HearingQA: A Dataset of Question and Answers in U.S. Committee Hearings

Anonymous ACL submission

Abstract

Existing literature on the quality of questions and answers focus on factual question answering and natural language understanding. However, there are no large scale datasets enabling the study of question and answer quality in public hearings and interview like settings. One challenge for constructing such a dataset is that public hearings and interviews are typically only accessible as unstructured transcripts in plain text. We develop a pipeline to extract utterances and identify utterances as questions and answers, which can then be used for downstream tasks of studying question and answer quality along different dimensions. Using this pipeline, we build a novel dataset constructed from committee hearings from the U.S. House of Representatives and Senate consisting of all questions and answers in transcripts from the 108th to the 117th congressional sessions. We find that it is possible to accurately distinguish the party affiliations of the questioners based on the question utterances alone indicating that committee members with different party affiliations use language differently when asking questions in committee hearings.

1 Introduction

Congressional *hearings* are a cornerstone of the legislative process in the United States. Elected representatives in the two chambers of the legislative chambers, the house of representatives and the senate have the authority to use subpoenas to compel evidence and summon individuals, called *witnesses*, to testify at these hearings. Congressional hearings help shape public policy (Oleszek et al., 2015), conduct investigations (Hamilton et al., 2007), oversee government actions (Dolan et al., 2014; Levin and Bean, 2018), ensure accountability of government agencies under the executive branch (Bussing and Pomirchy, 2022), and review the appointment of presidential nominees for key positions including judges (Collins Jr and Ringhand, 2016) and cabinet members (Ross, 1998). The public nature of these hearings provide members of the public a window into the complex, often contested process of lawmaking and oversight at the highest levels of government. This process is facilitated by the division of the elected representatives in each chamber into distinct *committees* with responsibilities and jurisdiction over legislation, activities and oversight on specific issues. Each party is typically represented in each committee proportionally to their strength in each chamber.

A key component of these hearings involves members of the committee at the hearing interviewing the witnesses. Each member is assigned a pre-determined time slots within which they may question the witnesses (Sachs, 2003). The effectiveness of questions asked in these hearings can have far-reaching consequences. For example, the failure of senate nomination hearing for supreme court justice Brett Kavanaugh in eliciting accurate information pertinent to an investigation into accusations of sexual assault and information about his stance on consequential constitutional issues and legal doctrine, its bearing on the committee's recommendation and ultimately the senate's vote on whether or not to accept the nomination have been extensively studied in the political science (Hemingway and Severino, 2019; Fredrickson, 2019) and discource analysis (Kaur, 2022) literature.

Despite the important role congressional hearings play in the democratic process, they are often perceived as dysfunctional (Lewallen et al., 2016). While hearings ought to perform the functions of fact finding, gathering expert opinions and conducting investigations (Huitt, 1954), they are instead sometimes a means to spread propaganda (Truman, 1951) and for elected representatives to seek attention and credit (Davidson et al., 2023) or grandstand (Lewallen et al., 2024) to interest groups and constituents. Indeed, committee members may enter a hearing with prepared questions (Oleszek et al., 2015), while witnesses may be invited strategically by committee members to support a certain view (Talbert et al., 1995), and may have prepared responses (Oleszek et al., 2015) which has been exacerbated by increasing partisanship in politics.

Party affiliation of a legislator can reflect different political agendas and strategic goals – ranging from seeking transparency and accountability to promoting or defending policy positions. We conjecture that these political agendas and strategic goals are reflected on the linguistic structures, tone, and style of the questions asked in the congress. In light of the dysfunctions in congressional hearings and the potential effect on questions and answers in congressional hearings, we seek to address the following research questions: (1) Does party affiliation impact the questions asked in hearings, and can we identify the party affiliation and determine whether their party is in the majority or minority? (2) Given an *answer utterance*, can we predict the party affiliation and standing of the *questioner* associated with the answer utterance.

1.1 Our Contributions

Our main conceptual contribution is the initiation of a new research direction on the large-scale and automatic analysis of the quality and use of language in questions and answers in interview like settings.

To initiate this line of research, we provide a novel dataset constructed from transcripts of committee hearings in the United States House of Representatives and Senate which identifies question and answer utterances. To facilitate future research, we also include human annotations of question and answer utterances, and a comprehensive set of linguistic features for each utterance.

We find that given a question utterance by a member of a committee, it is indeed possible to predict the questioner's party affiliation from the text of the question alone with an accuracy of 92%, and the standing of their party, i.e. whether the party is in the majority or minority in the respective chamber, with an accuracy of 75%. However, it turns out that identifying the affiliation and standing of the questioner from the answer obtained in response to a question appears to be a significantly harder problem.

We use a comprehensive collection of linguistic features (Horne et al., 2019) to find statistically significant differences in the use of language in questions from committee members with different party affiliations.

2 Related Work

We are not aware of datasets enabling the study of question and answer quality in interviews in general, or on political hearings or proceedings specifically, while the large-scale and automatic assessment of question and answer quality has been studied extensively in the question-answering literature in the context of answering factual questions (Zhang et al., 2023; Biancofiore et al., 2024; Yu et al., 2024), conversational question answering (Zaib et al., 2022) and natural language understanding (Kočiskỳ et al., 2018).

Hearing, Interview and Dialogue Datasets. The Congressional Committees Hearing dataset (CoCoHD) is perhaps the most similar dataset to ours and consists of transcripts from the congressional hearings from the 105th to the 118th congressional sessions (Hiray et al., 2024), with a focus on determining the impact of hearings in different committees on financial instruments related to issues under their jurisdiction. Our dataset differs from CoCoHD by separating individual utterances, identifying question and answer utterances within hearings from downstream question and answer quality assessment tasks, and providing utterance level annotations and linguistic features to facilitate further studies into the linguistic characteristics of different political actors.

The MediaSum (Zhu et al., 2021) and Interview (Majumder et al., 2020) datasets are perhaps the most prominent datasets consisting of real world dialogues and comprise of interviews conducted in mainstream media for dialogue summarization tasks. Xu et al. (2025) and Hoang (2025) provide a datasets of public remarks at local constituency and council meetings scraped from videos of proceedings and Fiva et al. (2025) provide transcripts of proceedings in the Norwegian parliament. Delano et al. archive congressional data including committee hearings and reports. The American Presidency Project (Woolley and Peters) includes transcripts of dialogues from campaign debates. Maher et al. (2020) analyze the testimonies of social scientist witnesses from 1946 to 2016 using a manual search process to identify relevant witnesses, and Fisher et al. (2013) analyze statements by committee members from the 109th and 110th congressional sessions from hearing that were related to climate change using discourse network analysis. However, these datasets do not identify question and answer utterances.

Arregui and Perarnaud (2022); Thomson et al. (2012); Thomson and Stokman (2006) analyze parliamentary and policy-making procedures at the European parliament through interviews of policy experts but not through the direct analysis of transcripts of the proceedings. Saliently, committee hearings are a common feature of many democratic institutions, and our framework for processing transcripts of proceedings to identify questions and answers withing hearings may be used to expand our dataset to include transcripts from those proceedings. We hope that this will enable comparative works on the effectiveness of the procedures and conduct in different democracies.

