

Increasing Robustness for Cross-domain Dialogue Act Classification on Social Media Data

Marcus Vielsted

Nikolaj Wallenius
IT University of Copenhagen
robv@itu.dk

Rob van der Goot

Abstract

Automatically detecting the intent of an utterance is important for various downstream natural language processing tasks. This task is also called Dialogue Act Classification (DAC) and was primarily researched on spoken one-to-one conversations. The rise of social media has made this an interesting data source to explore within DAC, although it comes with some difficulties: non-standard form, variety of language types (across and within platforms), and quickly evolving norms. We therefore investigate the robustness of DAC on social media data in this paper. More concretely, we provide a benchmark that includes cross-domain data splits, as well as a variety of improvements on our transformer-based baseline. Our experiments show that lexical normalization is not beneficial in this setup, balancing the labels through resampling is beneficial in some cases, and incorporating context is crucial for this task and leads to the highest performance improvements (~7 F1 percentage points in-domain and ~20 cross-domain).¹

1 Introduction

The rise of social media and digital assistants has led to new forms of communication, where an enormous amount of data is available, and automatically understanding this has become an important quest. Automatically identifying the intent of an utterance is therefore highly relevant for automatically interacting with humans (i.e. chatbots), or to analyze people’s behaviour online. This task is also called Dialogue Act Classification (DAC). An example of two social media utterances annotated for DAC is shown in Table 1. DAC has traditionally mainly been investigated in the context of one-to-one spoken conversations, which is drastically different from one-to-many written conversations.

¹Code/data available on <https://github.com/marcusvielsted/DialogueActClassification>

Utterance	Label
“We are free tomorrow night, right?”	<i>propositional question</i>
“No, the final Grand Prix is on!”	<i>disagreement</i>

Table 1: Example utterances annotated for DAC

On top of this, language use on social media is evolving rapidly (Eisenstein, 2013). This makes the automatic processing of this data complex, and the standard setup in Natural Language Processing (NLP), where taking train and test data from the same domain is less relevant. New platforms are created while old ones are abandoned, and each platform comes with its own language norms and varieties. Hence we argue for a setup with two test sets, one in-domain and one cross-domain, and aim to improve the robustness of the current state-of-the-art models in NLP, transformers (Wolf et al., 2020; Devlin et al., 2019).

This leads to our research question: **How can dialogue act classification models be made more robust for in-domain and cross-domain applications on social media data?**, followed by our sub-questions:

- *SQ1: Can lexical normalization improve the robustness of a DAC Model?*
- *SQ2: Can resampling of label distributions improve the robustness of a DAC Model?*
- *SQ3: Can incorporating utterance context improve the robustness of a DAC Model?*

Contributions 1) we provide an annotation schema adapted from the ISO 24617-2:2020 standard, which we modify to better fit the task of annotating social media data 2) we provide DAC-annotated datasets for two domains, one large enough to train on, and one from another

domain/time-span to evaluate for robustness. 3) We evaluate and compare three methods to improve the robustness of DAC models: lexical normalization, label resampling, and exploiting context

2 Related Work

Social Media Despite the prevalence of social media in modern society, little research presently exists on the application of dialogue act classification on social media domains. The task has primarily been researched with a focus on verbal communication. Recently, some work has evaluated a CNN for Twitter data (Saha et al., 2019) and LSTMs on Reddit and Facebook data (Dutta et al., 2019). Unfortunately, these datasets are not publicly available.

Cross-domain In relation to the task of DAC, there has been limited research into model performance when predicting across unseen domains. Given the plurality of social media domains and their differences in communication structure, it is an opportune target for cross-domain classification. Dutta et al. (2019) evaluated cross-domain performance from Reddit to Facebook, reporting a drop of 5 absolute points F1 score, showing that the domains are relatively close. Additionally, cross-domain transfer learning between Human-Human and Human-Machine communication has been tested by Ahmadvand et al. (2019), who managed to outperform a state-of-the-art Hidden Markov Model through the use of transfer learning.

Context While some research into DAC has been applied to a single utterance in isolation of its context, dialogue acts are often context-dependent or context-sensitive (Bothe et al., 2018b). Although merely applying the preceding utterance provides performance improvements, Bothe et al. (2018a) demonstrate that using an utterance-level attention-based bidirectional recurrent neural network to analyze the importance of preceding utterances to classify the current one, provides additional performance. This is underlined by Raheja and Tetreault (2019), who use a conditional random field for sequence labeling of preceding utterances in combination with a self-attention recurrent neural network for text classification to achieve similar performance gains.

Transformer-Based Language Models As is the case for most NLP tasks, transformer-based language models finetuned on the target tasks

have recently been shown to outperform previous approaches. This was shown by Duran et al. (2021), who comparatively analyzed six different supervised learning models and ten pre-trained language models on DAC; the best performance was obtained by BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models.

3 Data

3.1 Datasets

Dialogue acts are commonly annotated on the smallest functional segment of a text that conveys an intended goal. Therefore, in this work, we annotate (multiple) utterances within a single social media post. For instance, while it is possible to encapsulate the following utterance within one single social media post: “Hi! How are you? Did you see the final Grand Prix last night?”, should not be interpreted as a single utterance, but rather split up and interpreted as three separate utterances, as it aims to convey three different dialogue acts. Additionally, for all input datasets, we assume that context is provided. Therefore, if the information is not present in a dataset, we enrich the dataset with columns containing these.

We will make use of two social media domains. For in-domain DAC, this project utilized the “NPS Chat Corpus” (Eric Forsyth et al., 2008) as source domain, consisting of 10,567 textual utterances collected through various chat forums in 2006 and thus presents a unique collection of early-day social media data. As social media domains can vary substantially in language and structure, we considered the NPS Chat Corpus to be an interesting source domain for cross-domain application, as we hypothesize that the similarities in utterance structure compared to a modern domain, such as Reddit, would be small. Therefore, we would be able to investigate and evaluate our models against two drastically different social media domains, and test the robustness of a given model.

