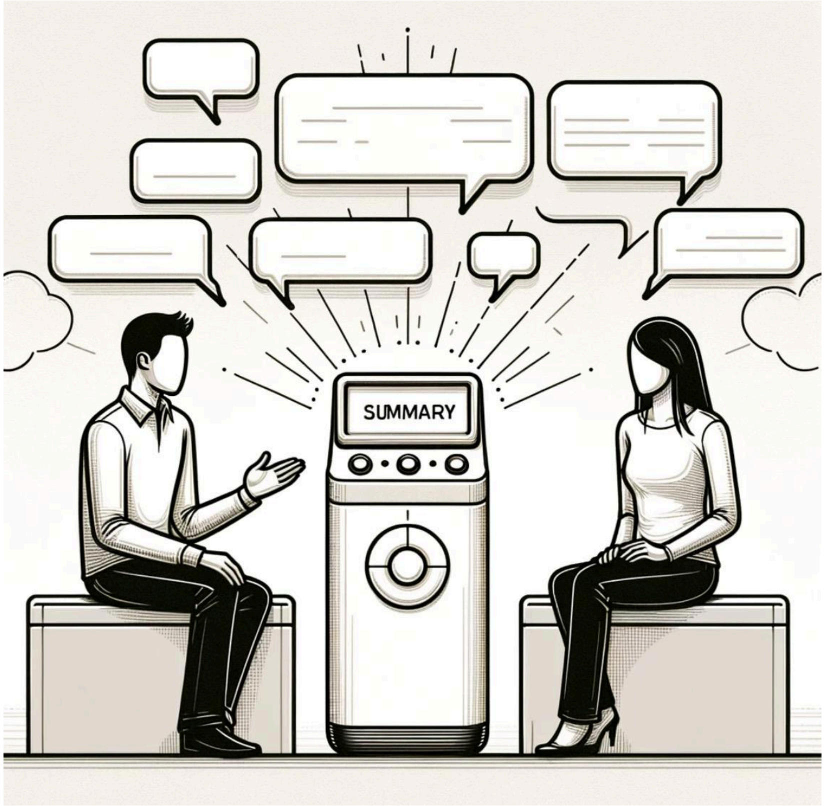
Conversational Summarization Using BART



INDENG 242B - Machine Learning and Data Analytics II
Final Project Report

Chloe Kang ID: 3039712860
Tunan Li ID:3039669622
Kairan Wang ID: 3034574973
Roxanne Wang ID: 3039679645
Iris Zheng ID: 3039679164

Introduction

In the realm of natural language processing, the quest for efficient and accurate text summarization has been a longstanding challenge. Leveraging cutting-edge technologies, our project delves into the domain of sentence summarization using the BART (Bidirectional and Auto-Regressive Transformer) architecture. This architecture, known for its prowess in natural language understanding and generation tasks, serves as the cornerstone of our endeavor.

Motivation

We fine-tune BART models facebook/bart-large-xsum and facebook/bart-large-cnn to distill their verbose text into short summaries. These models are all fine-tuned with a specific dataset to gain the power of summarization. We ensure that no stone is left unturned, from meticulous data preparation, custom training configurations, and rigorous evaluation, both using ROUGE scores. We ensure that the summaries not only give accuracy but also push the text summarization technology in its capabilities to make the results very insightful.

Data Collection & EDA

The Samsum-train.csv dataset contains 14,732 Training, 818 Validation, and 819 Test conversations with their respective IDs, dialogues, and summaries. There is only one missing value and it was handled and removed by simply ensuring that the data had conversational elements. In the EDA, different lengths of the dialogue were used (from 31 to 5492 characters) and summary (from 3 to 300 characters) to show different vocabulary patterns and distinct information from dialogue to summary. The dialogues are presented with key points and summaries in a structured manner with the common elements of dialogue. For the training set, these presented many patterns of words, which will be informative for the development of natural language processing-based models. Similar analysis and visualization were performed for the validation and test sets with consistency in data handling and observations across all the three subsets.

Modeling

Our project employs the BART (Bidirectional and Auto-Regressive Transformer) architecture for the task of sentence summarization, using two different pretrained models: \`facebook/bart-large-xsum\` and \`facebook/bart-large-cnn\`. The former is a pre-trained BART model developed by Facebook AI Research fine-tuned on the XSum dataset, which contains single-sentence news articles and their corresponding summaries. The latter was trained on a dataset including articles from the CNN news website and is suitable for various text generation tasks. These models have been selected for their robustness and effectiveness in natural language understanding and generation tasks.

Model Preparation

The datasets used include dialogues and respective summaries. We have divided our data into training, testing, and validation sets with 14731, 819, and 818 entries respectively. The datasets were converted from pandas dataframes into Hugging Face's `Dataset` objects to streamline the training and evaluation process.

For both models, we implemented a `preprocess_function` specifically tailored for preparing datasets to be preprocessed by BART. It handles the tokenization of both input and target variables, which is the first step in converting raw text into a form that the model can understand. For the inputs, which are the dialogues in our case, the function extracts each dialogue and encodes it into tokens using the tokenizer1, with a set maximum length of 1024 tokens, applying truncation as needed. The tokenizer is then configured to process the target variables, which are the summaries, encoding them into a tokenized form with a maximum length of 128 tokens, also truncating where necessary. The output of the tokenization includes the `input_ids` for BART, representing the numerical representation of the tokens.The function additionally facilitates labeling, which involves associating each input token with a corresponding label from the target tokenized summary. This is crucial for the model to learn the mapping from inputs to outputs during training.

To complement the preprocessing, a `DataCollatorForSeq2Seq` is used, which is a utility that organizes the tokenized data into batches. It ensures that within each batch, all sequences are padded to the same length, creating uniform input sizes for efficient processing by the sequence-to-sequence model during training. This padding is essential for leveraging the parallel computation capabilities of modern deep learning frameworks and for maintaining consistency across batches during training.

Model Training and Configuration

Both models are instantiated with their respective tokenizers and pre-trained weights. The BART model architecture comprises an encoder-decoder structure with multi-layer attention mechanisms, which is ideal for sequence-to-sequence tasks like summarization.

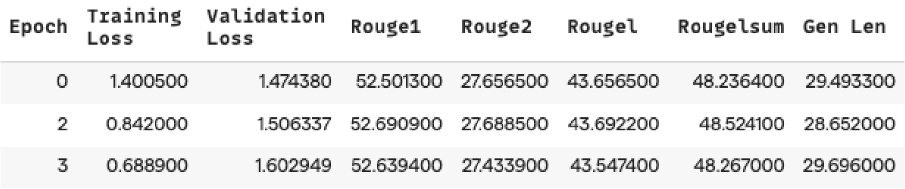
We employed a `Seq2SeqTrainer` from Hugging Face's Transformers library, configured with specific training arguments to optimize the learning process. Throughout the training process, we adjust hyperparameters, such as the learning rate and batch size, and monitor loss metrics to optimize the model's performance. Our arguments specify the number of training epochs, batch sizes, learning rates, and evaluation strategies, among other parameters. For instance, we set a learning rate of 2e-5, a train batch size of 4 per device, and employed gradient accumulation to handle higher effective batch sizes than what single GPU memory could allow. We utilize beam search (k=6) during generation to explore the space of possible summaries and select the one that the model predicts with the highest confidence.

Evaluation Metrics

Our custom evaluation function, compute_metrics1, decodes the model's predictions and computes the ROUGE scores to provide a quantitative measure of the summarization quality. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric is a commonly used metric for evaluating the quality of text summarization or generation tasks. It compares the overlap of n-grams between the generated summaries and the reference summaries. We measure ROUGE-1 (indicates overlap of sole words), ROUGE-2 (measures overlap of two-word sequences), and ROUGE-L (measures the Longest Common Subsequence between the two summaries) scores to assess the quality of our summarizations in terms of unigrams, bigrams, and the longest common subsequence, respectively.

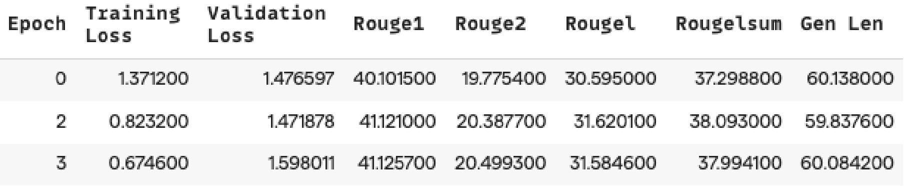
Fine-tuning and Results

Through iterative training and validation, we observed how the models adapted to the summarization task.



Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
0	1.400500	1.474380	52.501300	27.656500	43.656500	48.236400	29.493300
2	0.842000	1.506337	52.690900	27.688500	43.692200	48.524100	28.652000
3	0.688900	1.602949	52.639400	27.433900	43.547400	48.267000	29.696000

Model 1 (facebook/bart-large-xsum) Results



Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
0	1.371200	1.476597	40.101500	19.775400	30.595000	37.298800	60.138000
2	0.823200	1.471878	41.121000	20.387700	31.620100	38.093000	59.837600
3	0.674600	1.598011	41.125700	20.499300	31.584600	37.994100	60.084200

Model 2 (facebook/bart-large-cnn) Results

The models showed different levels of efficacy, with Model 1 `facebook/bart-large-xsum` generally performing better across the metrics, likely due to its training on a summarization-specific dataset (XSum). As we set `load_best_model_at_end = True` in the trainer's arguments, the training process utilizes a trainer that automatically selects the final model based on the smallest validation loss recorded during training. Therefore, for Model 1, the best-performing epoch according to this criterion would be epoch 0, as it has the lowest validation loss. Similarly, for Model 2, the trainer would select the model from epoch 2, where the validation loss is at its minimum.

Model 1’s performance surpasses standard baseline metrics, especially notable in ROUGE-1 scores, consistently above 50. This means that 52.639% of the unigrams in the reference summary are also present in the generated summary, a benchmark indicative of strong unigram matching. Similarly, Model 1's ROUGE-2 (-L) scores remain superior to Model 2's and above 27 (43), denoting an adeptness at capturing multigram relationships. The Gen Len metric (length of the generated summary) further reveals Model 1's capacity to generate succinct summaries, a desirable trait that underscores the model's ability to condense text into its most informative components. See the appendix for an example of input conversation and output summary generated by model 1.