Our framework may also be extended to identify questions and answers in interview like settings such as interviews with politicians or political actors on news media. While datasets consisting of transcripts of such media exist, including those extracted from videos (Birkenmaier et al., 2025), we are not aware of datasets that identify questions and answers within such transcripts.

Function and Dysfunction in Congressional Hearings. Lewallen et al. (2024) study congressional hearings to analyze political messages sent by the policymakers. Ray (2018) studies legislator behaviors in the Congressional hearings and whether they become lobbyist after they retire from representing. Wolfe (2012) study how media coverage affects the time it takes an introduced legislation to become law in the 109th congressional session.

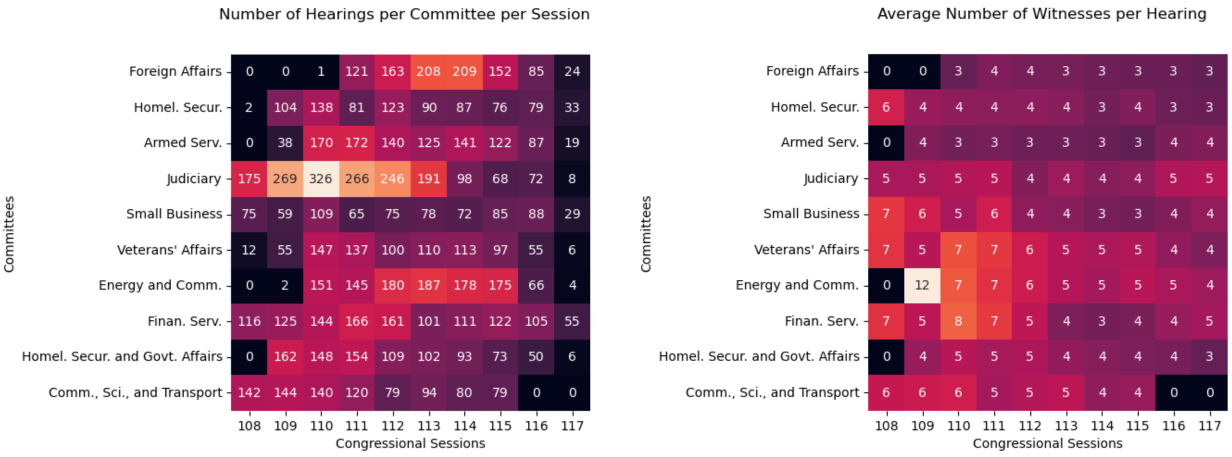
Esterling (2007) find through a study of major congressional hearings that contributions from interest groups incentivize committee members to ask analytical questions which induce a falsifiable response and engage in analytical discourse in certain major congressional hearings. Coil et al. (2024) study gender representation of witness in the congressional hearings, raising concerns of women being underrepresented. We hope that our dataset will enable fruther, large scale studies into similar behaviors in congressional hearings.

3 The HearingQA Dataset

Congressional hearings open to the public are made available in digital at `https://www.govinfo.gov/app/collection/chrg` between 2 months to 2 years after the hearing is held, and as Figure 2 illustrates, are published as a dialogue style transcript in plain text. Our dataset is constructed from all available transcripts as of April 2022, covering a span of 20 years spanning the 108th to the 117th congressional sessions with transcripts from 16,130 hearings by 76 committees involving 1,774 committee members and 69,149 witnesses in total.

Figure 1 (a) illustrates the distribution of hearings across congressional sessions and the top ten most active committees. We notice some clear trends of congressional committees responding to domestic and global events. The House Committee on Financial Services which oversees the financial services industry in the United States shows a clear increase in the number of hearings in response to the 2007–2008 global financial crisis between the 110th and 112th congressional sessions which span the years 2007 to 2013. We also see a clear increase in the average number of witnesses called to hearing within the same period in Figure 1 (b). A spike in the number of hearings by the Judicial committee in the 110th session coincides with growing concerns over domestic warrantless surveillance and wiretapping by law enforcement and national security agencies. Notice that some committees have been renamed over time although their roles and responsibilities typically have not changed over time. For example, the House Committee on Foreign Affairs operated as the Committee on International Relations from 1975 to 1978 and from 1995 to 2007. To avoid confusion, we maintain the name the committees operated under which they operated and only consider activity across committees when appropriate for further analysis.

Transcripts. Each transcript reflects the structure of a congressional hearing. A hearing typically starts with a roll call of the members present, followed by an address by the committee chair, followed by statements from the committee members and the witnesses. This is then followed by segments of back and forth exchanges, where the committee members ask questions which are to be answered by the witnesses. Some hearings have more structure with fixed time slots allocated to each member. Each transcript is associated with metadata identifying the hearing and information about the members in attendance, the committee(s) hosting the hearing, and witnesses present at the hearing along with their affiliations. There is also other data such as the date, time and address of the hearing and other administrative details of the hear-



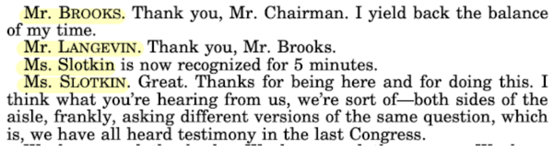
Number of Hearings per Committee per Session	108	109	110	111	112	113	114	115	116	117
Foreign Affairs	0	0	1	121	163	208	209	152	85	24
Homel. Secur.	2	104	138	81	123	90	87	76	79	33
Armed Serv.	0	38	170	172	140	125	141	122	87	19
Judiciary	175	269	326	266	246	191	98	68	72	8
Small Business	75	59	109	65	75	78	72	85	88	29
Veterans' Affairs	12	55	147	137	100	110	113	97	55	6
Energy and Comm.	0	2	151	145	180	187	178	175	66	4
Finan. Serv.	116	125	144	166	161	101	111	122	105	55
Homel. Secur. and Govt. Affairs	0	162	148	154	109	102	93	73	50	6
Comm., Sci., and Transport	142	144	140	120	79	94	80	79	0	0

(a)

Average Number of Witnesses per Hearing	108	109	110	111	112	113	114	115	116	117
Foreign Affairs	0	0	3	4	4	3	3	3	3	3
Homel. Secur.	6	4	4	4	4	4	3	4	3	3
Armed Serv.	0	4	3	3	3	3	3	3	4	4
Judiciary	5	5	5	5	4	4	4	4	5	5
Small Business	7	6	5	6	4	4	3	3	4	4
Veterans' Affairs	7	5	7	7	6	5	5	5	4	4
Energy and Comm.	0	12	7	7	6	5	5	5	5	4
Finan. Serv.	7	5	8	7	5	4	3	4	4	5
Homel. Secur. and Govt. Affairs	0	4	5	5	5	4	4	4	4	3
Comm., Sci., and Transport	6	6	6	5	5	5	4	4	0	0

(b)

Figure 1: Heatmaps of (a) the number of hearings and (b) the average number of witnesses by committee and congressional session


Mr. BROOKS. Thank you, Mr. Chairman. I yield back the balance of my time.