For our cross-domain target, we compiled a Reddit dataset from the “Reddit Corpus (small)” dataset from “Convokit” (CornellNLP, 2021). Reddit is particularly interesting as subreddits potentially have variances in their use of language, vocabulary, and communication structure. Therefore, we are able to get a broader representation of the social media landscape compared to using other social media domains. By imposing our rules for selecting relevant utterances, see Appendix A, we ended

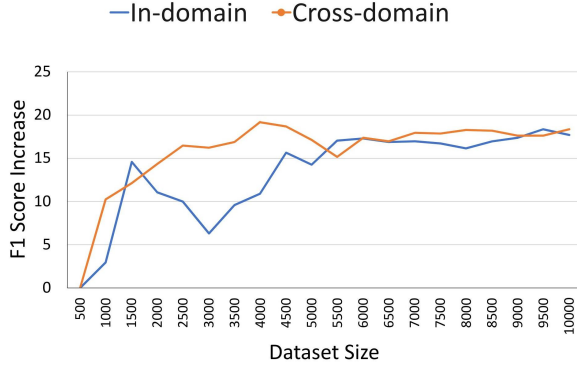


Figure 1: Learning curve of an SVM-based model on the NPS Chat corpus (taken from Wallenius and Vielsted (2021)).

up with a dataset consisting of 1,705 unique utterances from 50 unique threads within 37 different sub-reddits. Each thread is comprised of at least 20 consecutive utterances.

In previous research (Wallenius and Vielsted, 2021), it was shown that for a bag-of-words SVM model on the NPS-chat corpus, ~5,000 utterances is sufficient to train a competitive in-domain model (see Figure 1). Hence, we choose to annotate training data of a similar size (4,800 utterances).

Finally, in order to achieve a pure cross-domain setup for our model (i.e. no target domain data seen during development), we utilize a development split from the in-domain dataset for the tuning and model selection of all experiments. As such, the development set from the cross-domain dataset will only be used for the cross-domain model evaluations throughout the experiments. This is done in order to prevent overfitting of the cross-domain model improvement towards a single known cross-domain dataset. As such, this setup prevents over-estimation of performance for testing on additional domains not included in the experiments (Artetxe et al., 2020; Goot, 2021).

3.2 Guidelines

In relation to the task of DAC, various different annotation schemas have been developed, depending on the specific communication format for which a given task is being applied to. One of the first examples of DAC annotation schemas is the SWBD-DAMSL (Dan Jurafsky et al., 1997) intended for use on the Switchboard corpus (Potts, 1998). The Switchboard corpus consists of telephone conversations between two participants, i.e., a 1-1 bidirectional communicative relationship, which is reflected in the corresponding annotation schema.

Split	In-domain	Out-of-domain
train	4,800	—
dev	600	853
test	600	852

Table 2: Dataset splits for both in-domain and cross-domain.

Additional examples of annotation schemas include the MRDA tag set (Shriberg et al., 2004) created for dialogue annotation of large meetings, and the Posting Act Tagging schema (Wu et al., 2005) for early online chat forums. Though, in the examination of these annotation schemas, we deemed them unfit for this task as they emphasize traditional 1-1 bidirectional conversation formats. The Posting Act Tagging schema initially appeared to better utilized towards unidirectional 1-n social media data, as it was created to support tasks aimed towards earlier online chat forums.

However, having applied this tag-set in previous research, we found it inadequate for capturing many of the nuances of modern unidirectional social media posts and would skew the data towards one specific label. Therefore, the annotation schema utilized for this project has been adapted and modified from the ISO 24617-2:2020 dialogue act annotation standard (ISO 24617-2, 2020) in order to fit the specific task. This is done in an attempt to more accurately capture the nuances and account for the frequent use of unidirectional communication in modern social media data. The ISO 24617-2:2020 standard has specifically been made to address 3 shortcomings made from a previous version of the standard, as well as limitations from other annotation schemas. *“These experiences have brought to light (1) that the standard allowed dialogue act annotations that are slightly inaccurate in some respects, (2) that some applications would benefit from the availability of mechanisms for customizing the set of concepts defined in the standard, and (3) that certain use cases require the representation of functional dialogue act information to be extended with semantic content information.”*(ISO 24617-2, 2020). Thus, we have chosen to adapt and modify this standard, as the schema has been designed to account for these limitations by being domain-independent, and encouraging customization and extension as indicated in point (2). This allows us to create an annotation

Label	Example
propositionalQuestion	"r u serious?"
setQuestion	"what list should i put him in?"
choiceQuestion	"shaken or stirred?"
inform	"i wanna chat"
elaborate	"and dr phil said so."
continuer	"I know, but it threw me"
agreement	"i agree"
disagreement	"no, I didnt even look."
correction	"i meant to write the word may."
greeting	"hey ladies"
goodbye	"see u all laters"
positiveExpression	"yay!"
negativeExpression	"ewwww lol"
offer	"il get you a cheap flight to hell:)"
suggestion	"We should have a club"
instruct	"shut the fuck up."
acceptAction	"yeah i should toss it"
declineAction	"i don't wanna"
misc	:tongue:

Table 3: Tag-set adapted from ISO24617-2:2020 with examples from NPS Chat Corpus.

schema best suited for the task of labeling social media data. An overview of the resulting label set is shown in Table 3.