Discussion

Implications

The BART model could change the way educational content is being presented by providing brief synopsis that help the learning person understand tough material easily and hence provide more personalized ways of learning. Their effectiveness, however, is dependent more on accurate understanding of various academic disciplines and especially new vocabularies and concepts involved through constant learning, more so in fields that have highly technical jargon be it in legal studies or science. This, therefore, implies the necessity of an ongoing model fine-tuning process that would assure not only its relevance but effectiveness in light of actuality.

Limitations

Meanwhile, the heavy computation resources needed for BART models will be not only limiting their application to resource-poor environments but also symbolizing a much larger problem for the broad accessibility of advanced AI technologies. This could disadvantage institutions and human beings in poorer areas, who are getting pushed farther into a corner in order to take advantage of the great opportunities for educational or professional development that these powerful tools offer, and thereby increase the digital divide. This difference indicates the need to create more resource-efficient models that can still be fully functional but are available to a larger audience.

Further, the biases that are embedded in the training data for these models carry considerable risk. If the training data used to train these summarization models is not curated with a fine and careful eye to balance and comprehensiveness, then the produced summaries will do more than just reflect back existing prejudices; they may also introduce biases or distort and even drop important information. This can have dire consequences, especially in those circumstances where precise and objective information is paramount, such as in legal, medical, or educational fields. Correcting for these biases demands effort both in the design of the dataset and in actively monitoring and updating the model to produce fair and truthful output.

Conclusion

Ultimately, these BART models represent a major advancement beyond the state-of-the-art in NLP, yet their real-world effectiveness and ethical deployment will rely on continued progress in model training, ethical considerations, and computational optimization. Should further advance, it is recommended that the systems minimize biases and computational demands for wider accessibility and fair treatment during automated text summarization.

Appendix

Below is the text summarization simulation we tested:

Original Dialogue:

Marta: Hi, I'm at the supermarket now to make some shopping for todays dinner. Do you have any wishes?
Nick: Hm I don't know. I haven't eat spaghetti in a while
Marta: Oh no, I've got spaghetti yesterday by Patric and the day before too.
Nick: Okay maybe some fish?
Marta: Yeah fish is great, I'll go and search for something
Nick: Text me what do you find.
Marta: Actually there is one small fish left and I don't think we will be full from it.
Nick: Let's make a lasagne
Marta: Do you know how much work that is?
Nick: No i don't know
Marta: It's a lot...
Nick: Maybe I could help you?
Marta: That's not a bad idea
Marta: Did you cook something yet?
Nick: No, but I can learn really fast :D
Marta: Okay, I'll buy the meat and sauce an let's do it
Nick: Should I look for a recipe in the Internet?
Marta: No need I did lasagne many times before
Nick: I can't wait ntil you teach me how to cook
Marta: I hope it will be eatable :D

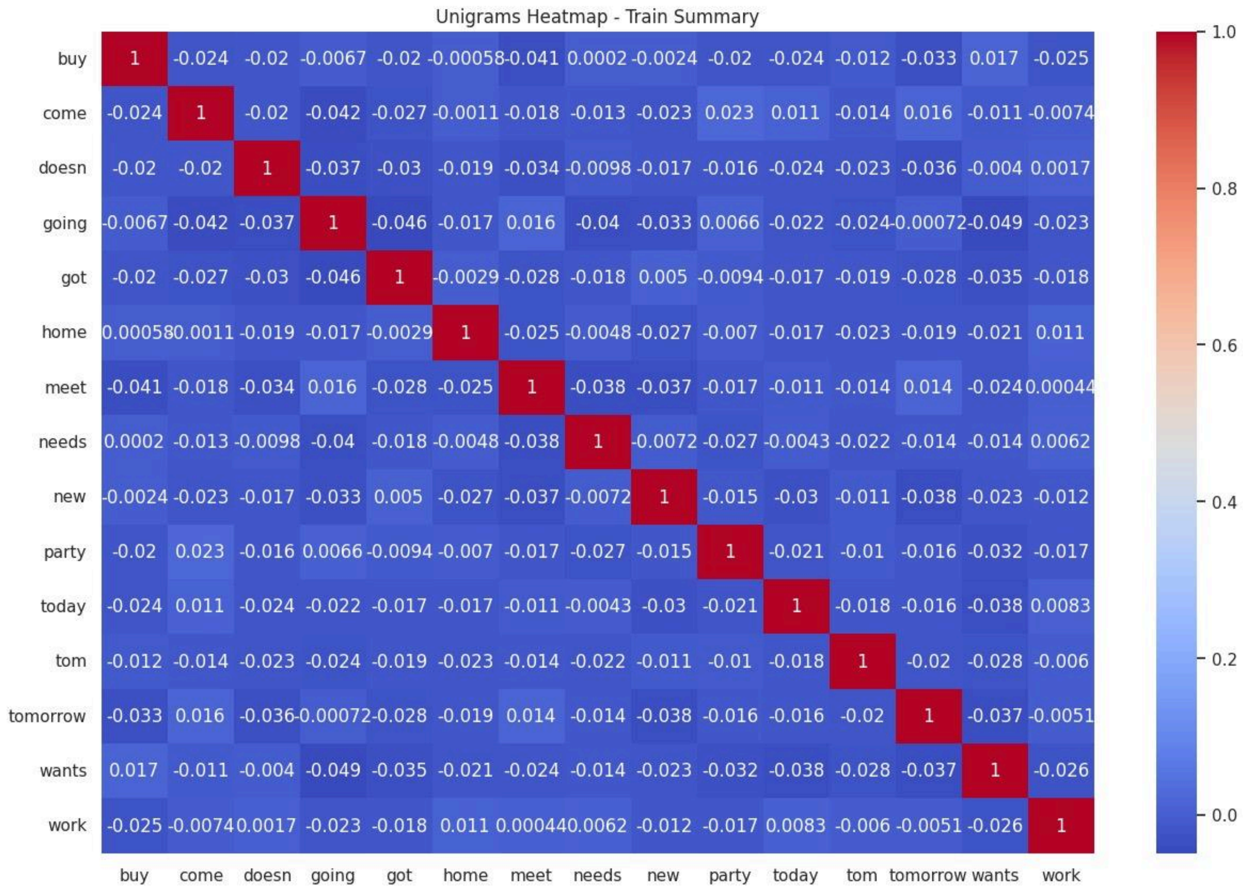
Reference Summary:

Marta is grocery shopping for dinner. She and Nick will make lasagne.

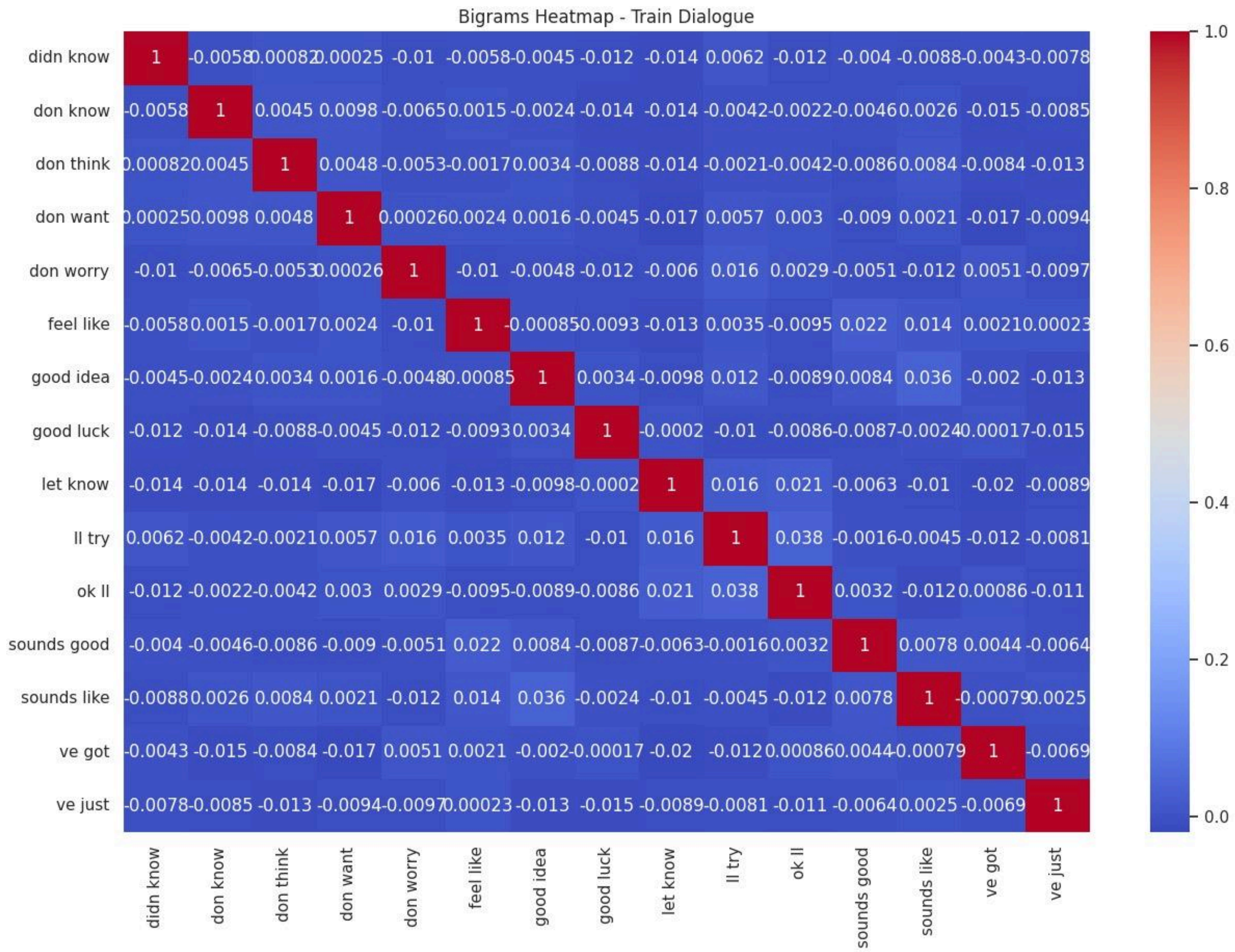
Model-generated Summary:

Marta is at the supermarket. She will buy fish, meat and sauce for today's dinner with Nick. They will make lasagna.