Mr. LANGEVIN. Thank you, Mr. Brooks.

Ms. Slotkin is now recognized for 5 minutes.

Ms. SLOTKIN. Great. Thanks for being here and for doing this. I think what you're hearing from us, we're sort of—both sides of the aisle, frankly, asking different versions of the same question, which is, we have all heard testimony in the last Congress.

Figure 2: An excerpt from a transcript of a congressional hearing

ing which we do not consider as it is not relevant to our present work.

3.1 Dataset Curation Methodology

Curating our dataset consisting of individual utterances together with linguistic features and identifying questions and answers from transcripts in plain text involves the following tasks.[1]

Task 1: Identify Utterances.

We identify individual utterances in a transcript through our pipeline summarized in Figure 3. This begins with a preprocessing step where we identify the beginning and end of the proceedings using rules developed through careful observation of the structure of the transcript, allowing us to dispose of parts of the transcript that convey redundant metadata. We observed that the proceedings usually began with stating the chamber of congress hosting the hearing, i.e., House of Representatives or hours of Senate. Followed by key words such as "committee met" or "subcommittees met". We use these keywords to flag the beginning of the proceedings. The end of the proceedings is flagged either by the end of file or the Appendix. With this we are able to isolate the part of the transcript which is relevant to us.

Our next step is to identify the beginnings and endings of an utterance. We use the word "utterance" to mean the consecutive words spoken by a single member or witness. As seen in Figure 2, the proceedings of the hearing in the transcript are in dialogue format where the name of the speaker (member or witness) precedes their utterance. The challenge, however, is that the standards for format of these markers varies by committee and the time at which the transcript was produced. For example, in identifying the speaker of an utterance, typically only the last name is used, preceded by an honorific. However, in some transcripts, the full name is used, sometimes followed by the speaker's affiliation. Names may be in sentence case or fully capitalized. We therefore take a hybrid approach of carefully designed rules based on the the rules of procedure and the typical structure of transcripts, coupled with automated Named Entity Recognition (NER) to identify these markers. NER, discussed below involves identifying and classifying entities, such as names of people, organizations, locations, dates, and other specific items within a text.

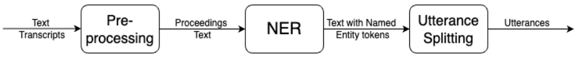


Figure 3: The utterance identification pipeline.

Named Entity Recognition. We use a combination of automated and heuristic NER techniques. The automated NER consists of a BERT-based NER due to Devlin et al. (2018). To augment

[1] https://tinyurl.com/3yudwner

this, we develop heuristic NER rules by careful observation of the congressional transcripts and the metadata which we implement using regular expressions. For example, a common pattern is that utterances begin with an honorific title followed by a name. Names of persons present at the hearing and their roles and affiliations can typically be found in the metadata associated with the hearing or as part of the Propublica Congress API https://projects.propublica.org/api-docs/congress-api/members/. Therefore, named entities appearing at the beginning of utterances can be tagged as either a member of the committee or a witness.

The code and documentation detailing all of the steps will be made publicly available if the paper is accepted. Code to reproduce results in this submission can be found at https://anonymous.4open.science/r/CongressHearings-6B0F/README.md and data can be found at https://tinyurl.com/3yudwner.

Task 2: Compute Features.

We compute some human engineered features derived from expert understanding of linguistic analysis: the NELA (News Ecosystem Analysis) (Horne et al., 2019) features, which were designed to assess the credibility and reliability of news sources and content. or may be computed by clustering commonly occurring patterns in the text. These linguistic features are characteristics of language that can be analyzed to understand patterns in text, such as syntax, semantics, and sentiment. These features are computed across the following liguistic attributes:

- **Style** These indicate the style and structure of the utterances, including POS tags, symbols, and punctuation usage.
- **Complexity** These indicate the readability and understandability of the utterances such as the lexical diversity, reading difficulty, length of words and sentences.
- **Affect** These indicate the sentiment scores of the utterances.
- **Bias** These are indicative of bias in the language like used of hedges, factives, opinions, assertatives, etc.
- **Event** indicates the mention of dates and locations.

The NELA toolkit's various feature computations can be leveraged to gain insight into the utterances in the congressional hearings and can be further used in differentiating different political affiliations, political standings, and quality of discourse.

Task 3: Identify Questions and Answers.

Next, we identify question and answer utterances and focus particularly on the question and answer sessions that follow the statements by the witnesses. We take a supervised transfer learning approach and use a pretrained BERT model described in Figure 4 from Huggingface https://huggingface.co/google-bert/bert-base-cased coupled with a dropout layer and a linear layer with ReLU activation to classify the utterances as questions and answers. Since the transcripts are not annotated, we train our model using a training dataset constructed from the Reddit Ask Me Anything (r/AMA) community and the U.K. Parliamentary dataset, which at a high level, consist of text that is easily identifiable as either questions or answers. More details can be found in Appendix B.

3.2 Dataset Curation Results

Task 1: Identify Utterances.

We evaluated the efficacy of our utterance identification manually as follows. First, we sample 50 hearings from each congressional session uniformly at random. Then, from each of these hearings, we sample 10 utterances from the output of our utterance identification pipeline, for a total of 500 utterances from each congressional session. We then manually verify whether each each of these utterances were indeed correctly identified.

Overall, 93.96% of the utterances identified by our pipeline were correct in our manual verification. The remaining incorrectly split utterances occur as either two utterances that are clubbed together or one utterance that is split into two. Table 1 summarizes our findings for the task of identifying utterances. In total, our dataset consists of 3,319,386 utterances.

We observe that the house committees on Financial Services, Oversight and Government Reform, and Energy and Commerce have the largest number of utterances per hearing on average (see Figure 5 (a)), while having relatively fewer number of words per utterance (Figure 5 (b)). In contrast, the senate Commerce, Science and Transport committee and the house Homeland Security committee have