Schema Adaptations The primary adaptation involves splitting the generalized `Inform` label into 3 separate labels. The standard definition of the “Inform” label is “*Communicative function of a dialogue act performed by the sender, S, in order to make the information contained in the semantic content available to the addressee, A; S assumes that the information is correct*” (ISO 24617-2, 2020). In addition to this, we wanted to incorporate context into the annotation schema and distinguish between what dialogue act a given utterance is responding to. The two labels `elaborate` and `continuer` were therefore added. These categories are distinguishable from `inform` in the context of the dialogue act. Both labels imply additional information being added to a given subject, while referencing an object previously mentioned in a conversation. The distinction between the two labels is that `elaborate` implies that a sender is elaborating upon their own previous utterance, whereas, `continuer` implies that a sender is continuing a previous utterance by a different sender. By splitting the `inform` label into 3 we are thus isolating `inform` for instances where the utter-

ance can be read and fully understood without any context, which is a common occurrence in social media data where utterances are often unidirectional. Utterances labeled `inform` will therefore never reference named entities from previous utterances.

Other adaptations to the standard involves generalizing and unifying specific labels to narrow down the total number of labels. This was done in order to ensure that all labels were represented in a dataset. For this purpose, labels encompassed by the “Action-discussion functions” category in the ISO 24617-2 standard (ISO 24617-2, 2020) were reduced to `offer`, `suggestion`, `instruct`, `acceptAction` and `declineAction`. All `accept` and `decline` “Action-discussions functions” labels in the ISO 24617-2 were combined into the two labels, `acceptAction` and `declineAction`. The labels `answer`, `confirm` and `disconfirm` were removed as their functions could be incorporated into `continuer`, `agreement` and `disagreement`. Lastly, “Social-Expression” was incorporated through the labels, `greeting`, `goodbye`, `negativeExpression` and `positiveExpression`, as it is a significant part of communication on social media.

3.3 Annotation

The NPS dataset was annotated with the new tag-set by two annotators. Across 10 iterations with 50 utterances each, they consistently reached a Cohen’s κ score of 0.83, which can be interpreted as an “almost perfect” agreement (Cohen, 1960). Given this trend and a stagnation in improvement, the remaining utterances were then annotated individually. The dataset statement (Bender and Friedman, 2018) can be found in Appendix B.

4 Models

4.1 Baseline Model

As this research explores methods to improve DAC performance and robustness for cross-domain social media data rather than reaching optimal scores for one specific domain, hyperparameters were not continuously optimized throughout the experiments. The hyperparameter setup for this project was therefore to establish an optimized baseline model and to freeze the hyperparameters in this configuration throughout the experiments. The method for obtaining the optimized hyperparam-

Hyperparameter	Value	Range
Batch Size	16	[16, 64]
Warmup Steps	125	[75, 125, 250]
Learning Rate	7e-5	[5e-5, 7e-5, 9e-5]
Weight Decay	0.5	[0.1, 0.5]

Table 4: Our hyperparameters test ranges and chosen values.

ters was a three-step process. Firstly, we used the BERT-base model (Devlin et al., 2019) to find the optimal set of hyperparameters fine-tuned for this specific task. The hyperparameters optimized in this project can be seen in Table 4.

Secondly, having established the optimal hyperparameter values for the BERT-base model, we tested a total of 17 different transformer models (see Appendix C for the full list), to determine the best performing model(s). Lastly, the five best-performing transformer models from the previous test were then re-tested with the same range of hyperparameters as step one, to achieve a single optimal baseline model. By doing this, we could limit the scope of our model fine-tuning to 180 models instead of 612 models. Using this setup, the following five transformer models produced the best results: “*deberta v3 base*”, “*deberta v3 large*” (He et al., 2021), “*bertweet-base*” (Quoc Nguyen et al., 2020), “*bertweet-large*” (Quoc Nguyen et al., 2020), and “*bert-base-uncased*” (Devlin et al., 2019). Doing hyperparameter optimization for the five models, we found *deberta-v3-large* to provide the best performance. However, we selected **deberta-v3-base** as the model for our experiments. This selection was made due to computational restrictions, limiting the number of large models that we would be able to fine-tune and test. To support this selection, the large version was shown to be only slightly better, with an F1 score of 0.325 percentage-points higher than the base model, which scored **77.11** F1 in-domain and **53.92** F1 cross-domain averaged over five seeds.

4.2 Lexical Normalization

As a result of the informal nature of social media data, utterances often include abbreviations, slang, and misspellings. These language variations already constitute a significant challenge for traditional NLP models trained on chronological text (Baldwin et al., 2013; Eisenstein, 2013). Moreover, language variations potentially pose an even greater

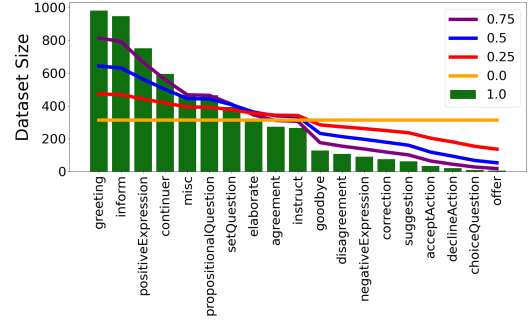


Figure 2: Label resampling of NPS Chat Corpus

challenge for cross-domain application, since it involves two different domains, likely resulting in more Out-Of-Vocabulary (OOV) tokens. Therefore, we hypothesized that applying lexical normalization on both our source and target domains could unify their vocabularies, thus increasing the token overlap and improving our model performance. An example of manually annotated normalization is:

Original: “any ladis wanna chat?”

LexNorm: “any ladies want to chat?”

For this task, we used the lexical normalization tool **MoNoise** (Van Der Goot, 2019), which produces performance on par with the state-of-the-art for English data (van der Goot et al., 2021). We use the publicly available MoNoise model for English, trained on data from Li and Liu (2014), to create parallel datasets for each domain with normalized text. This allows us to continuously test the results of lexical normalization, both in isolation and in combination with methods described in Section 4.

