension of phrases is out of the scope of this study. Also, LIWC's labeling policy may have a wrong interpretation for certain words. For example, it will categorize *love* in *I love this idea* as a *sexual* word, or *sect* in *comparisons involving the study introduced here (sect. 2.2)* as a *relig* word. In this context, the label of words is questionable. Last, only considering linguistic features cannot perfectly reflect reviewers' understanding of the paper's topic. Hendrycks et al. (8) created MATH benchmark for NeurIPS 2021. From their paper, even the latest natural language processing models can only solve 5% of advanced math problems, indicating that there is still a long way to go before computers can analyze human thoughts. Currently available methods cannot fully resolve authors' complaints. When future work is done, similar methods can be applied with new tools.

References

[1] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," tech. rep., 2015.

[2] P. van den Besselaar, U. Sandström, and H. Schiffbaenker, "Studying grant decision-making: a linguistic analysis of review reports," *Scientometrics*, vol. 117, no. 1, pp. 313–329, 2018.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] D. Kornbrot, *Point Biserial Correlation*. John Wiley Sons, Ltd, 2014.

[5] R. L. Moore, C.-J. Yen, and F. E. Powers, "Exploring the relationship between clout and cognitive processing in mooc discussion forums," *British Journal of Educational Technology*, vol. 52, no. 1, pp. 482–497, 2021.

[6] E. Kacewicz, J. W. Pennebaker, M. Davis, M. Jeon, and A. C. Graesser, "Pronoun use reflects standings in social hierarchies," *Journal of Language and Social Psychology*, vol. 33, no. 2, pp. 125–143, 2014.

[7] W. W. Xu and C. Zhang, "Sentiment, richness, authority, and relevance model of information sharing during social crises—the case of mh370 tweets," *Computers in Human Behavior*, vol. 89, pp. 199–206, 2018.

[8] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the MATH dataset," *CoRR*, vol. abs/2103.03874, 2021.

[9] M. Partch and C. Dykeman, "Text messaging as a mental health treatment intervention: A corpus-based study," 02 2019.

Appendix

LIWC variables	**Description/Examples** (1)
negate	no, not, never
Clout	Nontransparent LIWC feature. Reflect the level of confidence of the speaker (5; 6). Higher Clout score is marked by we and social words, fewer I, negations, and swear words (7).
health	clinic, flu, pill
adverb	very, really
leisure	cook, chat, movie
money	audit, cash, owe
motion	arrive, car, go
reward	take, prize, benefit
relig	altar, church
prep	to, with, above
power	superior, bully
sexual	horny, love, incest
assent	agree, OK, yes
ingest	dish, eat, pizza
function	all function words, including total pronouns, articles, prepositions, auxiliary verbs, common adverbs, conjunctions, negations
auxverb	am, will, have
i	I, me, mine
relativ	including motion, space, time: area, bend, exit
article	a, an, the
space	down, in, thin
differ	hasn't, but, else
time	end, until, season
Tone	Nontransparent LIWC feature. Emotional tone. A higher score means a positive tone. Scores below 50 mean a negative tone (9).
compare	greater, best, after
Dic	Words in LIWC dictionary.
posemo	love, nice, sweet

Table 5: Description of LIWC features in the model

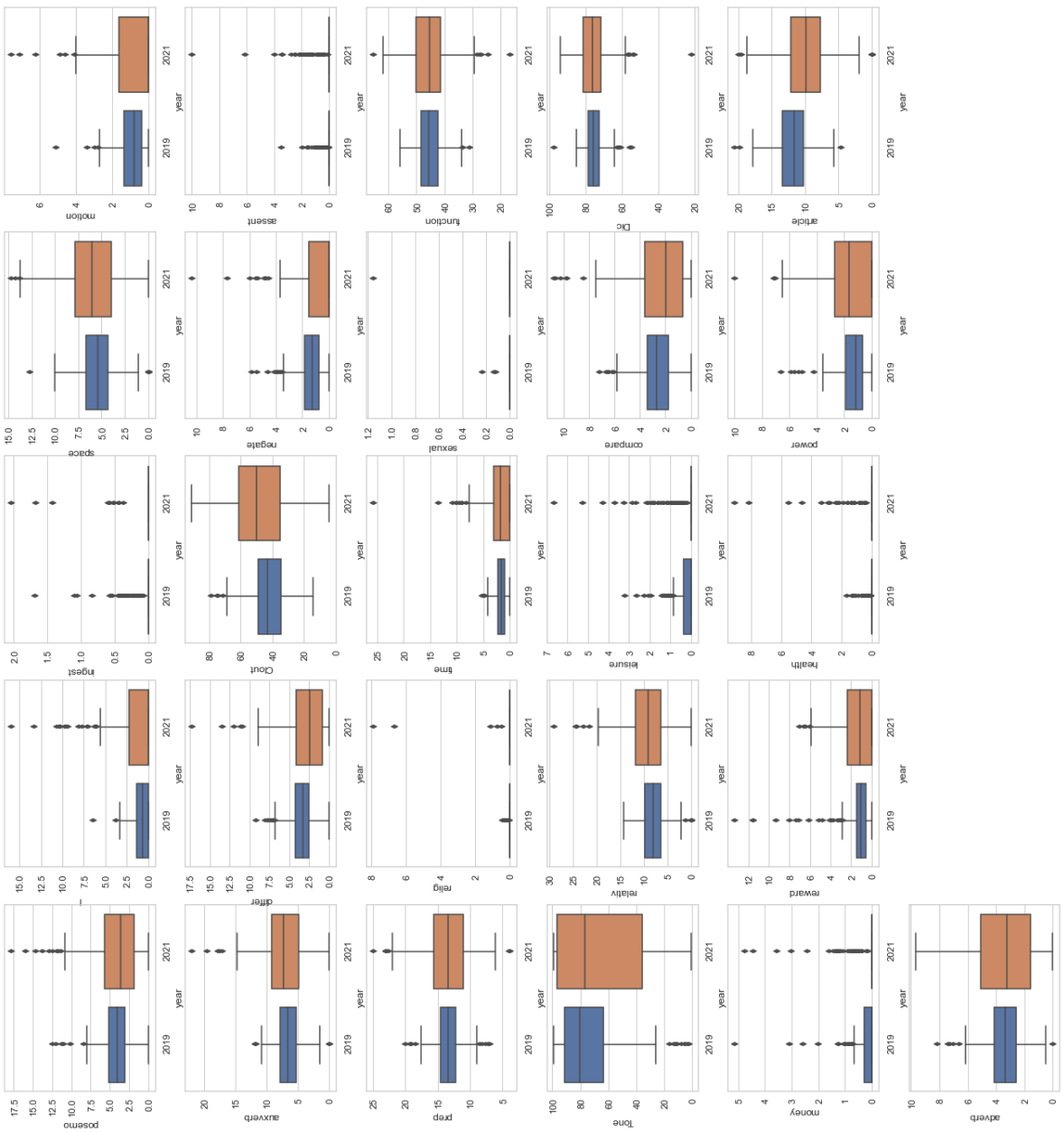


Figure 3: Distribution of selected features in two corpora