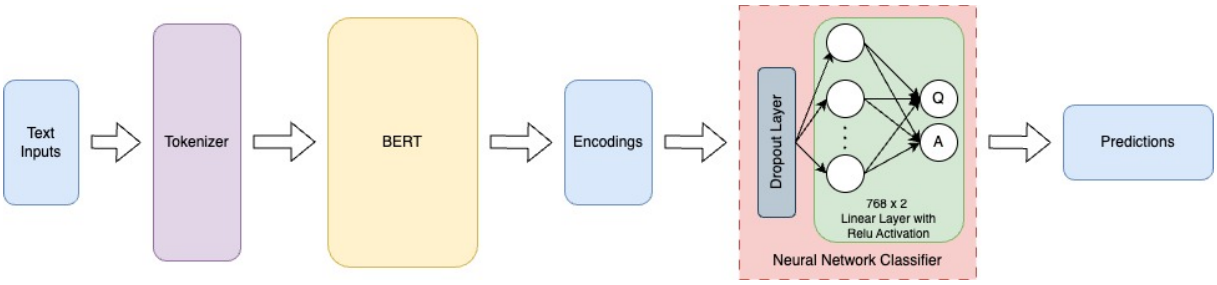
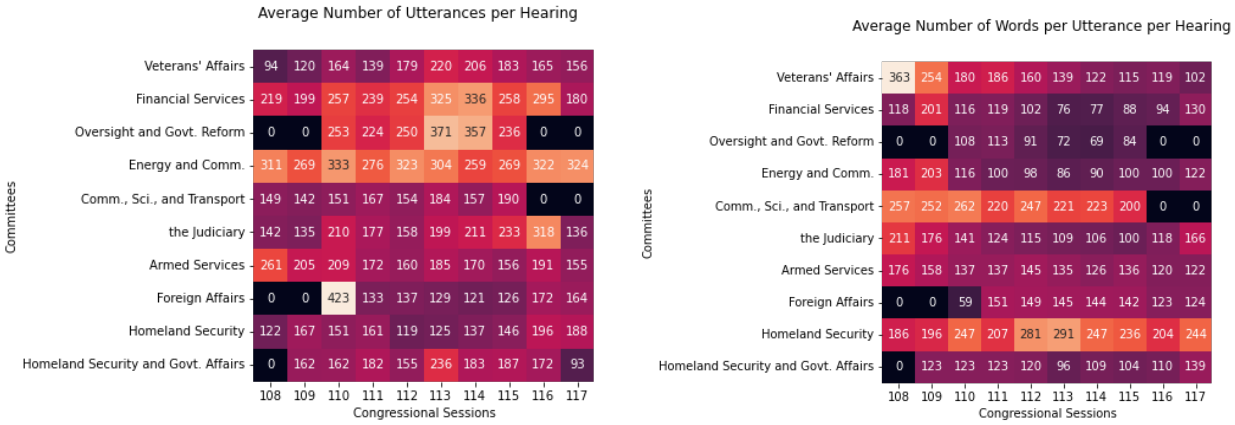


Figure 4: Architecture of the BERT classifier model



(a)

(b)

Figure 5: Heatmaps of (a) the number of utterances per hearing and (b) the average number of words per utterance, across committees and congressional sessions

fewer utterances but on average, a higher number of words per utterance.

Session	#Incorrect	#Clubbed	#Broken
108	32	12	20
109	16	8	8
110	20	15	5
111	12	8	4
112	4	2	2
113	39	36	3
114	27	18	9
115	23	18	5
116	52	40	12
117	77	69	8
Total	**302**	**226**	**76**

Table 1: Performance on Task 1 for utterance identification. In each congressional session, we sample 50 hearings, and sample 10 sequences identified as utterances by our method from each hearing, and manually verified. The column indicates a type of mistake. For example, over 10 sessions and 5000 manually verified utterances, 302 were incorrect.

Task 3: Identify Questions and Answers.

To evaluate our question and answer annotations, we sampled 800 utterances uniformly at random from among all utterances that occurred between the 114th through to the 117th congressional sessions. These 800 utterances were then hand-labeled, yielding 379 questions and 421 answers. We achieved an accuracy of 87.14% at the question-answer labeling task using our BERT-based classifier. Table 2 shows the confusion matrix for this task. Overall, 291 and 380 out of the 379 and 421 question and answer utterances respectively were classified correctly.

Session		**114**	**115**	**116**	**117**	**114–117**
Questions	True	20	105	101	65	291
	False	6	25	29	28	88
Answer	True	26	127	134	93	380
	False	0	3	3	5	11
Accuracy		**0.88**	**0.89**	**0.87**	**0.83**	**0.87**

Table 2: Performance on Task 3, identifying question and answer utterances, against 800 human labeled utterances. For each session indicated by a column, a row entry provides the number of utterances that were either correctly (True) or incorrectly (False) identified as a question or answer. For example, over sessions 114-117, 291 out of 379 utterances identified as questions by our classifier were truely question utterances.

4 Identifying Party Affiliation and Standing

4.1 Methodology

In order to accomplish this, we use both BERT-based classifier (similar to the one described in Figure 4) as well as classical machine learning classification techniques using NELA features (Horne et al., 2019). Similarly, we will also predict the *standing* of the party the questioner is affiliated to, whether the party may either be in the majority or minority in the House or Senate.

Our BERT based classifier consists of the BERT encoder followed by a droppout layer and a linear layer (768*2) with LeakyReLU activation. We used a learning rate of $1e^{-5}$ and chose the other hyperparameters of batch size and number of epochs to train through 5-fold cross-validation. We used the cross-entropy loss as our loss function and Adam optimizer.

For the NELA feature classification, we used the Random Forest classifier. We did a 5-fold cross-validation grid search over the n-estimators, max-depth, and min-sample split hyperparameters. We repeated this grid-search exhaustively for all combinations of Congressional session and Congressional Committee.

4.2 Results

We achieve an accuracy of **92.2%** in identifying the party affiliation of the questioner in the question utterances and an accuracy of **75%** in identifying the party standing (majority or minority) of the questioner using the BERT based classifier discussed in Section 4. Furthermore, the results for identification of party affiliation or standing of the questioner of the answer utterances were 58% and 63%, respectively. We conjecture that the identification of the party affiliation or the party standing of an answer utterance is more nuanced than that of the question utterances as we are trying to identify the party affiliation of the questioner of the question utterance in response to which the answer utterance had been spoken. Since results from LLMs like BERT are harder to interpret and derive insights from, we also use models that lend themselves to further in-depth post-hoc analysis in conjunction with more interpretable for the same prediction task. To this end, we input the NELA features into a Random Forest Classifier. Table 3 shows the results of this in greater detail.

	Question		Answer	
	Affiliation	Standing	Affiliation	Standing
Base	0.51	0.60	0.51	0.60
NELA	0.54	0.61	0.52	0.61
BERT	0.92	0.75	0.58	0.63

Table 3: Results for identification of the party affiliation and standing of the questioner of an utterance. For answer utterances, affiliation and standing labels refer to the party affiliation and standing of the questioner.

We further partitioned the dataset by sessions and committees. Table 4 shows the top five best combinations of the congressional session and committee on which random forest classifier using NELA features obtains the highest accuracy at classifying party affiliation, illustrating the efficacy of this approach. The Random Forest classifier on NELA features does better on the special committees. Further analysis is needed to explore this.

We used our BERT classifier to classify the question top 10 committees with the highest number of question utterances as shown in Table 5. We used the question utterances as is as well as question-utterances without stop-words for this and the classifier performed very well. We used batch-sizes of 2 and 64 and chose the classifier model which performed better on our validation set. The train, balidtaion and test sets were split in a ratio of 0.2, 0.05 and 0.75. We observed the classifier performing very well over these two version of the dataset. We hypothesised that this may have been due to the BERT classifier memorizing the names of the speakers as both these versions of the utterances begin with the name of the speaker. We tested this hypothesis by using the BERT classifier with the same architecture over the question utterances with the speaker name now removed. The 5th column in Table 5 shows that while the BERT classifier was not able to perform as well as with the utterances containing the speaker name, it was still outperforming the base classifier on almost all the committees depicting that there is indeed a difference between the linguistics used by the two parties.