4.3 Multinomial Resampling

For cross-domain applications, we assume that label distributions within the source domain and target domain differ. In order to negate a potential labelling bias towards specific classes, we hypothesized that having more balanced and aligned label distributions between the datasets would improve model robustness. For this purpose, we resampled our datasets with respect to the annotated labels and according to a multinomial distribution. Using a multinomial resampling algorithm, each label is resampled according to the probability of its occurrence in our dataset:

$$\frac{1}{p_i} * \frac{p_i^\alpha}{\sum_i p_i^\alpha}$$

Utterance: I have never used elbow	
Context	Utt. Label
You only scored the goal because you used your elbow	disagreement
Did you use your hands or elbows to get up there?	inform

Table 5: Impact of differing contexts on dialogue acts. Note that the label column indicates the label of the original utterance

where p_i represents the probability that a random sample corresponds to the label i and α is a hyperparameter, for our sample smoothing function, to determine the proportional degree to resample. $\alpha = 1.0$ corresponds to the pre-existing distribution, and $\alpha = 0.0$ corresponds to an equal distribution for all labels. The effect of α on the training data distribution is visualized in Figure 2.

4.4 Utterance Context

Motivated by the definition of the task and previous work (Section 2), we hypothesized that integrating context into the model might increase robustness. Table 5 exemplifies this: if isolated from context, it is unclear which label is correct for the utterance. We found three different context elements to be relevant. **context_Label**: the given classification label for the utterance to which a given utterance responds. **context_Text**: the actual utterance text a given utterance is responding to. **context_Sender**: a binary value specifying whether the sender of the context utterance is the same as the sender for a given utterance. This would be the case if a participant responds to their own prior utterance.

These three context elements are concatenated to the text and are separated with the [SEP] token. A total of 15 permutations of the three elements were then combined either pre or post the input text, thus resulting in 30 different context configurations. An additional permutation was added for no context elements (overview in Appendix D).

Gold vs. Predicted Context Label Out of our three context elements described above, context_Text & context_Sender can be easily generated using the utterance_ID. context_Label, however, is not that easily generated: it requires either manual annotation or an iteration of model predictions for the whole dataset. Therefore, our dataset has two different context_Label columns, *Gold* and *Predicted*. The *Gold* labels are used as a baseline to

	In-domain	Cross-domain
Base	77.11 \pm 1.85	53.92 \pm 1.06
+Norm	76.81 \pm 0.95	54.02 \pm 0.72

Table 6: Average results of lexical normalization in isolation in-domain & cross-domain (dev).

compare the performance for the *Predicted* labels, and are therefore only for analytical importance. The *Gold* labels were manually annotated, when the dataset was annotated. We obtained the *Predicted* labels through a five-fold cross-validation setup on our training data, where we trained on 80% and predicted a label on the remaining 20%. For each fold, we instantiated a new optimized baseline model, see section 4.1, so as to avoid overfitting.

5 Results

All results reported are average macro-F1 scores over 5 random seeds unless mentioned otherwise. As mentioned in Section 3, we always used the in-domain dev-set for model picking, as well as for hyperparameter tuning. We first evaluate each of our proposed improvements (Section 5.1-Section 5.3). Then, we attempt to combine our methods (Section 5.4), and confirm our findings on the test data (Section 5.5). We use Almost Stochastic Order (ASO) for significance testing (Dror et al., 2019) as implemented by Ulmer et al. (2022) over the random seeds, and with an epsilon (ϵ) smaller than 0.5 we reject the null hypothesis.

5.1 Lexical Normalization

As shown in Table 6, we observed a performance decrease in F1 score of .3 percentage points in-domain and a negligible gain of 0.1 percentage points for cross-domain when normalizing the train and dev data. Based on these scores, we conclude that lexical normalization is not beneficial for DAC in our setup. Because the in-domain results show an opposite trend as we hypothesized, namely that normalization is not useful, we do an ASO significance test to confirm whether using the original data is stochastically dominant over using the normalized data. This test resulted in a minimum epsilon of 0.0, and we can thus confirm that normalization leads to lower scores, whereas the cross-domain differences were shown not to be significant.

	In-domain	Cross-domain
Base	77.11 \pm 1.85	53.92 \pm 1.06
Resample	78.50 \pm 1.80	55.54 \pm 1.31

Table 7: Scores on the in-domain and cross-domain data when using resampling compared to our baseline (dev). Resampling α is 0.9 and 0.8 respectively.

Setup	#cont.	F1
Baseline In-domain	—	77.11
Gold In-domain	27	85.39
Pred In-domain	27	86.22
Baseline Cross-domain	—	53.92
Gold Cross-domain	14	76.35
Pred Cross-domain	22	76.35

Table 8: Best performing context configuration for all four setups. **The 3 context configs (#cont.) used are:** 27: [Context Sender + Text + Label] Post Input Text 22: [Context Label + Sender + Text] Post Input Text 14: [Context Sender + Label + Text] Pre Input Text.

5.2 Multinomial Resampling

For multinomial resampling, the best settings we found where $\alpha = 0.9$ for in-domain and $\alpha = 0.8$ for cross-domain (Section 6.3). Results show a substantial improvement, while keeping standard deviation in a similar range (Table 7). The resampling ratios (α) tested were 0.60, 0.80 and 0.90 for in-domain and 0.80, 0.85, 0.90 and 0.95 for cross-domain. These were selected, as they provided the best results for each domain. Significance tests resulted in a minimum epsilon value of 0.0 for in-domain and 0.02 for cross-domain compared to disabling the resampling ($\alpha = 1.0$), confirming that multinomial resampling is stochastically dominant.

5.3 Context

In Table 8, we report the results for the best permutation of all different elements of context we consider (Section 4.4). Full results can be found in Appendix D. Perhaps surprisingly, the predicted labels perform on par with the gold labels. We hypothesize that a partial explanation for this, is that our model performs very well on the informative labels our models require to learn context. i.e. labels `inform`, `instruct`, `offer`, `suggestion`, `setQuestion`, `propositionalQuestion` (see appendix E). We confirmed with an ASO

Feature	In-domain	Cross-domain
Context Conf.	[19, 23, 26, 27, 31]	[11, 19, 22, 23, 31]
Context label	[Gold, Predicted]	[Gold, Predicted]
Resampling α	[.95, .9, .75, .65, .6]	[.95, .9, .85, .7, .4]
Normalization	[+,-]	[+,-]

Table 9: Feature setup for combined models. The exact context configurations can be found in Appendix D, but all of the ones in this table use all three context elements.

