Question Type Prediction in Natural Debate

Zlata Kikteva Alexander Trautsch Steffen Herbold Annette Hautli-Janisz
Faculty of Computer Science and Mathematics
University of Passau
firstname.lastname@uni-passau.de

Abstract

In spontaneous natural debate, questions play a variety of crucial roles: they allow speakers to introduce new topics, seek other speakers' opinions or indeed confront them. A three-class question typology has previously been demonstrated to effectively capture details pertaining to the nature of questions and the different functions associated with them in a debate setting. We adopt this classification and investigate the performance of several machine learning approaches on this task by incorporating various sets of lexical, dialogical and argumentative features. We find that BERT demonstrates the best performance on the task, followed by a Random Forest model enriched with pragmatic features.

1 Introduction

Questions are at the core of human communication and can be used in a variety of ways, from simply eliciting information from the hearer to communicating a speaker's standpoint. They are also crucial in argumentation, where they make up around 5% of all speech acts and are rarely left ignored (Kikteva et al., 2022). However, question type prediction faces a number of challenges. First, the lexical surface does not correlate with the function of the question (e.g., 'Who would do that?' can either be a request for information on a set of entities or a rhetorical question communicating that no one would do that). Secondly, context is assumed to be crucial for their interpretation, however, it is not exactly clear what features in the context are indeed relevant. And lastly, a large majority of computational work assumes a bipartite distinction into information-seeking and rhetorical questions, a classification that does not capture the variety of functions that questions fulfil in debate. There is, in fact, a third category of questions referred to as assertive questions that has been theoretically motivated (Freed, 1994) and empirically tested (Visser et al., 2020; Hautli-Janisz et al., 2022b). Such questions are characterised by the speaker's intention to express their opinion while still seeking information (as opposed to information-seeking questions the main purpose of which is to elicit a response, and the rhetorical ones which do not necessitate one).

With this paper, we examine the impact of different (1) context configurations and (2) combinations of carefully selected lexical, pragmatic and argumentative features for the question type prediction. We explore deep learning approaches that tend to effectively capture lexical features as well as statistical models that while still capable of representing lexical information, also benefit from categorical feature inputs. Our results indicate the introduction of the third question type drastically increases the complexity of the task when compared to the binary classification. We find that BERT achieves the highest scores when the input is enriched with lexical features of either the preceding material or response. Furthermore, we report our second-best results with a Random Forest model that makes use of a rich pragmatic feature set.

2 Previous work

There is a significant body of work on questions in computational linguistics in the context of question-answering systems, i.e., question classification based on the type of information expected as an answer, which focuses predominantly on factoid questions. Earlier approaches include extensive use of lexical and syntactic features combined with traditional statistical approaches (Zhang and Lee, 2003; Metzler and Croft, 2005; Huang et al., 2008; Silva et al., 2011; Loni, 2011; Tayyar Madabushi and Lee, 2016), while recently more work utilizes extensive capabilities of the deep learning models (Kim, 2014; Iyyer et al., 2014; Sun et al., 2019; Anhar et al., 2019; Yilmaz and Toklu, 2020)

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 624–630
September 18–20, 2024. ©2024 Association for Computational Linguistics

Questions in natural communication, however, are often used to elicit more than just factual information from interlocutors and instead serve a variety of communicative purposes. Thus, to interpret the function of questions in discourse, researchers often adhere to a bipartite distinction: Harper et al. (2009) focus on identification of conversational versus factual questions, Bhattasali et al. (2015); Ranganath et al. (2016); Oraby et al. (2017) distinguish rhetorical questions from non-rhetorical ones; Kalouli et al. (2018, 2021) identifies information and non-information-seeking questions; Bagga et al. (2021) categorize questions into unpalatable and not unpalatable ones from the perspective of abusive language detection. They adopt a range of approaches such as the use of lexical features like n-grams, POS tags, speaker roles and word embeddings as well as the modelling of the context surrounding questions.

However, recent research in pragmatics suggests that question typology is a bit more complex than previously assumed. In their work, Hautli-Janisz et al. (2022a) and Kikteva et al. (2022) discuss a more fine-grained question typology that attempts to better capture the conversational functions that are fulfilled by questions in debate. In this work, we follow their distinction into information-seeking, rhetorical and assertive questions.