To further bolster our understanding of the differences in the use of various linguistic cues by members of different parties, we conducted a two-sample Kolmogorov-Smirnov test across the NELA features. Here, we have a sample for question utterances from members of each party, and the null hypothesis is that both samples are drawn from the same distribution. Our main findings are that, on average, both Democrats and Republicans use more complex language than Independents, as shown in

Random Forest	Majority Classifier	Session	Committee
1	0.548	109	Committee on Financial Services and Committee on Resources
0.81	0.52	113	Committee on Commerce, Science, and Transportation
0.79	0.56	110	Committee on Education and Labor and Committee on Health, Education, Labor, and Pensions
0.75	0.54	113	Committee on Armed Services Meeting Jointly with Subcommittee on Asia and the Pacific of the Committee on Foreign Affairs
0.72	0.50	117	Committee on House Administration

Table 4: Top results of identifying party affiliations on Question utterances after partitioning the dataset by congressional session and committee using the Random Forest Classifier. The Random Forest Classifier tends to do better on special committees.

All		No Stop Words		No Speaker		Committee on
BERT	Base	BERT	Base	BERT	Base	
98.31	54.21	98.40	54.20	64.48	54.75	Energy and Commerce
98.73	56.72	98.72	56.67	65.39	56.40	Financial Services
97.88	54.88	97.87	55.05	62.54	55.19	the Judiciary
98.80	69.48	98.84	69.42	72.48	70.21	Oversight and Government Reform
95.06	54.69	94.91	54.92	64.69	66.47	Ways and Means
98.43	63.71	98.18	63.59	67.13	63.99	Homeland Security and Governmental Affairs
98.94	61.34	99.07	61.43	60.95	60.54	Commerce, Science, and Transportation
95.82	51.75	95.68	51.62	59.27	51.20	Armed Services
98.82	50.45	99.03	50.10	61.04	50.32	Veterans' Affairs
97.10	52.91	97.01	53.38	60.00	50.80	Foreign Affairs

Table 5: Question utterance party affiliation prediction accuracy by session and committee by a BERT classifier

Figure 9 in Appendix C. The Democrats use more positive sentiment words, while the Republicans tend to use more neutral sentiment words in their utterances, as shown in Figure 11 in Appendix C. Figure 10 in Appendix C shows that the Republicans use more assertive and hedge words while the Democrats tend to use more implicative and positive opinion words.

The hues of each cell represent the difference between the means of the two distributions under study. If the difference between the two distributions is not statistically significant, the cell has been hatched with the crossed pattern. Here, the R stands for Republican, D for Democrat, I for Independent, M for Majority party, m for Minority party, R.M. for Republican in Majority, and D.M. for Democrat in Majority.

5 Discussion, Summary and Limitations

Committee reports are made publicly available and will help shed further light on the activities and responsiveness of committees on specific issues which is an exciting avenue of further research at the interface of computer science with the scholarship in the political economy. We plan to maintain our dataset and tools to serve as a resource for future work and stay up to date with the latest committee hearings as they are released, and in future work, expand our collection to public hearings from other democratic institutions.

Limitations. As of the time of writing, our dataset only includes hearings from the 108th to the 117th congressional sessions. As the release of transcripts is an ongoing process, we will expand our collection to include the latest committee hearings.

References

Javier Arregui and Clément Perarnaud. 2022. A new dataset on legislative decision-making in the european union: the deu iii dataset. *Journal of European Public Policy*, 29(1):12–22.

Giovanni Maria Biancofiore, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Fedelucio Narducci. 2024. Interactive question answering systems: Literature review. *ACM Computing Surveys*, 56(9):1–38.

Lukas Birkenmaier, Laureen Sieber, and Felix Bergstein. 2025. Polinterviews – a dataset of german politician public broadcast interviews. *Preprint*, arXiv:2501.04484.

Austin Bussing and Michael Pomirchy. 2022. Congressional oversight and electoral accountability. *Journal of Theoretical Politics*, 34(1):35–58.

Collin Coil, Caroline Bruckner, Natalie Williamson, Karen O'Connor, and Jeff Gill and. 2024. Still underrepresented? gender representation of witnesses at house and senate committee hearings. *Journal of Women, Politics & Policy*, 45(4):429–445.

Paul M Collins Jr and Lori A Ringhand. 2016. The institutionalization of supreme court confirmation hearings. *Law & Social Inquiry*, 41(1):126–151.

Roger H Davidson, Walter J Oleszek, Frances E Lee, Eric Schickler, and James M Curry. 2023. *Congress and Its Members*. CQ Press.

Ryan Delano, Aaron Rudkin, and In Song Kim. Bulk ingestion of congressional actions and materials dataset.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alissa M Dolan, Elaine Halchin, Todd Garvey, Walter J Oleszek, and Wendy Ginsberg. 2014. *Congressional oversight manual*. Congressional Research Service Washington, DC.

Kevin M. Esterling. 2007. Buying expertise: Campaign contributions and attention to policy analysis in congressional committees. *American Political Science Review*, 101(1):93–109.

Dana R. Fisher, Joseph Waggle, and Philip Leifeld. 2013. Where does political polarization come from? locating polarization within the u.s. climate change debate. *American Behavioral Scientist*, 57(1):70–92.

Jon H Fiva, Oda Nedregård, and Henning Øien. 2025. The norwegian parliamentary debates dataset. *Scientific Data*, 12(1):4.

Caroline Fredrickson. 2019. The kavanaugh hearings and the search for a just justice submission. *Golden Gate University Law Review*, 49:67.

James Hamilton, Robert F Muse, and Kevin R Amer. 2007. Congressional investigations: Politics and process. *American Criminal Law Review*, 44:1115.

Mollie Hemingway and Carrie Severino. 2019. *Justice on Trial: The Kavanaugh Confirmation and the Future of the Supreme Court*. Simon and Schuster.

Arnav Hiray, Yunsong Liu, Mingxiao Song, Agam Shah, and Sudheer Chava. 2024. CoCoHD: Congress committee hearing dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15529–15542, Miami, Florida, USA. Association for Computational Linguistics.

Bai Linh Hoang. 2025. "are you too busy to listen up?" legislative (dis)engagement from constituents in local public meetings. *Political Communication,*, 0(0):1–20.

Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust fake news detection over time and attack. *ACM Trans. Intell. Syst. Technol.*, 11(1).

Ralph K Huitt. 1954. The congressional committee: A case study. *American Political Science Review*, 48(2):340–365.

