# Problem Statement:

Create a conversational helper that can efficiently help humans in keeping the conversation running. Since the topic itself is too broad in the aspects of conversation as humans can communicate upon various topics, for master’s thesis, we narrow the conversation down to a specific topic. For this, I have obtained the following datasets upon which we can perform fluent conversations using a dialogue type dataset.

# Datasets:

1. Cornell Movie Dialog Corpus:
   1. A dataset consisting of 502 English dialogs with 12,000 annotated utterances between a user and an assistant discussing movie preferences in natural language.
   2. <http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html>
2. Google Coached Conversational Preference Elicitation
   1. A dataset consisting of 502 English dialogs with 12,000 annotated utterances between a user and an assistant discussing movie preferences in natural language.
   2. <https://research.google/tools/datasets/coached-conversational-preference-elicitation/>
3. Facebook Empathetic Dialogues
   1. Dataset of 25,000 conversations grounded in emotional situations used for producing empathetic responses.
   2. <https://github.com/facebookresearch/EmpatheticDialogues>
4. Dataset of Multi-Domain Dialogs for the Fast Adaptation of Conversation Models
   1. This dataset is particularly suited for meta-learning dialog models or fine-tuning models with transfer-learning approaches that help develop models capable of predicting user responses in unseen domains
   2. <https://www.microsoft.com/en-us/research/project/metalwoz/>
5. NLPL OpenSubtitle corpus
   1. Dataset of multi-lingual dialogs from movie scripts. Includes 62 languages.
   2. <http://opus.nlpl.eu/OpenSubtitles2018.php>
6. Persona Chat ConvAI2
   1. Non goal oriented chatbot dataset that consists of 164,356 utterances in over 10,981 dialogs.
   2. <http://convai.io/#personachat-convai2-dataset>
7. Stanford Collaborative Communicating Agents (CoCoA): Mutual Friends
   1. Two people communicate and find a mutual friend
   2. <https://github.com/stanfordnlp/cocoa>
8. Google Meena: Human like chatbot for open domain
   1. Create an open domain chatbot that can help in conversating freely. The paper is mentioned here called, “Towards a Human-like Open-Domain Chatbot”. There were two datasets used, OpenSubtitle and closed source one.
   2. https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html
9. Bonus list of datasets
   1. Found a list of 250 NLP Datasets that are sorted according to the goal.
   2. <https://quantumstat.com/dataset/dataset.html>

# Papers:

1. Open-Domain Knowledge and Image Grounded Conversational Agents
   1. Set of 12 tasks that measures if a conversational agent can communicate engagingly with personality and empathy, ask questions, answer questions by utilizing knowledge resources, discuss topics and situations, and perceive and converse about images.
   2. The datasets used are: ConvAI2, DailyDialog, Wiz. of Wikipedia Empathetic Dialog Cornell Movie LIGHT, ELI5, Ubuntu, Twitter pushshift.io Reddit Image Chat, and IGC.
   3. <https://arxiv.org/pdf/1911.03768.pdf>
2. All-in-One Image-Grounded Conversational Agents
   1. Combine images and words to form a great conversation starter using dialogue-based images. This can be used for creating conversations that also include emotions.
   2. <https://arxiv.org/pdf/1912.12394.pdf>
3. An Ensemble Model with Ranking for Social Dialogue
   1. Create a chatbot that can chat fluently and engagingly for 20 mins and find out which utterance fits best in a given context.
   2. <https://arxiv.org/pdf/1712.07558.pdf>
4. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences
   1. Provide movie recommendations by studying the dialogues in one domain and presented a brief quantitative analysis of how people describe movie preferences.
   2. <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/54521b4011d0c2a19eaade8005ff4a499f754301.pdf>
5. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset
   1. An empathetic chatbot that can understand emotions and respond to dialogues that are geared towards general conversation or chit-chat.
   2. <https://arxiv.org/pdf/1811.00207.pdf>
6. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling
   1. Fully labeled collection of human-human written conversations spanning over multiple domains and topics. At a size of 10k dialogues, it is at least one order of magnitude larger than all previous annotated task-oriented corpora.
   2. <https://arxiv.org/pdf/1810.00278.pdf>
7. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles
   1. The release is compiled from a large database of movie and TV subtitles and includes a total of 1689 bitexts spanning 2.6 billion sentences across 60 languages.
   2. <http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf>
8. Personalization in Goal-oriented Dialog
   1. The main goal of modeling human conversation is to create agents which can interact with people in both open-ended and goal-oriented scenarios.
   2. <https://arxiv.org/pdf/1706.07503.pdf>
   3. Data: <https://github.com/chaitjo/personalized-dialog>
9. Generative Chat Bot Implementation Using Deep Recurrent Neural Networks and Natural Language Understanding
   1. Implement a chatbot over open domain that is created using seq2seq model.
   2. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3362123>
10. The Second Conversational Intelligence Challenge (ConvAI2)
    1. The aim of our competition is therefore to establish a concrete scenario for testing chatbots that aim to engage humans and become a standard evaluation tool in order to make such systems directly comparable.
    2. <https://arxiv.org/pdf/1902.00098.pdf>
11. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings
    1. Contains a dataset of 11,000 human – human dialogues that exhibits interesting lexical, semantic, and strategic elements. A neural model with dynamic knowledge graph embeddings was created that can evolve as the dialogue progresses.
    2. <https://arxiv.org/pdf/1704.07130.pdf>
12. Towards a Human-like Open-Domain Chatbot.
    1. Create an open domain chatbot that can help in conversating freely. There were two datasets used, OpenSubtitle and closed source one. This is also called as Google Project Meena.
    2. <https://arxiv.org/pdf/2001.09977.pdf>