Information-seeking questions Also called pure questions (PQs), those are used to elicit information from an interlocutor. For instance, in Example 1, moderator Fiona Bruce seeks the views of the panel members on the matter of the voter ID in the UK.[1]

(1) Fiona Bruce: Will voter IDs protect the integrity of elections or just undermine the UK democracy?

Rhetorical questions With RQs, speakers express their own opinion or standpoint, illustrated in Example 2 by Liz Saville-Roberts's intention to communicate her dissatisfaction with the current state of the prison system in the UK.[2] The speaker poses the question without expecting to hear a response which is signalled by the fact that she continues talking.

(2) Liz Saville-Roberts: *The black population in Wales is over-represented by five times within the prison population of Wales, surely that is a desperate failure? That is an indication of the racism in our society in action.*

Assertive questions AQs serve the double purpose of communicating information and asking for confirmation/rejection from an interlocutor. In Example 3, Gillian Keegan expresses her frustration regarding the police having to inspect every package due to the new Brexit regulations, while at the same time expecting other panel members to agree with her opinion on the matter.[3]

(3) Gillian Keegan: *The police probably have the legal right to open every packet and inspect every sausage. Isn't that unreasonable?*

3 Data

The data underlying our investigation is taken from QT30 (Hautli-Janisz et al., 2022b), the largest ever dataset of broadcast political debate. The corpus comprises the transcriptions of 30 episodes of the UK's talk show 'Question Time' (QT) between June 2020 and November 2021 and is manually annotated with Inference Anchoring Theory (IAT) (Budzynska et al., 2014, 2016), a framework that captures how argumentation unfolds and is reacted to in dialogue which allows us to extract questions of the three types as well as pragmatic features associated with them for the analysis.

The questions dataset used in the current work contains 2 867 questions, with the split into training and test given in Table 1. PQs make up almost 70% of all questions, both RQs and AQs are significantly less frequent and make up about 14% and 16% of the data, respectively. Questions extracted from QT30 are used for training; an additional 10 episodes of QT that were broadcast and analyzed in 2022 are used for testing. With this time split, we train on about 77% of the data and evaluate on the rest.

Table 1: Training and test split.

	PQ	RQ	AQ	Total
Training	1 555	306	343	2 204
Test	446	87	130	663
Total	2 001	393	473	2 867

[1] http://corpora.aifdb.org/qt30, json ID: 21308

[2] http://corpora.aifdb.org/qt30, json ID: 18464

[3] http://corpora.aifdb.org/qt30, json ID: 23888

4 Question type prediction

4.1 Feature selection

Lexical features Lexical features include questions and corresponding preceding and response texts. We consider one locution, i.e., discourse unit, dialogically preceding the question to be the preceding context, while the response constitutes any number of locutions that are contributed by the same or different speaker than the question speaker, directly following the question until the end of that speaker's turn. We represent them as n-grams for statistical models and as embeddings for deep learning models. For n-grams, we extract all available unigrams, as well as bi- and tri-grams that appear in at least two documents. We further process n-grams to identify the most relevant features by applying TF-IDF vectorization to the data. The vectorization is performed per question type allowing us to model feature representations for each question type separately.

Pragmatic features in the response We extract the following argumentative and pragmatic features from the response to the question:

- **Speaker roles** Information on whether the question and response material comes from the moderator, a panel member or an audience member and on whether the question and response speakers are the same or not.
- **Answers** Statements instantiating the content of the question.
- **Propositional relations** Inference (support between two statements), conflict (attack between two statements) and rephrase (reformulation or refinement of a previous statement).
- **Epistemic markers** Indicators of speaker commitment.

The number of answers, propositional relations, and epistemic markers is normalised on the level of the locution, i.e., we encode the relative frequency of a feature per locution in the response. Speaker roles are represented as categorical values.

4.2 Modeling

Statistical models In order to model both the pragmatic features of the response and lexical features of the question with its adjacent context, we use Random Forest (RF) and Support Vector Machine (SVM) for classification. After hyperparameter tuning, we choose an entropy criterion and 200 estimators with a maximum depth of 8 and minimum sample split of 0.1 for RF; for SVM we use an RBF kernel with a one-vs-one multiclass classification strategy.

Table 2: Balanced accuracies for different context configurations. PREC, QU, and RESP stand for preceding, question, and response texts respectively.