Taneesh Kaur. 2022. Conversation analysis in a us senate judiciary hearing: Questioning brett kavanaugh. *Discourse Studies*, 24(4):423–444.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Carl Levin and Elise J Bean. 2018. Defining congressional oversight and measuring its effectiveness. *Wayne Law Review*, 64:1.

Jonathan Lewallen, Ju Yeon Park, and Sean M Theriault. 2024. The politics of problems versus solutions: Policymaking and grandstanding in congressional hearings. *Policy Studies Journal*, 52(3):515–531.

Jonathan Lewallen, Sean M Theriault, and Bryan D Jones. 2016. Congressional dysfunction: An information processing perspective. *Regulation & Governance*, 10(2):179–190.

Thomas V Maher, Charles Seguin, Yongjun Zhang, and Andrew P Davis. 2020. Social scientists' testimony before congress in the united states between 1946-2016, trends from a new dataset. *Plos one*, 15(3):e0230104.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online. Association for Computational Linguistics.

Walter J Oleszek, Mark J Oleszek, Elizabeth Rybicki, and Bill Heniff Jr. 2015. *Congressional procedures and the policy process*. CQ press.

John Ray. 2018. Walk this way, talk this way: Legislator speech and lobbying. *Interest Groups & Advocacy*, 7:150–172.

William G Ross. 1998. The senate's constitutional role in confirming cabinet nominees and other executive officers. *Syracuse Law Review*, 48:1123.

Richard C Sachs. 2003. *A guide to congressional hearings*. Nova Publishers.

Jeffery C Talbert, Bryan D Jones, and Frank R Baumgartner. 1995. Nonlegislative hearings and policy change in congress. *American Journal of Political Science*, pages 383–405.

Robert Thomson, Javier Arregui, Dirk Leuffen, Rory Costello, James Cross, Robin Hertz, and Thomas Jensen and. 2012. A new dataset on decision-making in the european union before and after the 2004 and 2007 enlargements (deuii). *Journal of European Public Policy*, 19(4):604–622.

Robert Thomson and Frans N. Stokman. 2006. *Research design: measuring actors' positions, saliences and capabilities*, page 25–53. Political Economy of Institutions and Decisions. Cambridge University Press.

David Bicknell Truman. 1951. *The governmental process: Political interests and public opinion*. Alfred A Knopf Inc.

Michelle Wolfe. 2012. Putting on the brakes or pressing on the gas? media attention and the speed of policy-making. *Policy Studies Journal*, 40(1):109–126.

John T Woolley and Gerhard Peters. The american presidency project. *Santa Barbara, CA: University of California (hosted), Gerhard Peters (database). Available from World Wide Web:(http://www. presidency. ucsb. edu/ws*.

Tianliang Xu, Eva Maxfield Brown, Dustin Dwyer, and Sabina Tomkins. 2025. Publicspeak: Hearing the public with a probabilistic framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28520–28529.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems*, 64(12):3151–3195.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

A Abbreviations

Feature Name	Description
quotes	Number of quotations
exclaim	Number of exclamation (!) symbols
allpunc	Number of punctuations
allcaps	Number of capitalized words
stops	
CC	conjunction, coordinating
CD	cardinal number
DT	determiner
EX	existential there
FW	foreign word
IN	conjunction, subordinating or preposition
JJ	adjective (English), other noun-modifier (Chinese)
JJR	adjective, comparative
JJS	adjective, superlative
LS	list item marker
MD	verb, modal auxiliary
NN	noun, singular or mass
NNS	noun, plural
NNP	noun, proper singular
NNPS	noun, proper plural
PDT	predeterminer
POS	possessive ending
PRP	pronoun, personal
PRP$	pronoun, possessive
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	adverb, particle
SYM	symbol
TO	infinitival “to”
UH	interjection
WP$	wh-pronoun, possessive
WRB	wh-adverb
VB	verb, base form
VBD	verb, past tense
VBG	verb, gerund or present participle
VBN	verb past participle
VBP	verb, non-3rd person singular present
VBZ	verb, 3rd person singular present
WDT	wh-determiner
WP	wh-pronoun, personal

Table 6: Description of the **Style category of NELA Features. The description of the parts-of-speech features are found in SpaCy’s glossary `https://github.com/explosion/spacy/blob/master/spacy/glossary.py`**

Feature Name(Abbreviated)	Description
ttr	Lexical Diversity also known as Type-Token Ratio
avgWlen	Average number of characters in a word
wCount	Number of words
FKGLvl	**Flesch–Kincaid grade level**: Standard readability measure computed by $0.39 * \frac{totalwords}{totalsentences} + 11.8 * \frac{totalsyllables}{totalwords} - 15.59$
SmgIn	**Smog Index**: Standard readability measure computed by $1.0430 * \sqrt{\#polysyllables * \frac{30}{\#sentences}} + 3.1291$
CLIn	**Coleman–Liau index**: Standard readability measure computed by $0.0588 * L - 0.296 * S - 15.8$ where L = avg # letters per 100 words and S = avg # sentences per 100 words
lix	**LIX**: Standard readability measure computed by $S + L$ where S = average sentence length and L = percentage of words with more than 6 letters. The scores usually range from 20 to 60.

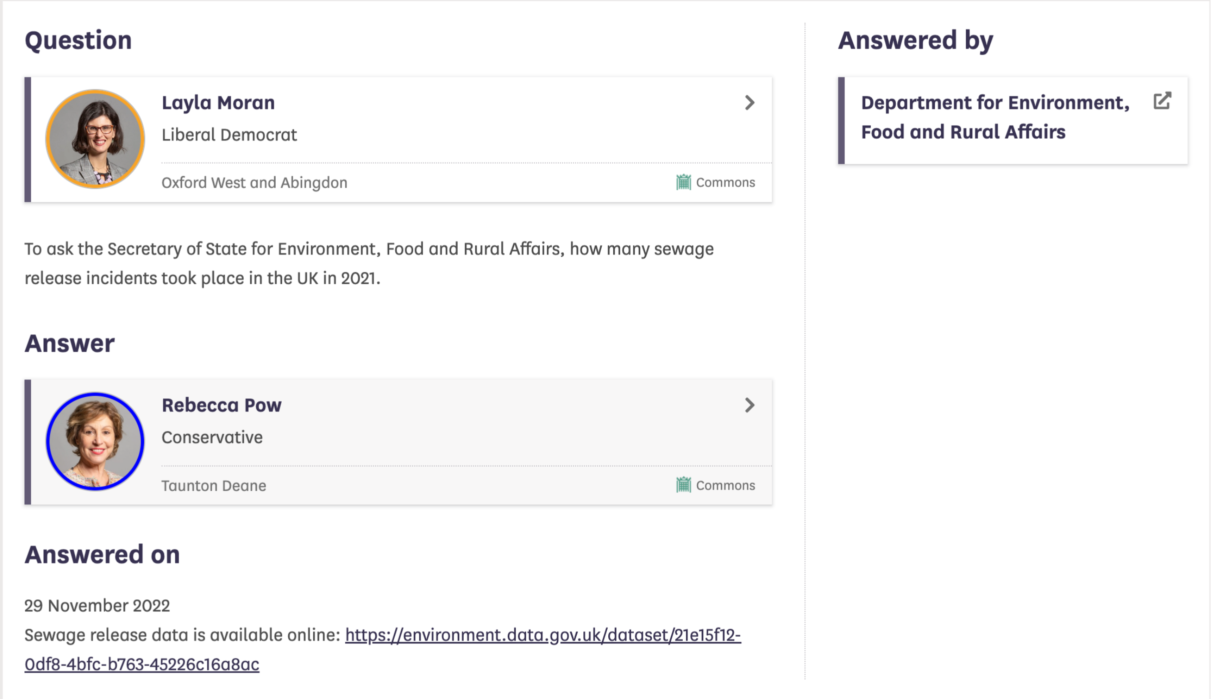
Table 7: Description of the **Complexity category of NELA Features**

