**Conversica - Research Statement**

***Product Overview:*** *Conversica AI engages with the leads/customers who are interested in buying a product, land or share, depends on the sector in which AI engages. AI changes this interest to intent through a conversation by email/chat. These intended leads are forwarded to human sales rep for further process in buying a product.*

***Problem Statement:***

***Use Case:*** *In any product marketing sector, generating leads or prospects and converting these leads to purchase orders is a biggest challenge. Our AI must handle the conversation by giving out statistics or why the specific product is good compared to other in the industry. For example, let’s consider convincing a client to buy a share – One of the question that we could ask as a lead is why this share? Our AI has to efficiently handle this question and give out comparative analysis like the share has been giving profits continuously, can expect long time growth etc. Give out that number. Similarly, we can follow the same approach in every industry. Potential questions must be analyzed before in hand about the product while communicating with the client.*

***Dataset:*** *Actual email/chat conversation history between sales/marketing representative with the leads. Classifying the conversation that lead to buying product, being positive and the one that failed, being negative. In this domain, text preprocessing is a challenging process where punctuations are removed [punctuations matter but considering it will make as a word and this could be handles in separate model], lemmatization has to be carried out but grouped together in tense wise, striping white space, sparse words etc.*

***Approach:***  *Let us discuss the solution for the use case mentioned above. The stages mentioned here would be the same as in every other NLP algorithm but the difference being in layers, hyper parameter tuning, methods considered.*

***Padding Sequences:*** *Let us pad dataset to represent each question/answer of same length to set the vocabulary.*

***Tokenization:*** *Each word to be mapped to a token*

***Embedding Layer:*** *Embedding is performed to present words in lower dimensions. Features can be extracted by converting word to vector of questions and answers using word2vec – using CBOW [continuous bag of words]. This helps to predict each word given the surrounding words, maximizing the average log probability. This layer can be designed using SoftMax regression and negative sampling technique. This layer will be helpful to predict the semantic similarity among words.*

***Tree Structure:*** *Creating dependency parse tree as described in De Marneffe et al., 2006. The dependency tree for a question is mapped to the answer dependency tree holding its embedding structure. Using nonlinear activation function such as tanh we can create RNN function to derive complete answer by tracing to the head of the tree. Each question and answer will define in their probability space with their weights that could alter based on their occurrence or relativity with the whole dataset.*

***Model 1:*** *Question vector and answer vector is sent to the model to for the network as mentioned in the above tree structure. In addition to this each epoch is modified to approximate correct answers to wrong answers weighted approximate rank pairwise (warp) loss proposed in Weston et al. (2011). Through this we could find correct answers to for each question in the test.*

***Model 2:*** *Classification task could be performed training a question with an answer as feature matrix giving out whether this pair lead to buying a product or not as mentioned in Minaee et al. (2017). Regression task could be performed to compute a factor for an answer that finds to be a best reply for a question. Threshold factor could be set to be displayed as output.*

***References:***

1. *S Minaee, Z Liu, "Automatic Question-Answering Using A Deep Similarity Neural Network", arXiv preprint arXiv:1708.01713, 2017.*
2. *M Iyyer, JL Boyd-Graber, LMB Claudino, R Socher, IH Daume, “A Neural Network for Factoid Question Answering over Paragraphs”, In Conference on Empirical Methods on Natural Language Processing, 2014.*
3. *Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. "Generating typed dependency parses from phrase structure parses". In LREC, 2006.*
4. *Jason Weston, Samy Bengio, and Nicolas Usunier, Wsabie: Scaling up to large vocabulary image annotation, In IJCAI, 2011.*
5. *J Weston, S Chopra, A Bordes, “Memory networks”. arXiv preprint arXiv:1410.3916, 2014.*
6. *M Feng, B Xiang, MR Glass, L Wang, B Zhou, “Applying deep learning to answer selection: A study and an open task”, IEEE Workshop on Automatic Speech Recognition and Understanding, 2015.*
7. *M Tan, B Xiang, B Zhou, “LSTM-based Deep Learning Models for non-factoid answer selection”, arXiv preprint arXiv:1511.04108, 2015.*

***Question:***  *How leads are identified in the first place? How our AI identifies these candidates shows interest?*

***Papers/publication:*** *The paper which I have contributed to drive it to compeletion.*

* *A Deep Learning Approach for Predicting Early Hospital Readmission of CHF Patients.*
  + *Used LSTM RNN [Long Short-Term Memory Recurrent Neural Networks], which takes features values in sequence to predict the possibility of readmission of patients*
  + *P. Grandhe, C.P. Hon, M. De Cock, A Deep Learning Approach for Predicting Early Hospital Readmission of CHF Patients, under review*