Context	LSTM	BERT	RF	SVM
QU	0.37	0.47	0.40	0.40
PREC-QU	0.31	**0.48**	0.35	0.36
QU-RESP	0.42	**0.48**	0.35	0.35
PREC-RESP	0.34	0.40	0.34	0.36
PREC-QU-RESP	0.36	0.42	0.35	0.37

Language models To gain insight into how deep learning models compare to more traditional machine learning approaches, we use an LSTM model and an LLM. For LSTM we use softmax activation with categorical cross-entropy as a loss function and the Adam optimizer with a batch size of 64, a maximum sequence length of 400 and 100-dimensional embeddings trained over 6 epochs. For an LLM we use a cased, large variant of BERT (Devlin et al., 2018) with 336M parameters which we retrieved from the Huggingface Model Hub.[4] We adopt the same configuration as used by Huggingface to evaluate BERT on the GLUE benchmark (Wang et al., 2019). We train for three epochs, with a learning rate of 2e-05 and a batch size of 32 with a maximum sequence length of 400, which is sufficiently large for all inputs.

Testing In order to mitigate how unbalanced the dataset is, we resort to an oversampling technique for the training set by matching the number of underrepresented RQs and AQs to the number of PQs. For the same reason, for the evaluation of the models' performance, we use balanced accuracy. The code is publicly available at https://github.com/ZlataKikteva/sigdial2024-questions.

5 Results

5.1 Context

We first model the impact of the lexical features in context on the multiclass question type prediction task as it has been observed to improve the

[4] 'bert-large-cased'

Table 3: Balanced accuracies for pragmatic features (in response) and lexical features (question text). The highest scores for each feature set are underlined; the highest score overall is in bold.

		Pragmatic Features				
Model	**Feature Set**	**Speakers**	**Answers**	**Prop.rel.**	**Ep. markers**	**All**
RF	Pragmatic Features only	0.41	0.43	0.38	0.37	0.40
	Pragmatic Features & Unigrams	0.41	**0.44**	0.43	0.42	0.42
	Pragmatic Features & Uni- and Bigrams	0.42	0.43	0.42	0.41	0.43
	Pragmatic Features & Uni-, Bi and Trigrams	0.41	0.42	0.41	0.41	0.43
SVM	Pragmatic Features only	0.42	0.43	0.38	0.37	0.37
	Pragmatic Features & Unigrams	0.40	0.35	0.36	0.34	0.41
	Pragmatic Features & Uni- and Bigrams	0.40	0.35	0.35	0.34	0.42
	Pragmatic Features & Uni-, Bi and Trigrams	0.41	0.36	0.36	0.35	0.41

predictions in a binary classification setting (Bhattasali et al., 2015; Kalouli et al., 2021). We model all possible context combinations of the question, preceding, and response material. The results are reported in Table 2. For this task, we employ both language models as well as RF and SVM models, for the latter unigrams are selected as input features the use of which results in better performance than bi-, trigrams, or any n-gram combination.

In this setting, BERT achieves the highest score overall with 0.48 for PREC-QU and QU-RESP context combinations confirming the positive effect of context on question type classification.[5] Notably, we see an even stronger impact of QU-RESP combination on LSTM performance with its score increasing from 0.37 to 0.42. However, with the statistical approach, the inclusion of context does not benefit either of the models.

From these results, we infer the following: (1) the use of context for multiclass question type prediction seems to be beneficial only in some settings; (2) the gap between the performance levels of an LLM and statistical models is not as large as it could be expected considering the disparity in the amounts of computational power required for using the latter. With this in mind, we explore in the next section the possibility of further improving RF and SVM results by incorporating pragmatic features.

5.2 Pragmatic and lexical features

The results of the statistical models using the pragmatic and lexical features are presented in Table 3. For this set of experiments, we use only question text for extracting relevant n-grams based on the results reported in Table 2. With this set of experiments, we find that the inclusion of pragmatic features in addition to n-grams improves the performances of both, RF and SVM, with the former achieving the highest balanced accuracy score for statistical models of 0.44. Overall, the RF model has comparable results for all feature sets with the presence of an answer seemingly having a slight edge in terms of its impact on the performance. However, the SVM tends to rely more heavily on the speaker roles as well as the combination of all of the pragmatic features. Finally, a relatively high score of 0.43 can be observed in the setting with the pragmatic features only. However, after further inspection, we find that models in this setting predict only two of the classes, indicating that lexical features carry valuable information that cannot be overlooked.