In-domain			Cross-domain		
#cont.	α	F1	#cont.	α	F1
27	1.0	86.22	22	0.95	77.17
26	1.0	85.51	11	0.90	76.58
31	1.0	85.48	31	0.70	76.58
23	1.0	85.47	31	0.95	76.48
19	1.0	85.28	31	0.40	76.43

Table 10: The feature values for our best performing models for both our In-domain(ID) and for our Cross-domain(CD). All 10 models are using *predicted context labels* and the *non-normalized* dataset. #cont. refers to the context configuration.

significance test that gold labels are not out-performing predicted labels with an epsilon of 1.0.

5.4 Combining

We evaluated all combinations for the (max.) five setups for each of our robustness proposals, which are summarized in Table 9. Our best performing model reached a performance of 82.09 (in-domain) with the following setup from Table 9: [19, Predicted, 0.95, Original]. This score is lower than our previous highest score, achieved when only using context, see Table 8. On average, the top five feature setups combining resampling and context tested 2.4 percentage points below the same feature setup without resampling. This reduction in score when combining context and resampling could potentially be explained by the two features achieving improvements in similar situations, and are thus not complementary. For the cross-domain experiments, the label resampling still contributes, as the best five combined models (shown in Table 10) outperform the 76.35 reported in Table 8.

5.5 Test Data

On the test data (Table 11), we see that the model slightly overfits on the in-domain dev data (from the lower scores on test), but this does not transfer

Model	In-domain		Cross-domain	
	Dev	Test	Dev	Test
Baseline	77.11	76.27	53.92	55.17
LexNorm	76.81	74.99	54.02	53.35
Resample	78.17	79.09	54.91	55.67
Context	84.42	83.83	75.01	75.18
Best	84.42	83.83	75.70	75.64

Table 11: Development and Test scores for In-Domain & Cross-Domain for selected models.

to the cross-domain setup (where $\text{test} > \text{dev}$). Furthermore, the results align with our observations on the test data: normalization is not useful, resampling benefits performance to some extent, and the context is most crucial for performance. Furthermore, combining context and resampling does not lead to improved performance.

6 Analysis

As our standard deviations were relatively small and to simplify our analyses (and for computational efficiency), all results in this section are obtained over one seed.

6.1 Baseline model

In Appendix E, Figure 6 and 7 we show confusion matrices for both domain’s baseline models predictions. As can be seen from these matrices, our baseline models have mostly certain classes with reoccurring mislabeling. These classes with reoccurring mislabeling cover the labels where context is a distinguishing factor, and where the variations of the input text between the labels is often minor. For this reason, it was expected for the baseline to struggle to distinguish between the labels `inform`, `continuer` & `elaborate`, without the addition of context.

The classes with reoccurring mislabeling are most evident for the cross-domain baseline model. We hypothesize that one reason for this could be that the utterance variations within each label between our in-domain train-set and dev-set is smaller compared to the difference between our in-domain train-set and our cross-domain dev-set. Therefore, our in-domain model would more likely have been trained on similar label tendencies as the one present in the in-domain dev-set, compared to our cross-domain model. This argument underlines that social media domains are constantly chang-

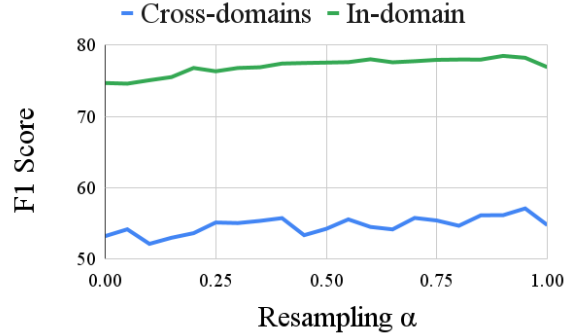


Figure 3: Results of Resampling in isolation in-domain & cross-domain.

ing, and confirms the importance of cross-domain evaluation.

6.2 Lexical Normalization

The, perhaps surprisingly, negative results obtained when using lexical normalization can be explained by a variety of reasons: 1) performance of the normalization model, as it is used out-of-domain (it is trained on Twitter), performance might be sub-optimal. Upon inspection, we found that the model is too conservative, and many non-standard words are not normalized. 2) removal of information, by normalizing we are essentially removing information, for example: “YOU DID” could be interpreted as a `propositionalQuestion` whereas writing “you did” is more likely to be interpreted as a `continuer`. 3) perhaps word overlap has become less important since modern language models use sub-words.

6.3 Resampling

As the resampling α determines the degree to which the label distribution is normalized, we have tested the full range of resampling ratio (from $\alpha = 0.0$ to $\alpha = 1.0$) in increments of 0.05. From these results, as shown in Figure 3, we were able to identify the trend that a lower resampling ratio (i.e., higher α) provides the biggest performance increase for both in-domain and cross-domain. Focusing on the in-domain line, we see that lower rates consistently perform worse, which can be explained by the fact that the label distribution of the in-domain dev data is similar to the train data, and changing this makes the distribution more distant. On the dev data, there is a slightly increasing trend up to $\alpha = 0.90$. The difference between 1.0 is larger compared to the in-domain line. There is a drop > 0.90 on both dev-sets. As shown in

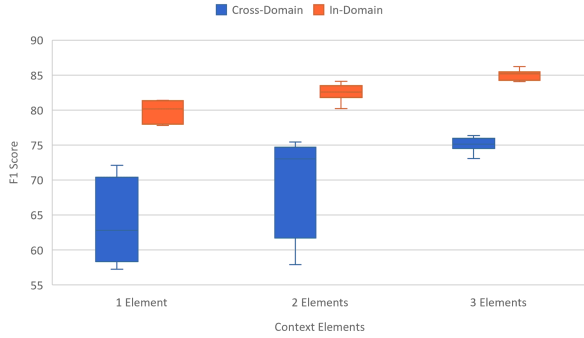


Figure 4: Effect of using different amount of context elements on F1 score.