Feature Name	Description
vneg	Negative sentiment score using Vadar Sentimet
vneu	Neutral sentiment score using Vadar Sentiment
vpos	Positive sentiment score using Vadar Sentiment
wneg	Number of weak negative words
wpos	Number of weak positive words
wneu	Number of weak neutral words
sneg	Number of strong negative words
spos	Number of strong positive words
sneu	Number of strong neutral words

Table 8: Description of the **Affect category of NELA Features**

Feature Name	Description
bias	The number of bias words
assert	the number of assertive verbs
facts	The number of factive verbs
hedges	The number of hedge words
implctv	The number of implicatives
repVerb	Count of report verbs
poWords	number of positive pinion words
noWords	number of negative opinion words

Table 9: Description of the **Bias category of NELA Features**



Question

Layla Moran
Liberal Democrat
Oxford West and Abingdon
Commons

To ask the Secretary of State for Environment, Food and Rural Affairs, how many sewage release incidents took place in the UK in 2021.

Answer

Rebecca Pow
Conservative
Taunton Deane
Commons

Answered on

29 November 2022
Sewage release data is available online: https://environment.data.gov.uk/dataset/21e15f12-0df8-4bfc-b763-45226c16a8ac

Answered by

Department for Environment, Food and Rural Affairs

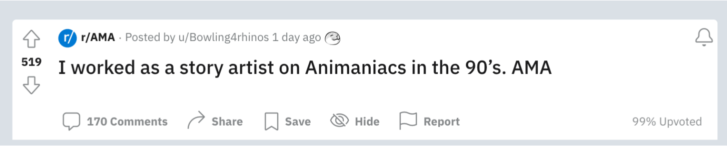
Figure 6: Answered-Question in the UK Parliament

B Identifying Questions and Answers

B.1 Datasets used for transfer learning

U.K. Parliamentary Question Hour The U.K. Parliamentary written answered questions are an important aspect of the democratic process in the United Kingdom. Members of Parliament (M.P.s) can submit written questions to government ministers on any issue within their portfolio. These questions are typically answered in writing within a set timeframe and are publicly available online on the U.K. Parliament's official website. This provides an opportunity for M.P.s to hold ministers accountable for their actions, even when Parliament is not in session. This dataset also provides an excellent source for labeled examples of questions and answers. We have used 4688 such examples (2344 questions and 2344 answers) to test our question-answer prediction model. Figure 6 is an example of an answered question from the House of Commons.

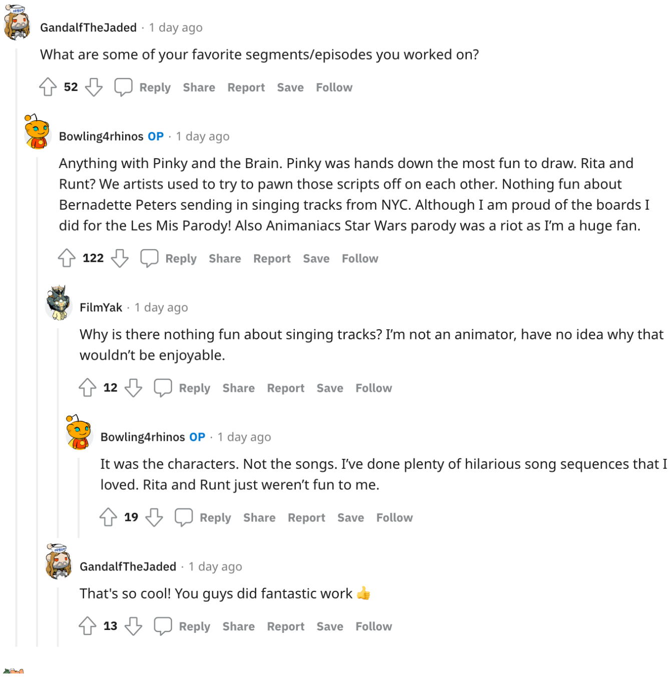
r/AskMeAnything The subreddit AskMeAnything, also known as AMA, is a popular online platform on Reddit that allows users to host a question-answering session that allows other users to ask the host any question. Hosts may include celebrities, politicians, scientists, athletes, and everyday individuals with unique experiences or perspectives to share. AMA sessions are typically hosted in a question-answer format, where the person answering questions, also known as the "host," responds to questions posted by users in real time. This makes the posts a very rich source of labeled question-answer examples. Each post is an introduction where the expert introduces themselves; Figure 7 is an example of such a post. Figure 8 is an example of a comment on the post in Figure 7. The first level of comment on these posts is the questions, and the second level of comments is the answer to the questions. This is how the dataset has been labeled. This dataset has 121512 examples, 60756 Questions, and 60756 Answers. We used it to train our questions-answer prediction model.



r/AMA · Posted by u/Bowling4rhinos 1 day ago

I worked as a story artist on Animaniacs in the 90's. AMA

Figure 7: An Ask Me Anything (AMA) post in the subreddit r/AMA



GandalfTheJaded · 1 day ago

What are some of your favorite segments/episodes you worked on?

Bowling4rhinos OP · 1 day ago

Anything with Pinky and the Brain. Pinky was hands down the most fun to draw. Rita and Runt? We artists used to try to pawn those scripts off on each other. Nothing fun about Bernadette Peters sending in singing tracks from NYC. Although I am proud of the boards I did for the Les Mis Parody! Also Animaniacs Star Wars parody was a riot as I'm a huge fan.

FilmYak · 1 day ago

Why is there nothing fun about singing tracks? I'm not an animator, have no idea why that wouldn't be enjoyable.