We also note that when adopting a multiclass approach to question typology, the performance of the models drops considerably when compared to previous research focused on two classes of questions at a time. Kalouli et al. (2021) achieve accuracy of about 0.88 when identifying information and non-information-seeking questions; Ranganath et al. (2016) and Oraby et al. (2017) distinguish between rhetorical and non-rhetorical questions with F1-scores of about 0.64 and 0.76 respectively. Our results suggest that an introduction of the third category of questions increases the task complexity.

5.3 Error analysis

In order to further investigate the results we conduct an error analysis by examining confusion matrices for the best-performing models in both settings (see Appendices A and B). Unsurprisingly, we find that the unbalanced nature of the dataset with a higher number of PQs compared to AQs and RQs results in models demonstrating better performance in the case of the over-represented category. With respect to the BERT models that

[5] We tested other models which all yielded inferior results, including RoBERTa, DeBERTaV3, and a zero-shot setting with Vicuna.

take into account context, there is also a relatively high number of AQs and RQs misclassified as PQs (about 40% for AQs and over 50% for RQs) and a considerably lower rate of misclassification between AQs and RQs (around 20%). As for the RF model that considers lexical features and answers, we observe that it is much better at identifying RQs than the BERT model (almost 40% of correct predictions compared to around 20%). In particular, it is more successful at distinguishing them from the PQs which can be attributed to the fact that RQs are less likely to be answered because of their nature. Overall, the results of the error analysis indicate that in the case of both approaches, the models exhibit better performance when it comes to the PQs while displaying varying degrees of difficulty with the other two question types.

6 Conclusion

In this paper, we introduce a task of question type prediction using a more fine-grained typology than the typically adopted bipartite distinction and find that the introduction of the third class increases the complexity of the task drastically. While questions can be tricky to interpret as speaker intention, which is notoriously hard to capture, is often one of the main indicators of the question type, the task complexity is further increased by a high class imbalance in naturally occurring data.

We tackle the task by adopting several approaches conventionally used for the binary question type prediction such as the use of lexical features and the incorporation of question-adjacent context as well as by using novel for this task pragmatic features including propositional relations and epistemic markers in responses to questions. We find that BERT exhibits the best performance with an RF model trained on a combination of pragmatic features and unigrams taking second place. However, neither of these results is truly satisfactory, suggesting that current machine learning approaches are not yet powerful enough to reason about the nature of a question if we adopt a finer granularity into three types.

Acknowledgements

The work reported on in this paper was partially funded by the VolkswagenStiftung under grant Az. 98544 'Deliberation Laboratory'.

References

Refany Anhar, Teguh Bharata Adji, and Noor Akhmad Setiawan. 2019. Question classification on question-answer system using bidirectional-lstm. In *2019 5th International Conference on Science and Technology (ICST)*, volume 1, pages 1–5. IEEE.

Sunyam Bagga, Andrew Piper, and Derek Ruths. 2021. "are you kidding me?": Detecting unpalatable questions on reddit. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2083–2099.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–749.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 185–196. IOS Press.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

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.

Alice F Freed. 1994. The form and function of questions in informal dyadic conversation. *Journal of pragmatics*, 21(6):621–644.

F Maxwell Harper, Daniel Moy, and Joseph A Konstan. 2009. Facts or friends? distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 759–768.

Annette Hautli-Janisz, Katarzyna Budzynska, Conor McKillop, Brian Plüss, Valentin Gold, and Chris Reed. 2022a. Questions in argumentative dialogue. *Journal of Pragmatics*, 188:56–79.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022b. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the 2008 Conference on empirical methods in natural language processing*, pages 927–936.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 633–644, Doha, Qatar. Association for Computational Linguistics.

Aikaterini-Lida Kalouli, Katharina Kaiser, Annette Hautli-Janisz, Georg A Kaiser, and Miriam Butt. 2018. A multingual approach to question classification. In *Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2715–2720.

Aikaterini-Lida Kalouli, Rebecca Kehlbeck, Rita Sevastjanova, Oliver Deussen, Daniel A Keim, and Miriam Butt. 2021. Is that really a question?: Going beyond factoid questions in nlp. In *14th International Conference on Computational Semantics: IWCS 2021*, pages 132–143.