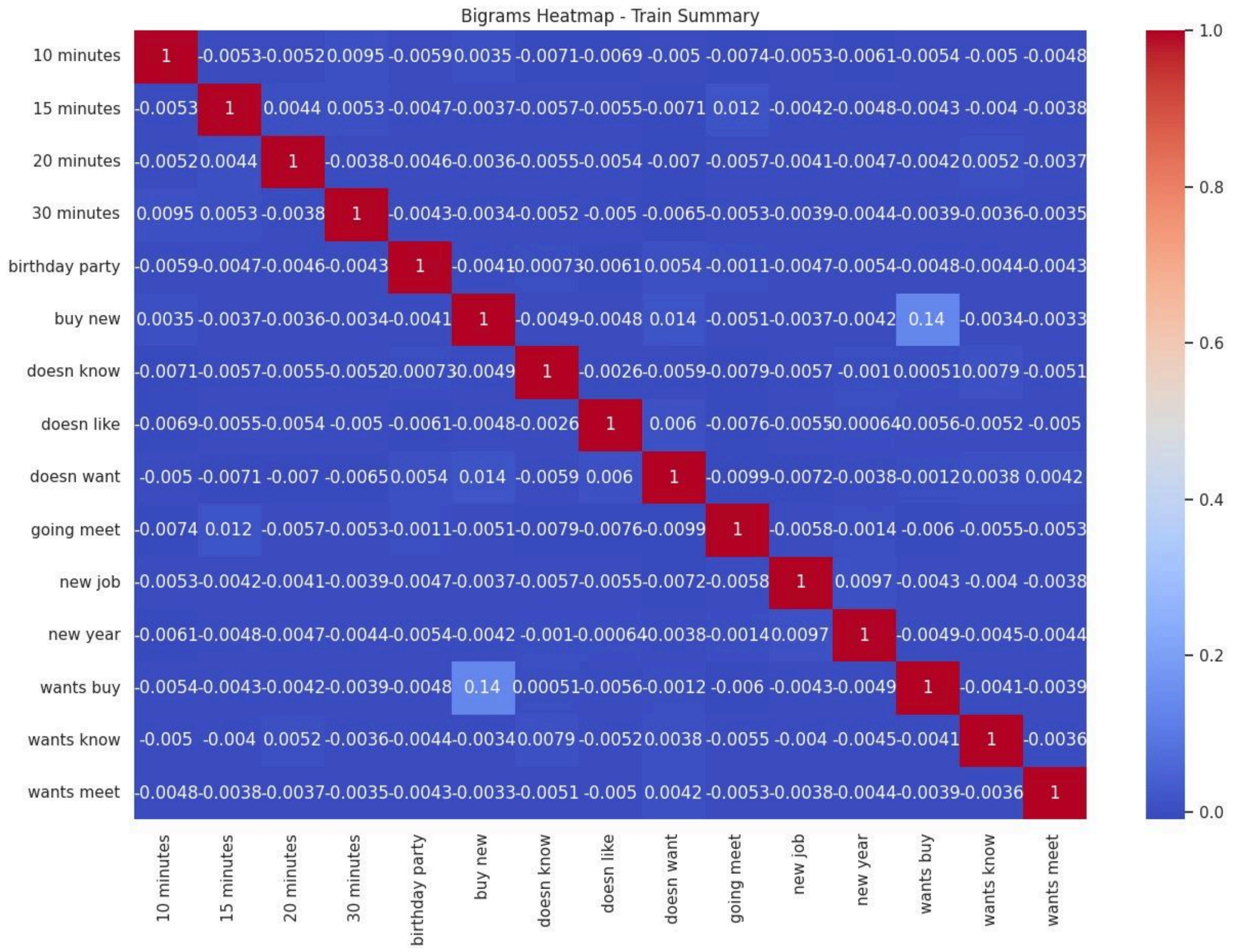
Appendix on heatmaps:
Heatmaps of dialogues and summary for unigram, bigram, and trigram in train set