Figure 2, our dataset has an unequal distribution in label classes, where labels such as `offer` and `choiceQuestion` are much less occurring than `greeting` and `inform`. We hypothesize that with such a label distribution as our NPS dataset, our model does not have enough training data for these rare classes, thus not being able to accurately predict on these. By applying a resampling ratio where one achieves a balance between aligning the training data with the dev-data and increasing the occurrences of rare labels, we get the most out of multinomial resampling.

6.4 Context

The results of all our permutations of using the context are plotted in boxplots in Figure 4. We summarize the results over the number of added elements, and plot quartiles. The trend is clear: more context information leads to higher performance, and in the case of the cross-domain results, all elements are necessary to obtain stable results. When inspecting the individual scores (Appendix D), we can conclude that the text of the previous utterance is the most important context feature.

7 Conclusion

Firstly, we found that lexical normalization does not constitute a stochastically dominant feature for cross-domain applications, but rather had a negative effect on F1 performance. By applying lexical normalization, the performance dropped for both our in-domain data, while staying the same for our cross-domain dataset. Additionally, while it decreased the standard deviation for our in-domain model, it almost doubled the standard deviations for our cross-domain model. We can therefore state that lexical normalization does not improve the robustness nor increase the F1 score of either in-

domain or cross-domain DAC model when trained on social media data.

Secondly, we have seen that multinomial resampling is a stochastically dominant feature in isolation with regard to increasing the F1 score for both in-domain and cross-domain. However, for cross-domain applications it increased the standard deviation drastically, while maintaining the same standard deviation for in-domain usage. Used in combination with context for cross-domain applications, we were able to both increase the F1 score marginally, while reducing the standard deviation from 0.92% to 0.42%. We can therefore conclude that multinomial resampling can increase the performance and robustness of a DAC model for cross-domain applications when combined with context as a feature, but should not be included for in-domain models.

Thirdly, we have observed that context has been the most significant contributing factor to the large performance increase of both in-domain and cross-domain models. We have shown that additional context elements in our setup increase robustness and constitutes a stochastically dominant feature compared to fewer context elements. Improving the F1 scores by 7.28 percentage-points for in-domain and 21.23 percentage-points for cross-domain, while also reducing the standard deviation to 0.65% for in-domain and 0.92% cross-domain, we have proved that context can improve the robustness of DAC models for cross-domain as well as in-domain applications.

Acknowledgements

We would like to thank Mike Zhang and the anonymous reviewers for their comments on earlier versions of this paper. We also acknowledge the IT University of Copenhagen HPC for the resources made available for conducting the research reported in this paper.

References

- Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. 2019. [Contextual dialogue act classification for open-domain conversational agents](#). *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1273–1276.
- Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. [A call for more](#)

- rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.
- Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. [How Noisy Social Media Text, How Different Social Media Sources?](#)
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter. 2018a. [Conversational Analysis using Utterance-level Attention-based Bidirectional Recurrent Neural Networks](#). *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018-September:996–1000.
- Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018b. [A context-based approach for dialogue act recognition using simple recurrent neural networks](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jacob Cohen. 1960. [A Coefficient of Agreement for Nominal Scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- CornellNLP. 2021. [ConvoKit version 2.5.1](#).
- Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. [Switchboard SWBD-DAMSL Shallow-DiscourseFunction Annotation Coders Manual](#). *SRI International*.
- Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.
- Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. [Deep Dominance - How to Properly Compare Deep Neural Models](#). *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2773–2785.
- Nathan Duran, Steve Battle, Jim Smith, and Nathan Duran. 2021. [Sentence encoding for Dialogue Act classification](#). *Natural Language Engineering*, pages 1–30.
- Subhabrata Dutta, Tanmoy Chakraborty, and Dipankar Das. 2019. [How Did the Discussion Go: Discourse Act Classification in Social Media Conversations](#). In *Linking and Mining Heterogeneous and Multi-view Data*, chapter 6, pages 137–160. Springer International Publishing.
- Jacob Eisenstein. 2013. [What to do about bad language on the internet](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369. Association for Computational Linguistics.
- Eric Forsyth, Jane Lin, and Craig Martell. 2008. [The NPS Chat Corpus](#).
- Rob van der Goot. 2021. [We Need to Talk About train-dev-test Splits](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing](#).
- ISO 24617-2. 2020. ISO 24617-2 - Language resource management - Semantic annotation framework - Part 2: Dialogue acts. *ISO*, pages 1–96.
- Chen Li and Yang Liu. 2014. [Improving text normalization via unsupervised model and discriminative reranking](#). In *Proceedings of the ACL 2014 Student Research Workshop*, pages 86–93, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Christopher Potts. 1998. [The Switchboard Dialog Act Corpus](#).
- Dat Quoc Nguyen, Thanh Vu, Anh Tuan Nguyen, and VinAI Research. 2020. [BERTweet: A pre-trained language model for English Tweets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. Association for Computational Linguistics.
- Vipul Raheja and Joel Tetreault. 2019. [Dialogue Act Classification with Context-Aware Self-Attention](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, volume 1, pages 3727–3733. Association for Computational Linguistics.
- Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2019. [Tweet Act Classification : A Deep Learning](#)

based Classifier for Recognizing Speech Acts in Twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, volume 2019-July. Institute of Electrical and Electronics Engineers Inc.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. [The ICSI Meeting Recorder Dialog Act \(MRDA\) Corpus](#). In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100. OSD or Non-Service DoD Agency, Association for Computational Linguistics.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

Rob Van Der Goot. 2019. [MoNoise: A Multi-lingual and Easy-to-use Lexical Normalization Tool](#). *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pages 201–206.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. [MultiLexNorm: A shared task on multilingual lexical normalization](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Nikolaj Wallenius and Marcus Lind Vielsted. 2021. Dialogue act classification for social media data.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler, and William M. Pottenger. 2005. [Posting Act Tagging Using Transformation-Based Learning](#). In *Foundations of Data Mining and knowledge Discovery*, pages 319–331. Springer, Berlin, Heidelberg.