Bowling4rhinos OP · 1 day ago

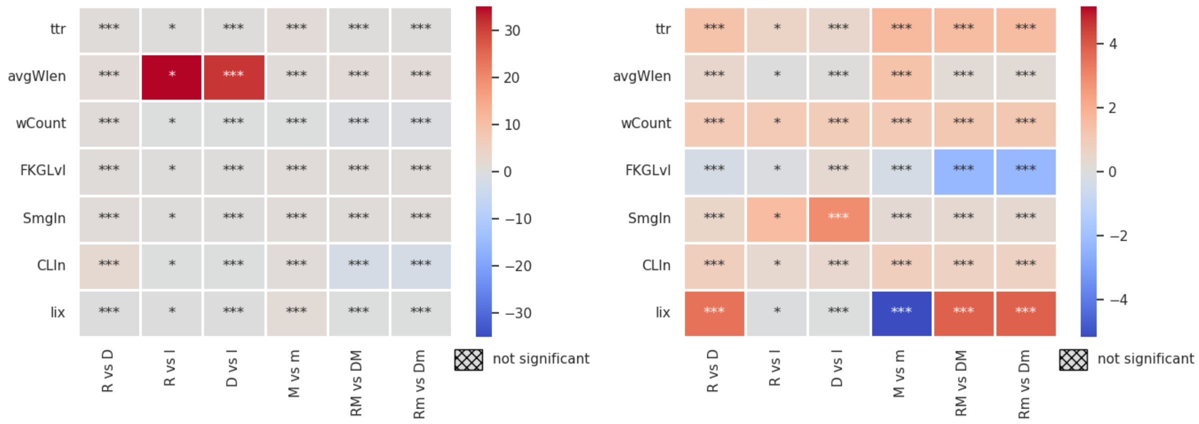
It was the characters. Not the songs. I've done plenty of hilarious song sequences that I loved. Rita and Runt just weren't fun to me.

GandalfTheJaded · 1 day ago

That's so cool! You guys did fantastic work 👍

Figure 8: The subsequent comments on the AMA post in 7

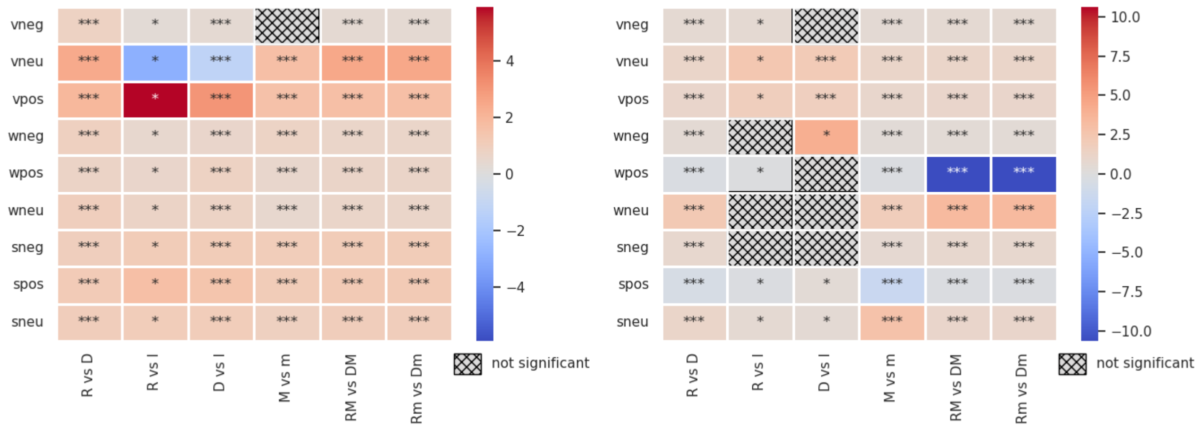
C Statistical Testing with NELA Features



(a) Statistical Testing of the Answer Utterances

(b) Statistical Testing of the Question Utterances

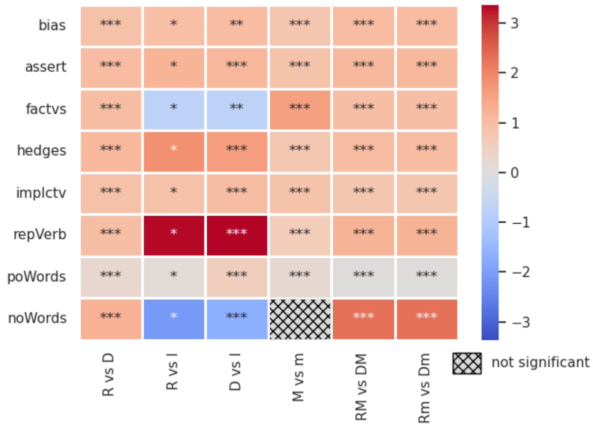
Figure 9: Statistical Testing of the Complexity NELA Features using two-sample Kolmogorov-Smirnov test. The hue of each cell signifies the ratio of the means of the two samples being studied; red hue signifies that the left sample is larger where as the blue signifies that the right side sample has a larger mean. The number of asterisks ‘*’ indicates the value of confidence for that statistical test. ‘***’ $\implies p \in [0, 0.001)$, ‘**’ $\implies p \in [0.001, 0.01)$, ‘*’ $\implies p \in [0.01, 0.05)$. Differences that are not statistically significant are indicated with a cross-hatched pattern.



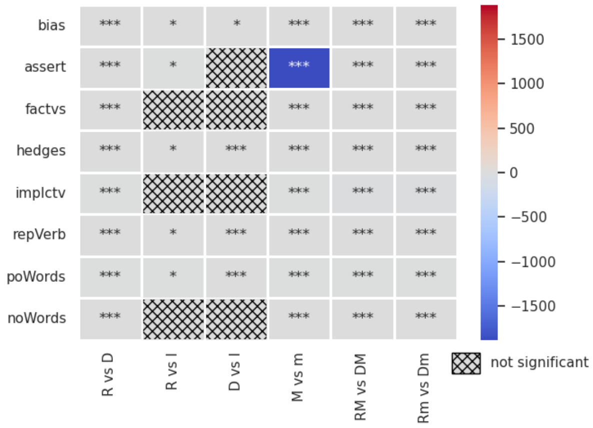
(a) Statistical Testing of the Answer Utterances

(b) Statistical Testing of the Question Utterances

Figure 10: Two-sample Kolmogorov-Smirnov test on the Affect NELA Features using. The hue of each cell signifies the ratio of the means of the two samples being studied; red hue signifies that the left sample is larger where as the blue signifies that the right side sample has a larger mean. The number of asterisks ‘*’ indicates the value of confidence for that statistical test. ‘***’ $\implies p \in [0, 0.001)$, ‘**’ $\implies p \in [0.001, 0.01)$, ‘*’ $\implies p \in [0.01, 0.05)$. Differences that are not statistically significant are indicated with a cross-hatched pattern.



(a) Statistical Testing of the Answer Utterances



(b) Statistical Testing of the Question Utterances

Figure 11: Two-sample Kolmogorov-Smirnov test on the Bias NELA Features using. The hue of each cell signifies the ratio of the means of the two samples being studied; red hue signifies that the left sample is larger where as the blue signifies that the right side sample has a larger mean. The number of asterisks '*' indicates the value of confidence for that statistical test. '***' $\implies p \in [0, 0.001)$, '**' $\implies p \in [0.001, 0.01)$, '*' $\implies p \in [0.01, 0.05)$. Differences that are not statistically significant are indicated with a cross-hatched pattern.