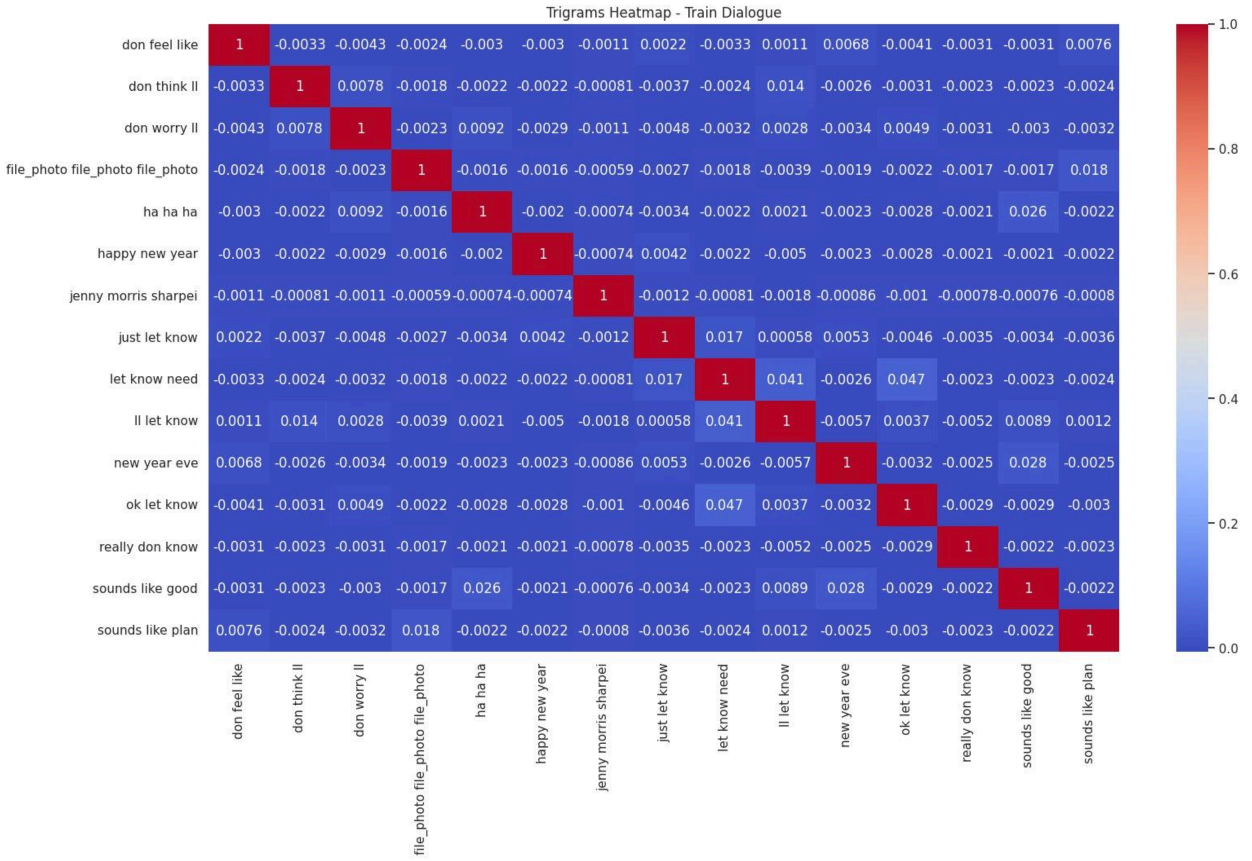


Unigrams Heatmap - Train Summary

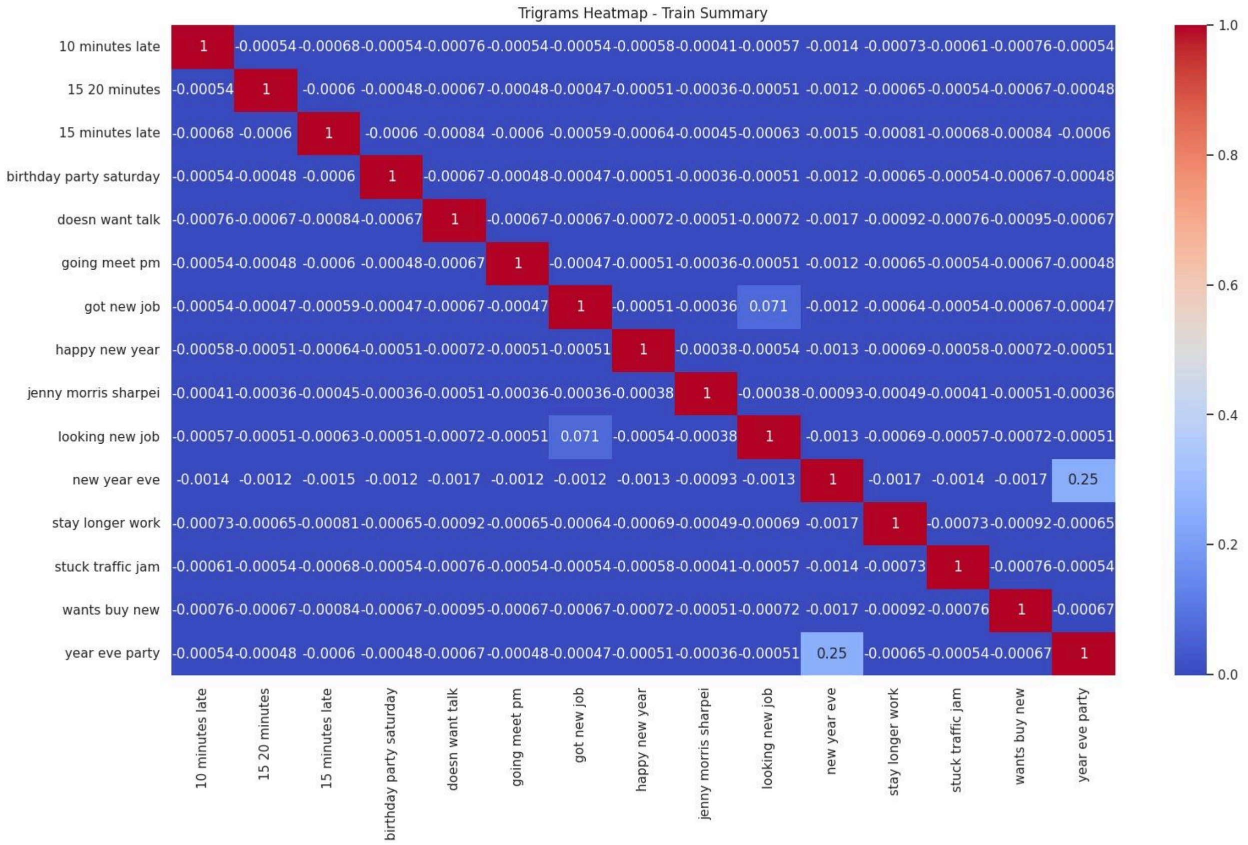


	didn know	don know	don think	don want	don worry	feel like	good idea	good luck	let know	ll try	ok ll	sounds good	sounds like	ve got	ve just
didn know	1	-0.0058	0.00082	0.00025	-0.01	-0.0058	-0.0045	-0.012	-0.014	0.0062	-0.012	-0.004	-0.0088	-0.0043	-0.0078
don know	-0.0058	1	0.0045	0.0098	-0.0065	0.0015	-0.0024	-0.014	-0.014	-0.0042	-0.0022	-0.0046	0.0026	-0.015	-0.0085
don think	0.00082	0.0045	1	0.0048	-0.0053	-0.0017	0.0034	-0.0088	-0.014	-0.0021	-0.0042	-0.0086	0.0084	-0.0084	-0.013
don want	0.00025	0.0098	0.0048	1	0.00026	0.0024	0.0016	-0.0045	-0.017	0.0057	0.003	-0.009	0.0021	-0.017	-0.0094
don worry	-0.01	-0.0065	-0.0053	0.00026	1	-0.01	-0.0048	-0.012	-0.006	0.016	0.0029	-0.0051	-0.012	0.0051	-0.0097
feel like	-0.0058	0.0015	-0.0017	0.0024	-0.01	1	-0.00085	-0.0093	-0.013	0.0035	-0.0095	0.022	0.014	0.0021	0.00023
good idea	-0.0045	-0.0024	0.0034	0.0016	-0.0048	-0.00085	1	0.0034	-0.0098	0.012	-0.0089	0.0084	0.036	-0.002	-0.013
good luck	-0.012	-0.014	-0.0088	-0.0045	-0.012	-0.0093	0.0034	1	-0.0002	-0.01	-0.0086	-0.0087	-0.0024	-0.00017	-0.015
let know	-0.014	-0.014	-0.014	-0.017	-0.006	-0.013	-0.0098	-0.0002	1	0.016	0.021	-0.0063	-0.01	-0.02	-0.0089
ll try	0.0062	-0.0042	-0.0021	0.0057	0.016	0.0035	0.012	-0.01	0.016	1	0.038	-0.0016	-0.0045	-0.012	-0.0081
ok ll	-0.012	-0.0022	-0.0042	0.003	0.0029	-0.0095	-0.0089	-0.0086	0.021	0.038	1	0.0032	-0.012	0.00086	-0.011
sounds good	-0.004	-0.0046	-0.0086	-0.009	-0.0051	0.022	0.0084	-0.0087	-0.0063	-0.0016	0.0032	1	0.0078	0.0044	-0.0064
sounds like	-0.0088	0.0026	0.0084	0.0021	-0.012	0.014	0.036	-0.0024	-0.01	-0.0045	-0.012	0.0078	1	-0.00079	0.0025
ve got	-0.0043	-0.015	-0.0084	-0.017	0.0051	0.0021	-0.002	-0.00017	-0.02	-0.012	0.00086	0.0044	-0.00079	1	-0.0069
ve just	-0.0078	-0.0085	-0.013	-0.0094	-0.0097	0.00023	-0.013	-0.015	-0.0089	-0.0081	-0.011	-0.0064	0.0025	-0.0069	1

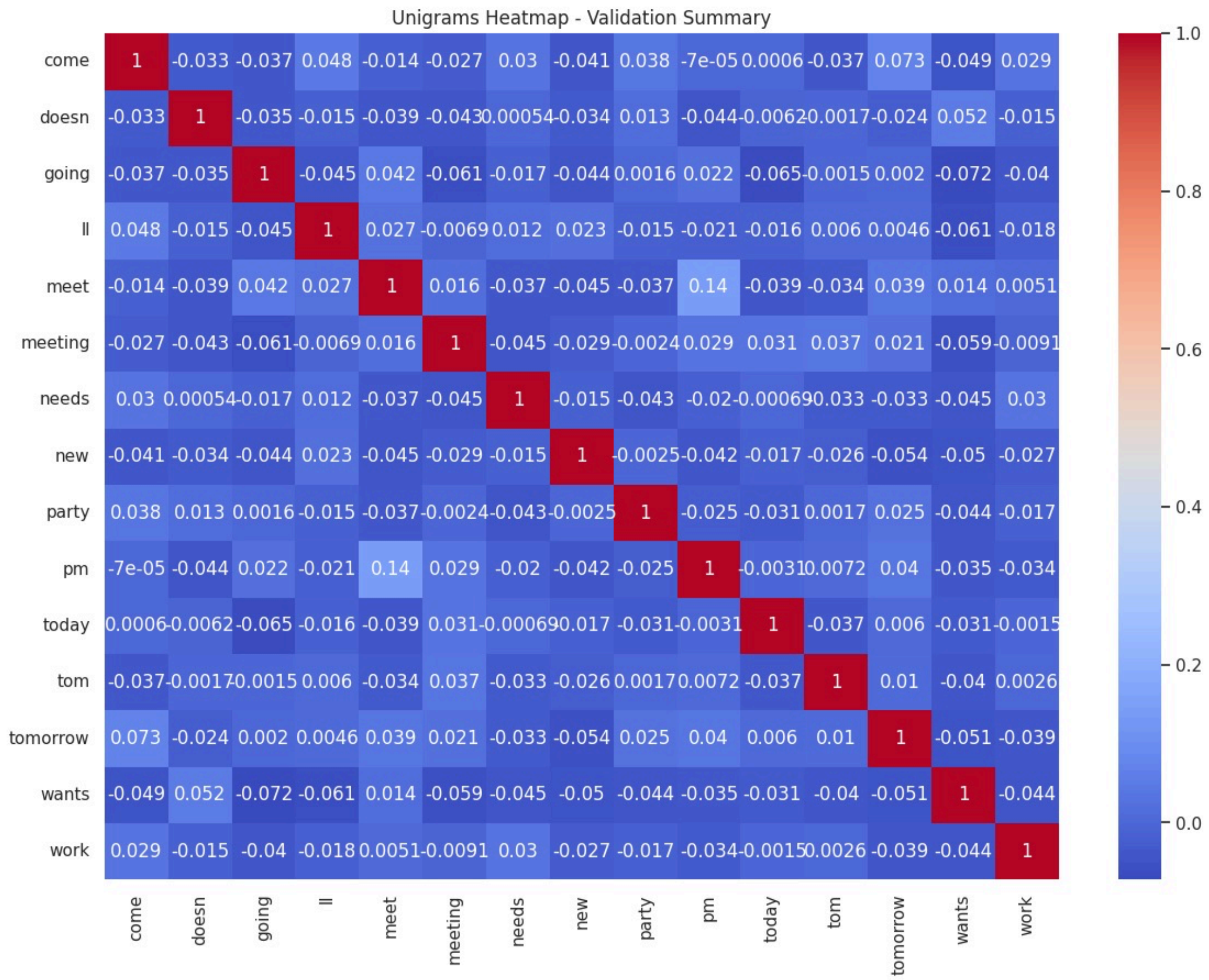

Bigrams Heatmap - Train Summary
10 minutes 1 -0.0053 -0.0052 0.0095 -0.0059 0.0035 -0.0071 -0.0069 -0.005 -0.0074 -0.0053 -0.0061 -0.0054 -0.005 -0.0048
15 minutes -0.0053 1 0.0044 0.0053 -0.0047 -0.0037 -0.0057 -0.0055 -0.0071 0.012 -0.0042 -0.0048 -0.0043 -0.004 -0.0038
20 minutes -0.0052 0.0044 1 -0.0038 -0.0046 -0.0036 -0.0055 -0.0054 -0.007 -0.0057 -0.0041 -0.0047 -0.0042 0.0052 -0.0037
30 minutes 0.0095 0.0053 -0.0038 1 -0.0043 -0.0034 -0.0052 -0.005 -0.0065 -0.0053 -0.0039 -0.0044 -0.0039 -0.0036 -0.0035
birthday party -0.0059 -0.0047 -0.0046 -0.0043 1 -0.0041 0.00073 0.0061 0.0054 -0.0011 -0.0047 -0.0054 -0.0048 -0.0044 -0.0043
buy new 0.0035 -0.0037 -0.0036 -0.0034 -0.0041 1 -0.0049 -0.0048 0.014 -0.0051 -0.0037 -0.0042 0.14 -0.0034 -0.0033
doesn know -0.0071 -0.0057 -0.0055 -0.0052 0.00073 0.0049 1 -0.0026 -0.0059 -0.0079 -0.0057 -0.001 0.00051 0.0079 -0.0051
doesn like -0.0069 -0.0055 -0.0054 -0.005 -0.0061 -0.0048 -0.0026 1 0.006 -0.0076 -0.0055 0.00064 0.0056 -0.0052 -0.005
doesn want -0.005 -0.0071 -0.007 -0.0065 0.0054 0.014 -0.0059 0.006 1 -0.0099 -0.0072 -0.0038 -0.0012 0.0038 0.0042
going meet -0.0074 0.012 -0.0057 -0.0053 -0.0011 -0.0051 -0.0079 -0.0076 -0.0099 1 -0.0058 -0.0014 -0.006 -0.0055 -0.0053
new job -0.0053 -0.0042 -0.0041 -0.0039 -0.0047 -0.0037 -0.0057 -0.0055 -0.0072 -0.0058 1 0.0097 -0.0043 -0.004 -0.0038
new year -0.0061 -0.0048 -0.0047 -0.0044 -0.0054 -0.0042 -0.001 -0.00064 0.0038 -0.0014 0.0097 1 -0.0049 -0.0045 -0.0044
wants buy -0.0054 -0.0043 -0.0042 -0.0039 -0.0048 0.14 0.00051 -0.0056 -0.0012 -0.006 -0.0043 -0.0049 1 -0.0041 -0.0039
wants know -0.005 -0.004 0.0052 -0.0036 -0.0044 -0.0034 0.0079 -0.0052 0.0038 -0.0055 -0.004 -0.0045 -0.0041 1 -0.0036
wants meet -0.0048 -0.0038 -0.0037 -0.0035 -0.0043 -0.0033 -0.0051 -0.005 0.0042 -0.0053 -0.0038 -0.0044 -0.0039 -0.0036 1
10 minutes
15 minutes
20 minutes
30 minutes
birthday party
buy new
doesn know
doesn like
doesn want
going meet
new job
new year
wants buy
wants know
wants meet
1.0
0.8
0.6
0.4
0.2
0.0