Appendix

A Reddit Dataset Creation

“Reddit Corpus (small)” is composed of 297,132 utterances (Posts) from 8,286 conversations (threads) within 100 unique Subreddits. From this starting point, we identified all threads with more than 40 utterances. This limit was set in order to get longer threads with potentially more conversations containing bidirectional communication, which would enable us to better investigate context among utterances. Despite this project splitting each social media post up into multiple different utterances based on its shortest possible functional segments, we still wanted posts with a shorter length, as we hypothesized this would entail a more accurate real world version of singular communicative functions. Furthermore, we hypothesized that relying on shorter posts would balance out the label distribution, as this implies a greater number of different speakers. Therefore, we chose to include all threads where the mean length of posts were between 50 and 150 characters.

Having compiled a list of available threads abiding by the aforementioned rules, we selected 50 random threads, and included the first 20 consecutive utterances from these. This resulted in a total of 1000 posts which, after splitting each post up into its smallest possible communicative function, returns a dataset consisting of 1705 unique utterances from 50 Subreddits.

B Dataset Statement

Following (Bender and Friedman, 2018), the following outlines the data statement:

A. CURATION RATIONALE Collection of text samples from different social media domains. The first part (NPS Chat corpus) was sampled from a variety of platforms in 2006, and the collection of the Reddit samples is a random sample of long threads taken from 100 different Subreddits (more detail in Appendix A)

B. LANGUAGE VARIETY Most of the data is filtered to be English, it is unknown which variety of English is dominant.

C. SPEAKER DEMOGRAPHIC Unknown.

D. ANNOTATOR DEMOGRAPHIC Two software-design master students, both have previous experience with annotating for dialogue act classification. Native language: Danish, but proficient in English.

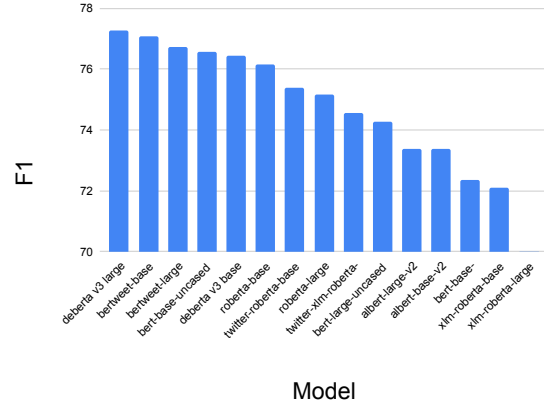


Figure 5: List of all the models used for testing within our project

E. SPEECH SITUATION Online, most probably the old data is typed from a keyboard, whereas the Reddit part of the data might come from a larger variety of devices.

F. TEXT CHARACTERISTICS There could be a variety of noise in the utterances, as well as utterances that contain mainly canonical text. No filtering on style was done.

C Model Selections

Figure 5 show the performance of all of the 17 language models evaluated on the in-domain dev data.

D Context Configurations

Table 12 shows the exact order of each of our context configurations, and their corresponding scores when using gold or predicted context labels for both in-domain and cross-domain. Note that these results are not averaged over 5 random seeds as the results in the main paper, thus the context configurations that do not use context labels still differ between Gold and Pred.

E Confusion matrices

Figure 6 and Figure 7 show the confusion matrices of our baseline model on the In-domain and Cross-domain dev sets.

Id	Context config	Position	# elems.	In-domain		Cross-domain	
				Gold	Pred	Gold	Pred
1	No Context Included		0	76.32	81.41	54.21	52.69
2	Only Context Label	Pre	1	78.02	76.77	58.35	63.54
3	Only Context Sender	Pre	1	79.27	80.84	70.39	69.80
4	Only Context Text	Pre	1	80.72	77.79	60.47	58.72
5	Only Context Label	Post	1	78.76	81.35	60.24	62.12
6	Only Context Sender	Post	1	79.90	79.51	71.37	72.11
7	Only Context Text	Post	1	79.39	78.05	61.76	57.23
8	Context [Label + Sender]	Pre	2	82.50	83.53	72.82	74.01
9	Context [Label + Text]	Pre	2	81.07	82.55	61.11	61.28
10	Context [Label + Sender + Text]	Pre	3	82.93	84.17	74.49	74.41
11	Context [Label + Text + Sender]	Pre	3	83.71	85.28	73.97	75.77
12	Context [Sender + Label]	Pre	2	81.38	83.35	73.17	74.98
13	Context [Sender + Text]	Pre	2	83.73	80.23	75.19	74.71
14	Context [Sender + Label + Text]	Pre	3	83.68	84.09	76.35	74.84
15	Context [Sender + Text + Label]	Pre	3	84.21	84.42	73.77	75.23
16	Context [Text + Sender]	Pre	2	83.86	81.32	75.89	70.99
17	Context [Text + Label]	Pre	2	81.37	82.64	60.97	62.90
18	Context [Text + Sender + Label]	Pre	3	83.19	85.17	75.74	73.07
19	Context [Text + Label + Sender]	Pre	3	83.96	85.28	75.90	76.28
20	Context [Label + Sender]	Post	2	82.54	83.63	71.22	74.33
21	Context [Label + Text]	Post	2	79.83	82.14	59.39	61.33
22	Context [Label + Sender + Text]	Post	3	83.43	84.12	75.07	76.35
23	Context [Label + Text + Sender]	Post	3	84.41	85.47	76.14	75.72
24	Context [Sender + Label]	Post	2	82.05	84.10	75.58	75.45
25	Context [Sender + Text]	Post	2	83.18	81.70	76.02	74.68
26	Context [Sender + Label + Text]	Post	3	83.42	85.51	74.97	74.15
27	Context [Sender + Text + Label]	Post	3	85.40	86.22	73.64	74.99
28	Context [Text + Sender]	Post	2	83.13	82.71	74.59	72.03
29	Context [Text + Label]	Post	2	78.64	82.27	59.28	57.91
30	Context [Text + Sender + Label]	Post	3 3	82.76	84.92	74.08	75.00
31	Context [Text + Label + Sender]	Post	3	83.97	85.48	75.20	76.02

Table 12: Results for all our context configurations (Dev). Position: pre means before the input text, post behind the input text.