Zlata Kikteva, Kamila Gorska, Wassiliki Siskou, Annette Hautli-Janisz, and Chris Reed. 2022. The keystone role played by questions in debate. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 54–63, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Babak Loni. 2011. A survey of state-of-the-art methods on question classification. *Delft University of Technology, Tech. Rep*, 55:57.

Donald Metzler and W Bruce Croft. 2005. Analysis of statistical question classification for fact-based questions. *Information Retrieval*, 8:481–504.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319, Saarbrücken, Germany. Association for Computational Linguistics.

Suhas Ranganath, Xia Hu, Jiliang Tang, Suhang Wang, and Huan Liu. 2016. Identifying rhetorical questions in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 667–670.

Joao Silva, Luísa Coheur, Ana Cristina Mendes, and Andreas Wichert. 2011. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35:137–154.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.

Harish Tayyar Madabushi and Mark Lee. 2016. High accuracy rule-based question classification using question syntax and semantics. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1220–1230, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Seyhmus Yilmaz and Sinan Toklu. 2020. A deep learning analysis on question classification task using word2vec representations. *Neural Computing and Applications*, 32(7):2909–2928.

Dell Zhang and Wee Sun Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 26–32.

A Confusion matrices for the best-performing models in the context setting

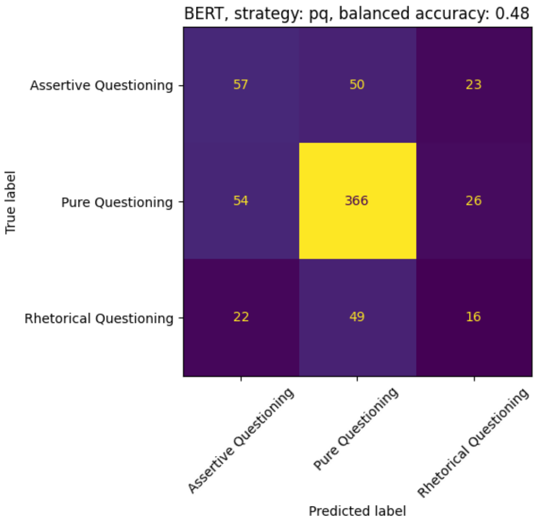

Figure 1: BERT in PREC-QU configuration

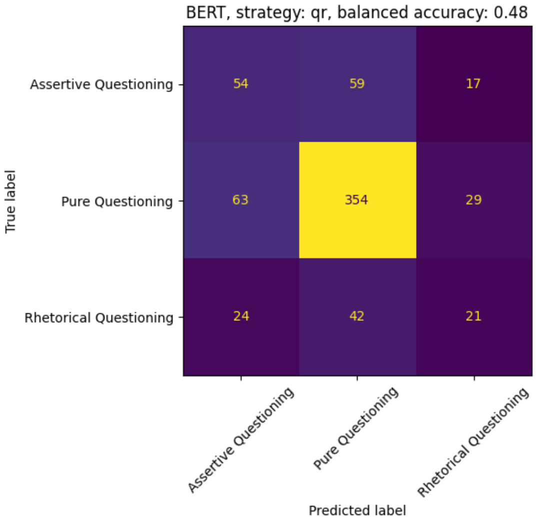

Figure 2: BERT in QU-RESP configuration

B Confusion matrix for the best-performing model in the pragmatic and lexical feature setting

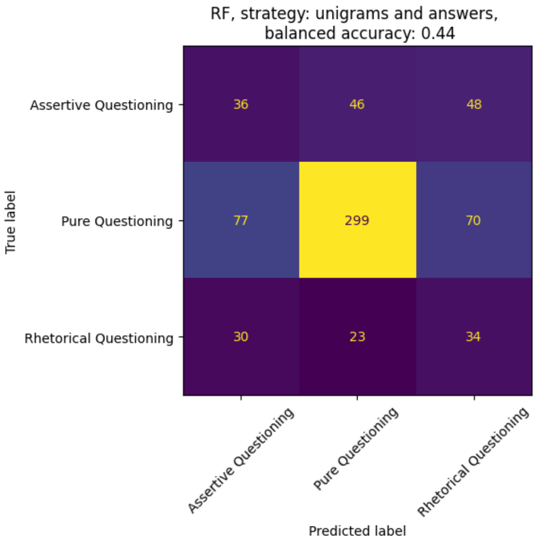

Figure 3: RF in Answers & Unigrams configuration