	don feel like	don think ll	don worry ll	file_photo file_photo file_photo	ha ha ha	happy new year	jenny morris sharpei	just let know	let know need	ll let know	new year eve	ok let know	really don know	sounds like good	sounds like plan
don feel like	1	-0.0033	-0.0043	-0.0024	-0.003	-0.003	-0.0011	0.0022	-0.0033	0.0011	0.0068	-0.0041	-0.0031	-0.0031	0.0076
don think ll	-0.0033	1	0.0078	-0.0018	-0.0022	-0.0022	-0.00081	-0.0037	-0.0024	0.014	-0.0026	-0.0031	-0.0023	-0.0023	-0.0024
don worry ll	-0.0043	0.0078	1	-0.0023	0.0092	-0.0029	-0.0011	-0.0048	-0.0032	0.0028	-0.0034	0.0049	-0.0031	-0.003	-0.0032
file_photo file_photo file_photo	-0.0024	-0.0018	-0.0023	1	-0.0016	-0.0016	-0.00059	-0.0027	-0.0018	-0.0039	-0.0019	-0.0022	-0.0017	-0.0017	0.018
ha ha ha	-0.003	-0.0022	0.0092	-0.0016	1	-0.002	-0.00074	-0.0034	-0.0022	0.0021	-0.0023	-0.0028	-0.0021	0.026	-0.0022
happy new year	-0.003	-0.0022	-0.0029	-0.0016	-0.002	1	-0.00074	0.0042	-0.0022	-0.005	-0.0023	-0.0028	-0.0021	-0.0021	-0.0022
jenny morris sharpei	-0.0011	-0.00081	-0.0011	-0.00059	-0.00074	-0.00074	1	-0.0012	-0.00081	-0.0018	-0.00086	-0.001	-0.00078	-0.00076	-0.0008
just let know	0.0022	-0.0037	-0.0048	-0.0027	-0.0034	0.0042	-0.0012	1	0.017	0.00058	0.0053	-0.0046	-0.0035	-0.0034	-0.0036
let know need	-0.0033	-0.0024	-0.0032	-0.0018	-0.0022	-0.0022	-0.00081	0.017	1	0.041	-0.0026	0.047	-0.0023	-0.0023	-0.0024
ll let know	0.0011	0.014	0.0028	-0.0039	0.0021	-0.005	-0.0018	0.00058	0.041	1	-0.0057	0.0037	-0.0052	0.0089	0.0012
new year eve	0.0068	-0.0026	-0.0034	-0.0019	-0.0023	-0.0023	-0.00086	0.0053	-0.0026	-0.0057	1	-0.0032	-0.0025	0.028	-0.0025
ok let know	-0.0041	-0.0031	0.0049	-0.0022	-0.0028	-0.0028	-0.001	-0.0046	0.047	0.0037	-0.0032	1	-0.0029	-0.0029	-0.003
really don know	-0.0031	-0.0023	-0.0031	-0.0017	-0.0021	-0.0021	-0.00078	-0.0035	-0.0023	-0.0052	-0.0025	-0.0029	1	-0.0022	-0.0023
sounds like good	-0.0031	-0.0023	-0.003	-0.0017	0.026	-0.0021	-0.00076	-0.0034	-0.0023	0.0089	0.028	-0.0029	-0.0022	1	-0.0022
sounds like plan	0.0076	-0.0024	-0.0032	0.018	-0.0022	-0.0022	-0.0008	-0.0036	-0.0024	0.0012	-0.0025	-0.003	-0.0023	-0.0022	1