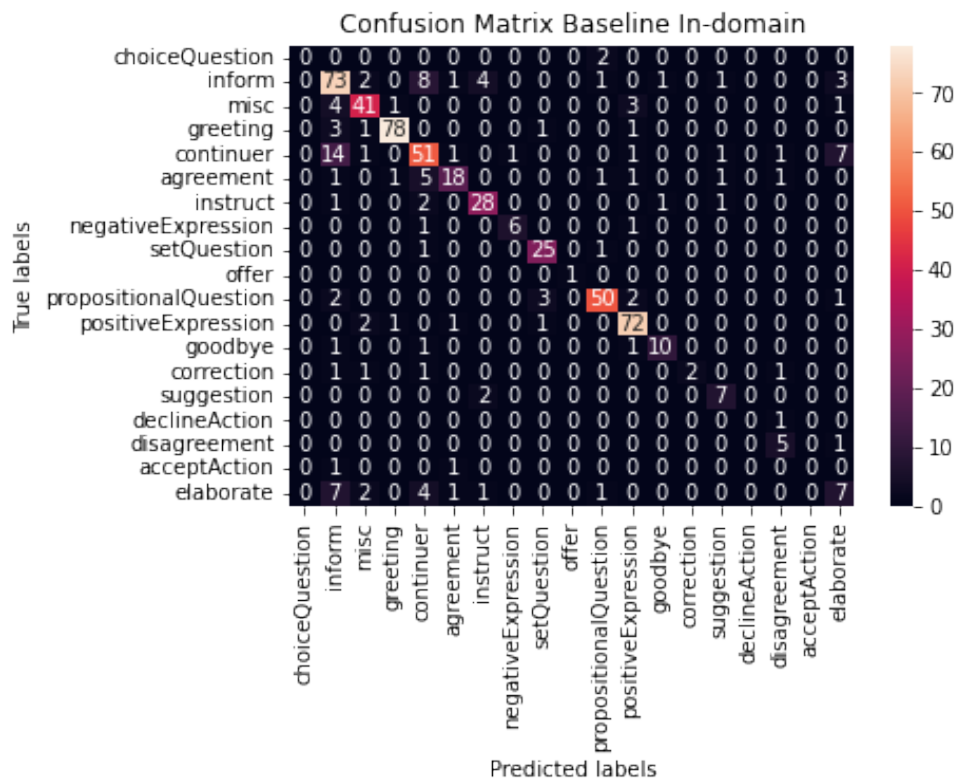


Figure 6: Confusion matrix for our baseline model for In-domain

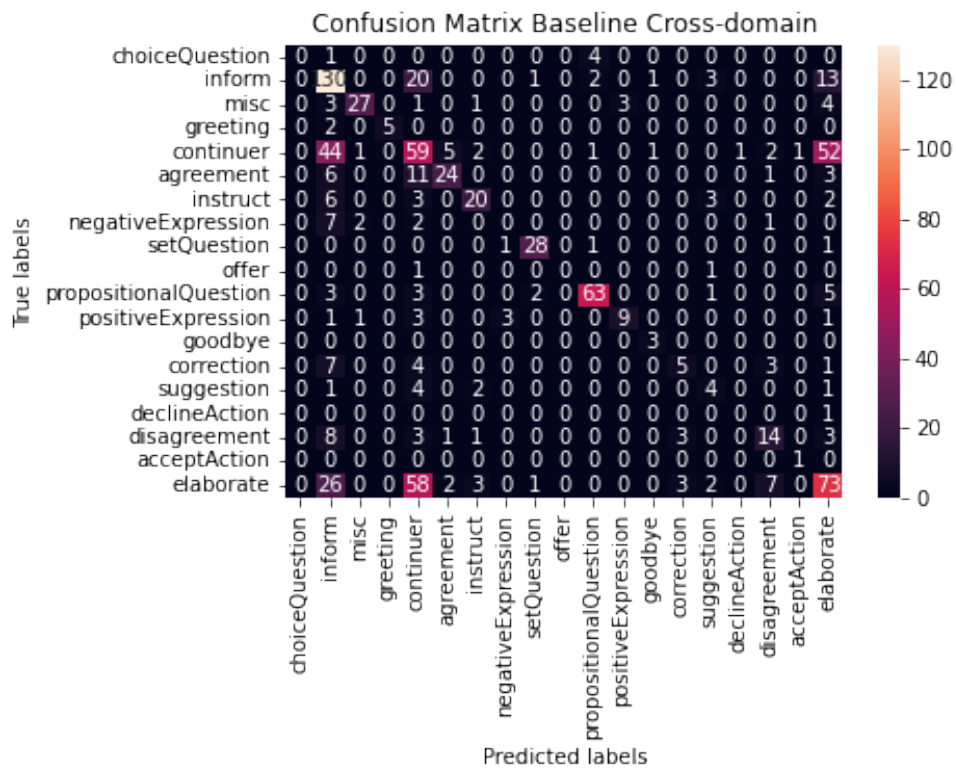


Figure 7: Confusion matrix for our baseline model for Cross-domain