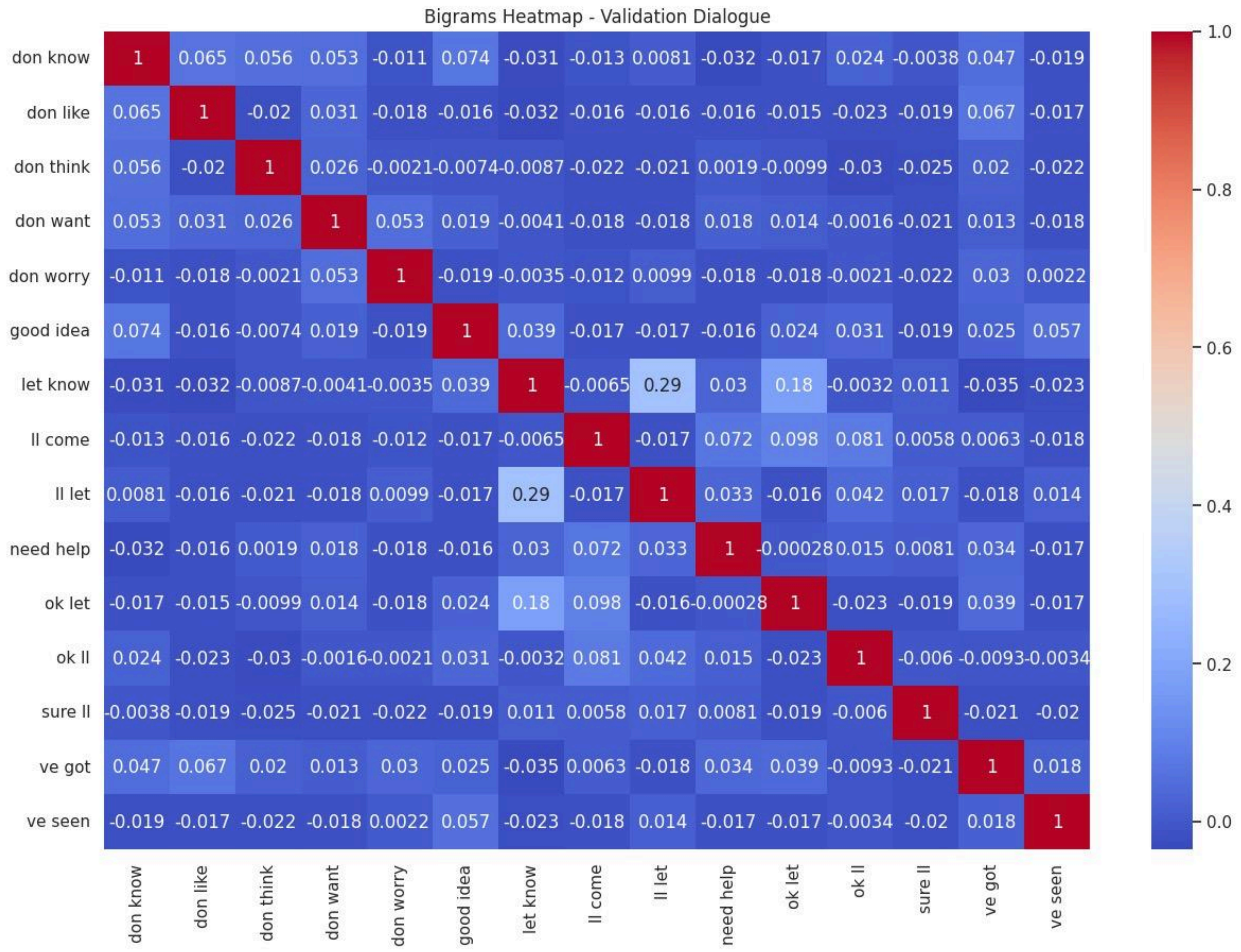


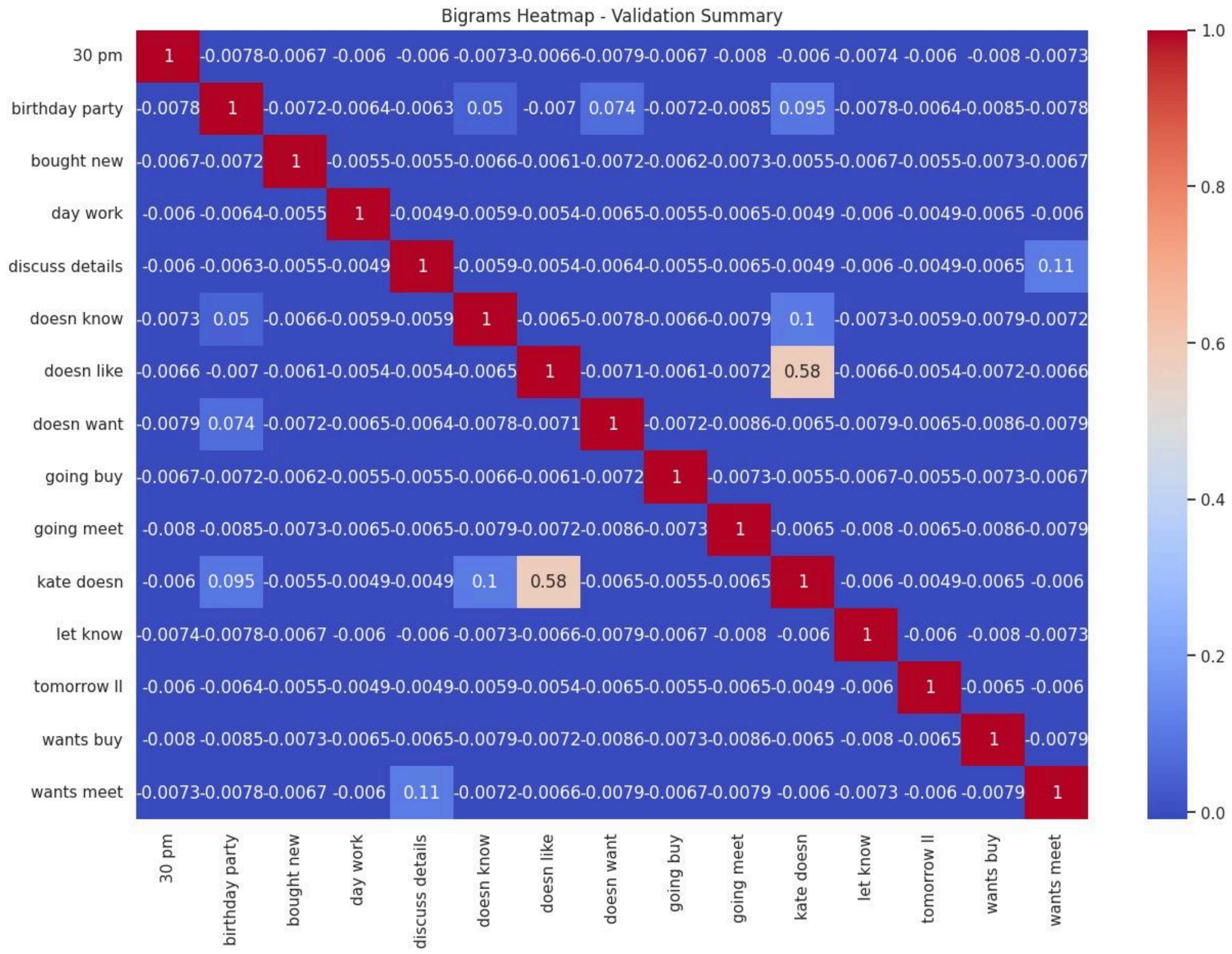
	10 minutes late	15 20 minutes	15 minutes late	birthday party saturday	doesn want talk	going meet pm	got new job	happy new year	jenny morris sharpei	looking new job	new year eve	stay longer work	stuck traffic jam	wants buy new	year eve party
10 minutes late	1	-0.00054	-0.00068	-0.00054	-0.00076	-0.00054	-0.00054	-0.00058	-0.00041	-0.00057	-0.0014	-0.00073	-0.00061	-0.00076	-0.00054
15 20 minutes	-0.00054	1	-0.0006	-0.00048	-0.00067	-0.00048	-0.00047	-0.00051	-0.00036	-0.00051	-0.0012	-0.00065	-0.00054	-0.00067	-0.00048
15 minutes late	-0.00068	-0.0006	1	-0.0006	-0.00084	-0.0006	-0.00059	-0.00064	-0.00045	-0.00063	-0.0015	-0.00081	-0.00068	-0.00084	-0.0006
birthday party saturday	-0.00054	-0.00048	-0.0006	1	-0.00067	-0.00048	-0.00047	-0.00051	-0.00036	-0.00051	-0.0012	-0.00065	-0.00054	-0.00067	-0.00048
doesn want talk	-0.00076	-0.00067	-0.00084	-0.00067	1	-0.00067	-0.00067	-0.00072	-0.00051	-0.00072	-0.0017	-0.00092	-0.00076	-0.00095	-0.00067
going meet pm	-0.00054	-0.00048	-0.0006	-0.00048	-0.00067	1	-0.00047	-0.00051	-0.00036	-0.00051	-0.0012	-0.00065	-0.00054	-0.00067	-0.00048
got new job	-0.00054	-0.00047	-0.00059	-0.00047	-0.00067	-0.00047	1	-0.00051	-0.00036	0.071	-0.0012	-0.00064	-0.00054	-0.00067	-0.00047
happy new year	-0.00058	-0.00051	-0.00064	-0.00051	-0.00072	-0.00051	-0.00051	1	-0.00038	-0.00054	-0.0013	-0.00069	-0.00058	-0.00072	-0.00051
jenny morris sharpei	-0.00041	-0.00036	-0.00045	-0.00036	-0.00051	-0.00036	-0.00036	-0.00038	1	-0.00038	-0.00093	-0.00049	-0.00041	-0.00051	-0.00036
looking new job	-0.00057	-0.00051	-0.00063	-0.00051	-0.00072	-0.00051	0.071	-0.00054	-0.00038	1	-0.0013	-0.00069	-0.00057	-0.00072	-0.00051
new year eve	-0.0014	-0.0012	-0.0015	-0.0012	-0.0017	-0.0012	-0.0012	-0.0013	-0.00093	-0.0013	1	-0.0017	-0.0014	-0.0017	0.25
stay longer work	-0.00073	-0.00065	-0.00081	-0.00065	-0.00092	-0.00065	-0.00064	-0.00069	-0.00049	-0.00069	-0.0017	1	-0.00073	-0.00092	-0.00065
stuck traffic jam	-0.00061	-0.00054	-0.00068	-0.00054	-0.00076	-0.00054	-0.00054	-0.00058	-0.00041	-0.00057	-0.0014	-0.00073	1	-0.00076	-0.00054
wants buy new	-0.00076	-0.00067	-0.00084	-0.00067	-0.00095	-0.00067	-0.00067	-0.00072	-0.00051	-0.00072	-0.0017	-0.00092	-0.00076	1	-0.00067
year eve party	-0.00054	-0.00048	-0.0006	-0.00048	-0.00067	-0.00048	-0.00047	-0.00051	-0.00036	-0.00051	0.25	-0.00065	-0.00054	-0.00067	1

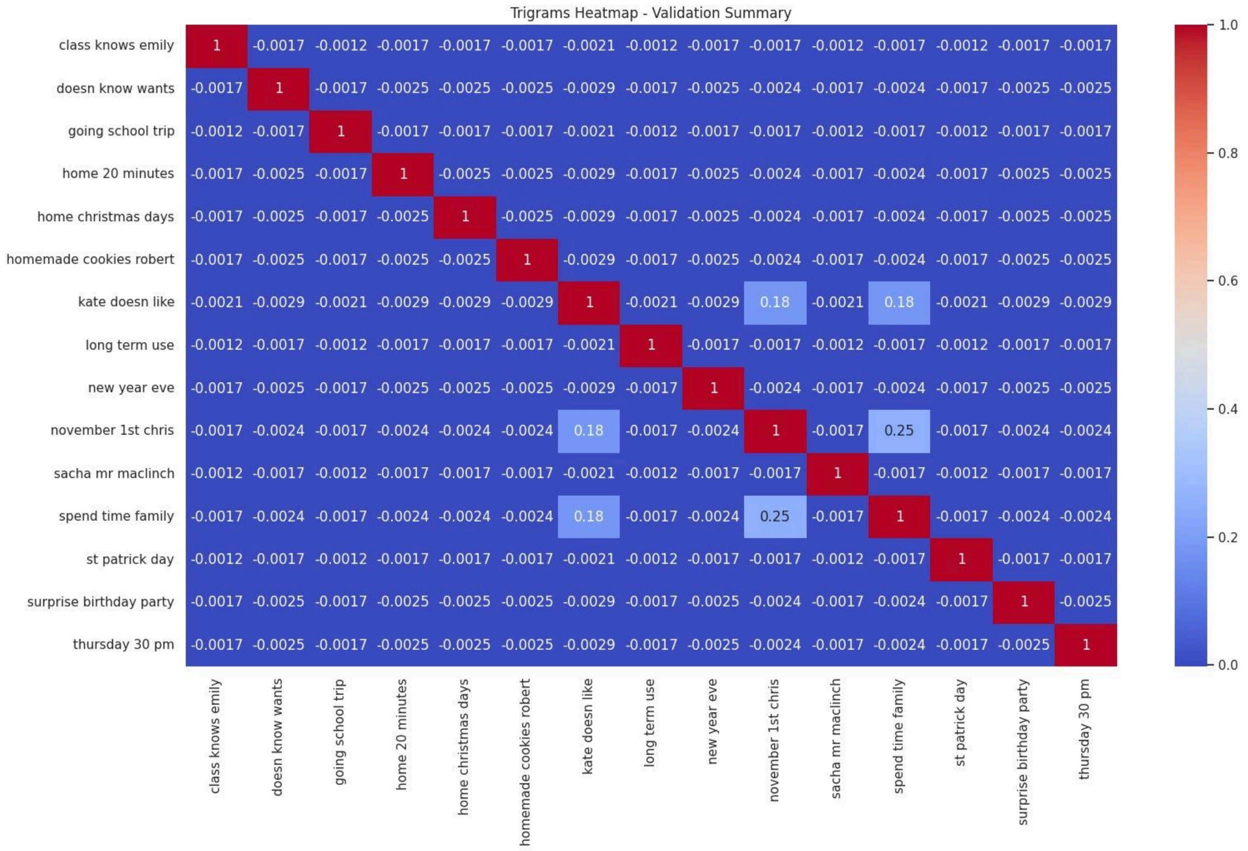


Unigrams Heatmap - Validation Summary

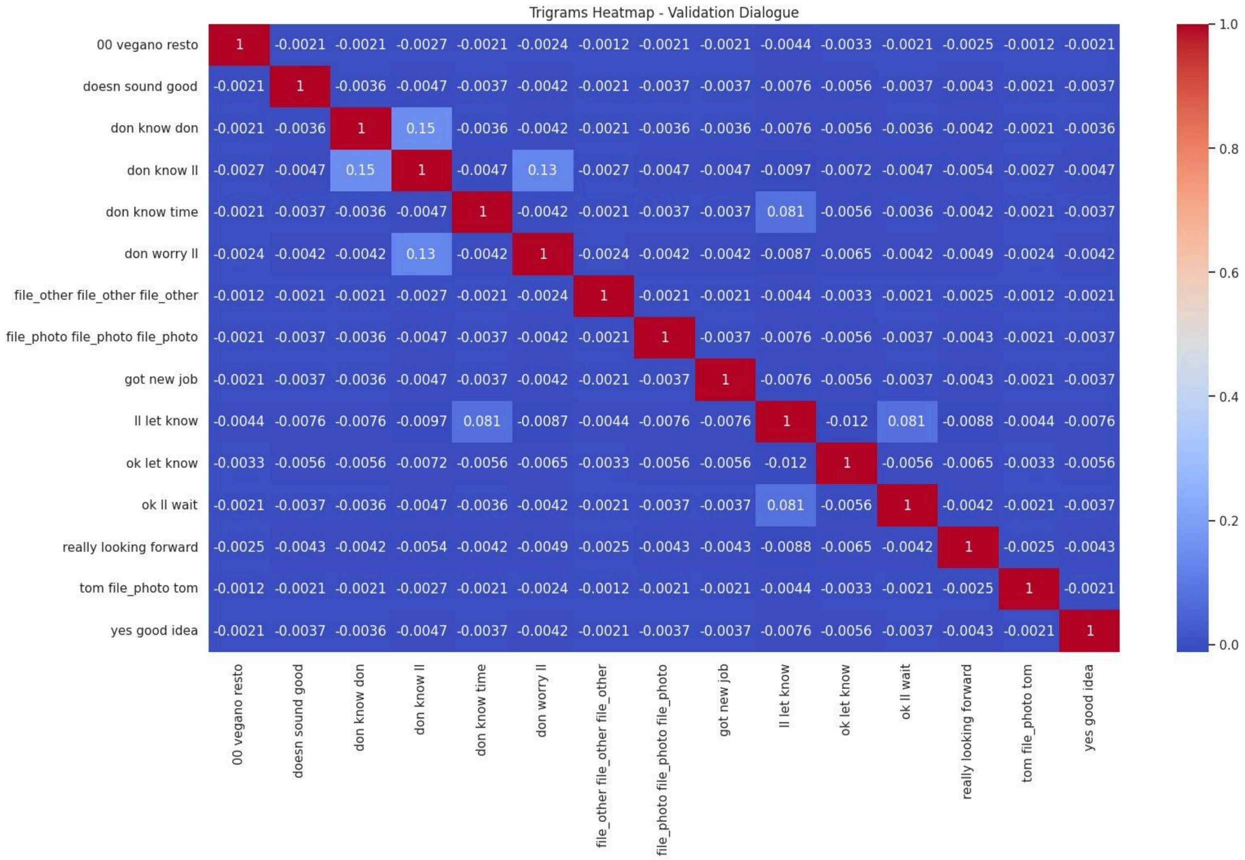
	come	doesn	going	ll	meet	meeting	needs	new	party	pm	today	tom	tomorrow	wants	work
come	1	-0.033	-0.037	0.048	-0.014	-0.027	0.03	-0.041	0.038	-7e-05	0.0006	-0.037	0.073	-0.049	0.029
doesn	-0.033	1	-0.035	-0.015	-0.039	-0.043	0.00054	-0.034	0.013	-0.044	-0.0062	-0.0017	-0.024	0.052	-0.015
going	-0.037	-0.035	1	-0.045	0.042	-0.061	-0.017	-0.044	0.0016	0.022	-0.065	-0.0015	0.002	-0.072	-0.04
ll	0.048	-0.015	-0.045	1	0.027	-0.0069	0.012	0.023	-0.015	-0.021	-0.016	0.006	0.0046	-0.061	-0.018
meet	-0.014	-0.039	0.042	0.027	1	0.016	-0.037	-0.045	-0.037	0.14	-0.039	-0.034	0.039	0.014	0.0051
meeting	-0.027	-0.043	-0.061	-0.0069	0.016	1	-0.045	-0.029	-0.0024	0.029	0.031	0.037	0.021	-0.059	-0.0091
needs	0.03	0.00054	-0.017	0.012	-0.037	-0.045	1	-0.015	-0.043	-0.02	-0.00069	-0.033	-0.033	-0.045	0.03
new	-0.041	-0.034	-0.044	0.023	-0.045	-0.029	-0.015	1	-0.0025	-0.042	-0.017	-0.026	-0.054	-0.05	-0.027
party	0.038	0.013	0.0016	-0.015	-0.037	-0.0024	-0.043	-0.0025	1	-0.025	-0.031	0.0017	0.025	-0.044	-0.017
pm	-7e-05	-0.044	0.022	-0.021	0.14	0.029	-0.02	-0.042	-0.025	1	-0.0031	0.0072	0.04	-0.035	-0.034
today	0.0006	-0.0062	-0.065	-0.016	-0.039	0.031	-0.00069	-0.017	-0.031	-0.0031	1	-0.037	0.006	-0.031	-0.0015
tom	-0.037	-0.0017	-0.0015	0.006	-0.034	0.037	-0.033	-0.026	0.0017	0.0072	-0.037	1	0.01	-0.04	0.0026
tomorrow	0.073	-0.024	0.002	0.0046	0.039	0.021	-0.033	-0.054	0.025	0.04	0.006	0.01	1	-0.051	-0.039
wants	-0.049	0.052	-0.072	-0.061	0.014	-0.059	-0.045	-0.05	-0.044	-0.035	-0.031	-0.04	-0.051	1	-0.044
work	0.029	-0.015	-0.04	-0.018	0.0051	-0.0091	0.03	-0.027	-0.017	-0.034	-0.0015	0.0026	-0.039	-0.044	1


Bigrams Heatmap - Validation Dialogue
don know 1 0.065 0.056 0.053 -0.011 0.074 -0.031 -0.013 0.0081 -0.032 -0.017 0.024 -0.0038 0.047 -0.019
don like 0.065 1 -0.02 0.031 -0.018 -0.016 -0.032 -0.016 -0.016 -0.016 -0.015 -0.023 -0.019 0.067 -0.017
don think 0.056 -0.02 1 0.026 -0.0021 -0.0074 -0.0087 -0.022 -0.021 0.0019 -0.0099 -0.03 -0.025 0.02 -0.022
don want 0.053 0.031 0.026 1 0.053 0.019 -0.0041 -0.018 -0.018 0.018 0.014 -0.0016 -0.021 0.013 -0.018
don worry -0.011 -0.018 -0.0021 0.053 1 -0.019 -0.0035 -0.012 0.0099 -0.018 -0.018 -0.0021 -0.022 0.03 0.0022
good idea 0.074 -0.016 -0.0074 0.019 -0.019 1 0.039 -0.017 -0.017 -0.016 0.024 0.031 -0.019 0.025 0.057
let know -0.031 -0.032 -0.0087 -0.0041 -0.0035 0.039 1 -0.0065 0.29 0.03 0.18 -0.0032 0.011 -0.035 -0.023
ll come -0.013 -0.016 -0.022 -0.018 -0.012 -0.017 -0.0065 1 -0.017 0.072 0.098 0.081 0.0058 0.0063 -0.018
ll let 0.0081 -0.016 -0.021 -0.018 0.0099 -0.017 0.29 -0.017 1 0.033 -0.016 0.042 0.017 -0.018 0.014
need help -0.032 -0.016 0.0019 0.018 -0.018 -0.016 0.03 0.072 0.033 1 -0.00028 0.015 0.0081 0.034 -0.017
ok let -0.017 -0.015 -0.0099 0.014 -0.018 0.024 0.18 0.098 -0.016 -0.00028 1 -0.023 -0.019 0.039 -0.017
ok ll 0.024 -0.023 -0.03 -0.0016 -0.0021 0.031 -0.0032 0.081 0.042 0.015 -0.023 1 -0.006 -0.0093 -0.0034
sure ll -0.0038 -0.019 -0.025 -0.021 -0.022 -0.019 0.011 0.0058 0.017 0.0081 -0.019 -0.006 1 -0.021 -0.02
ve got 0.047 0.067 0.02 0.013 0.03 0.025 -0.035 0.0063 -0.018 0.034 0.039 -0.0093 -0.021 1 0.018
ve seen -0.019 -0.017 -0.022 -0.018 0.0022 0.057 -0.023 -0.018 0.014 -0.017 -0.017 -0.0034 -0.02 0.018 1
don know
don like
don think
don want
don worry
good idea
let know
ll come
ll let
need help
ok let
ok ll
sure ll
ve got
ve seen
1.0
0.8
0.6
0.4
0.2
0.0

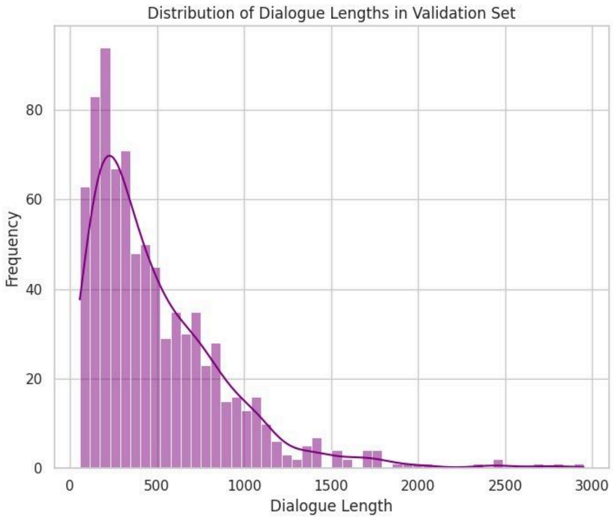

Bigrams Heatmap - Validation Summary

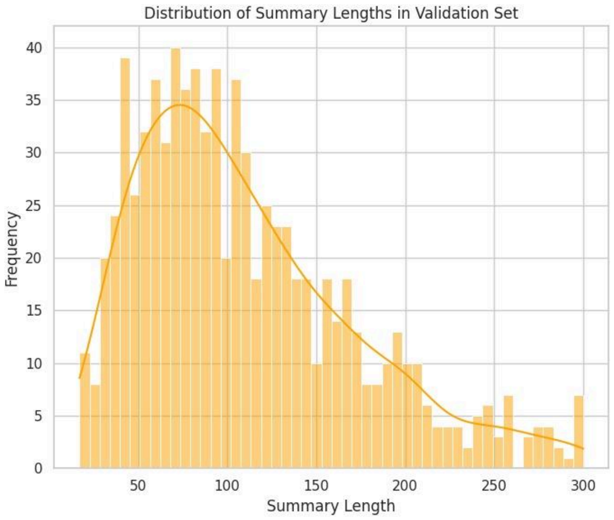


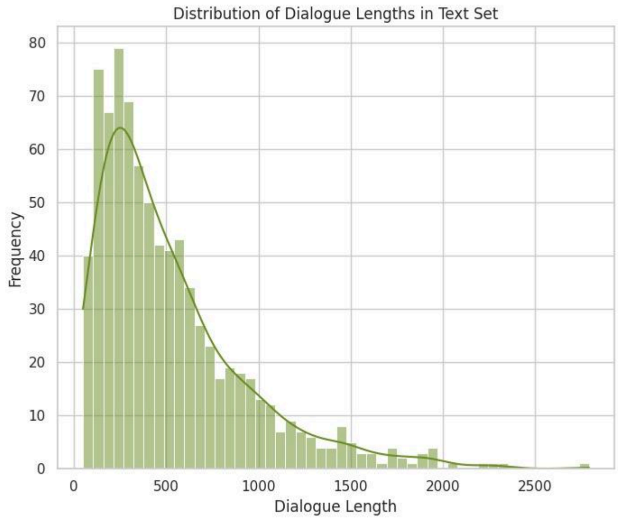
	class knows emily	doesn know wants	going school trip	home 20 minutes	home christmas days	homemade cookies robert	kate doesn like	long term use	new year eve	november 1st chris	sacha mr maclinch	spend time family	st patrick day	surprise birthday party	thursday 30 pm
class knows emily	1	-0.0017	-0.0012	-0.0017	-0.0017	-0.0017	-0.0021	-0.0012	-0.0017	-0.0017	-0.0012	-0.0017	-0.0012	-0.0017	-0.0017
doesn know wants	-0.0017	1	-0.0017	-0.0025	-0.0025	-0.0025	-0.0029	-0.0017	-0.0025	-0.0024	-0.0017	-0.0024	-0.0017	-0.0025	-0.0025
going school trip	-0.0012	-0.0017	1	-0.0017	-0.0017	-0.0017	-0.0021	-0.0012	-0.0017	-0.0017	-0.0012	-0.0017	-0.0012	-0.0017	-0.0017
home 20 minutes	-0.0017	-0.0025	-0.0017	1	-0.0025	-0.0025	-0.0029	-0.0017	-0.0025	-0.0024	-0.0017	-0.0024	-0.0017	-0.0025	-0.0025
home christmas days	-0.0017	-0.0025	-0.0017	-0.0025	1	-0.0025	-0.0029	-0.0017	-0.0025	-0.0024	-0.0017	-0.0024	-0.0017	-0.0025	-0.0025
homemade cookies robert	-0.0017	-0.0025	-0.0017	-0.0025	-0.0025	1	-0.0029	-0.0017	-0.0025	-0.0024	-0.0017	-0.0024	-0.0017	-0.0025	-0.0025
kate doesn like	-0.0021	-0.0029	-0.0021	-0.0029	-0.0029	-0.0029	1	-0.0021	-0.0029	0.18	-0.0021	0.18	-0.0021	-0.0029	-0.0029
long term use	-0.0012	-0.0017	-0.0012	-0.0017	-0.0017	-0.0017	-0.0021	1	-0.0017	-0.0017	-0.0012	-0.0017	-0.0012	-0.0017	-0.0017
new year eve	-0.0017	-0.0025	-0.0017	-0.0025	-0.0025	-0.0025	-0.0029	-0.0017	1	-0.0024	-0.0017	-0.0024	-0.0017	-0.0025	-0.0025
november 1st chris	-0.0017	-0.0024	-0.0017	-0.0024	-0.0024	-0.0024	0.18	-0.0017	-0.0024	1	-0.0017	0.25	-0.0017	-0.0024	-0.0024
sacha mr maclinch	-0.0012	-0.0017	-0.0012	-0.0017	-0.0017	-0.0017	-0.0021	-0.0012	-0.0017	-0.0017	1	-0.0017	-0.0012	-0.0017	-0.0017
spend time family	-0.0017	-0.0024	-0.0017	-0.0024	-0.0024	-0.0024	0.18	-0.0017	-0.0024	0.25	-0.0017	1	-0.0017	-0.0024	-0.0024
st patrick day	-0.0012	-0.0017	-0.0012	-0.0017	-0.0017	-0.0017	-0.0021	-0.0012	-0.0017	-0.0017	-0.0012	-0.0017	1	-0.0017	-0.0017
surprise birthday party	-0.0017	-0.0025	-0.0017	-0.0025	-0.0025	-0.0025	-0.0029	-0.0017	-0.0025	-0.0024	-0.0017	-0.0024	-0.0017	1	-0.0025
thursday 30 pm	-0.0017	-0.0025	-0.0017	-0.0025	-0.0025	-0.0025	-0.0029	-0.0017	-0.0025	-0.0024	-0.0017	-0.0024	-0.0017	-0.0025	1

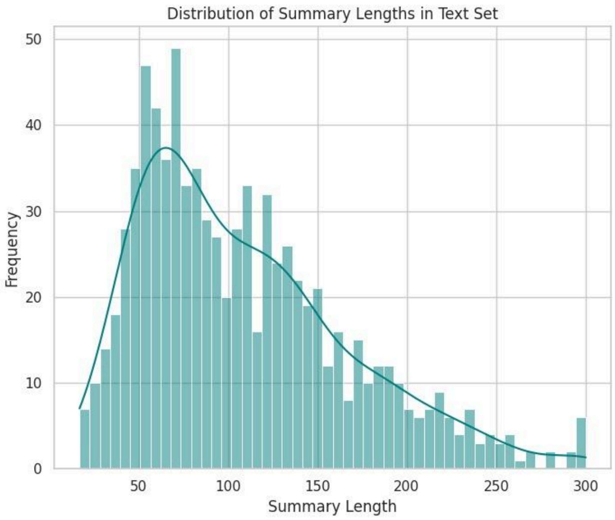


	00 vegano resto	doesn sound good	don know don	don know ll	don know time	don worry ll	file_other file_other file_other	file_photo file_photo file_photo	got new job	ll let know	ok let know	ok ll wait	really looking forward	tom file_photo tom	yes good idea
00 vegano resto	1	-0.0021	-0.0021	-0.0027	-0.0021	-0.0024	-0.0012	-0.0021	-0.0021	-0.0044	-0.0033	-0.0021	-0.0025	-0.0012	-0.0021
doesn sound good	-0.0021	1	-0.0036	-0.0047	-0.0037	-0.0042	-0.0021	-0.0037	-0.0037	-0.0076	-0.0056	-0.0037	-0.0043	-0.0021	-0.0037
don know don	-0.0021	-0.0036	1	0.15	-0.0036	-0.0042	-0.0021	-0.0036	-0.0036	-0.0076	-0.0056	-0.0036	-0.0042	-0.0021	-0.0036
don know ll	-0.0027	-0.0047	0.15	1	-0.0047	0.13	-0.0027	-0.0047	-0.0047	-0.0097	-0.0072	-0.0047	-0.0054	-0.0027	-0.0047
don know time	-0.0021	-0.0037	-0.0036	-0.0047	1	-0.0042	-0.0021	-0.0037	-0.0037	0.081	-0.0056	-0.0036	-0.0042	-0.0021	-0.0037
don worry ll	-0.0024	-0.0042	-0.0042	0.13	-0.0042	1	-0.0024	-0.0042	-0.0042	-0.0087	-0.0065	-0.0042	-0.0049	-0.0024	-0.0042
file_other file_other file_other	-0.0012	-0.0021	-0.0021	-0.0027	-0.0021	-0.0024	1	-0.0021	-0.0021	-0.0044	-0.0033	-0.0021	-0.0025	-0.0012	-0.0021
file_photo file_photo file_photo	-0.0021	-0.0037	-0.0036	-0.0047	-0.0037	-0.0042	-0.0021	1	-0.0037	-0.0076	-0.0056	-0.0037	-0.0043	-0.0021	-0.0037
got new job	-0.0021	-0.0037	-0.0036	-0.0047	-0.0037	-0.0042	-0.0021	-0.0037	1	-0.0076	-0.0056	-0.0037	-0.0043	-0.0021	-0.0037
ll let know	-0.0044	-0.0076	-0.0076	-0.0097	0.081	-0.0087	-0.0044	-0.0076	-0.0076	1	-0.012	0.081	-0.0088	-0.0044	-0.0076
ok let know	-0.0033	-0.0056	-0.0056	-0.0072	-0.0056	-0.0065	-0.0033	-0.0056	-0.0056	-0.012	1	-0.0056	-0.0065	-0.0033	-0.0056
ok ll wait	-0.0021	-0.0037	-0.0036	-0.0047	-0.0036	-0.0042	-0.0021	-0.0037	-0.0037	0.081	-0.0056	1	-0.0042	-0.0021	-0.0037
really looking forward	-0.0025	-0.0043	-0.0042	-0.0054	-0.0042	-0.0049	-0.0025	-0.0043	-0.0043	-0.0088	-0.0065	-0.0042	1	-0.0025	-0.0043
tom file_photo tom	-0.0012	-0.0021	-0.0021	-0.0027	-0.0021	-0.0024	-0.0012	-0.0021	-0.0021	-0.0044	-0.0033	-0.0021	-0.0025	1	-0.0021
yes good idea	-0.0021	-0.0037	-0.0036	-0.0047	-0.0037	-0.0042	-0.0021	-0.0037	-0.0037	-0.0076	-0.0056	-0.0037	-0.0043	-0.0021	1






Distribution of Dialogue Lengths in Text Set
Frequency
0
10
20
30
40
50
60
70
80
0
500
1000
1500
2000
2500
Dialogue Length


Distribution of Summary Lengths in Text Set
Frequency
0
10
20
30
40
50
50
100
150
200
250
300
Summary Length