**Project Proposal**

**Team Members:**

|  |  |  |
| --- | --- | --- |
| **Member Name** | **NetID** | **Email** |
| Wenxi Fan (Captain) | wenxif2 | wenxif2@illinois.edu |
| Abhishek Jain | aj26 | aj26@illinois.edu |

**Competition to join:** Text Classification

**Language to use:** Python

Our team's key motivation to join the **Text Classification Competition** is to go beyond what has been taught in the course, study state-of-the-art NLP tools and apply it on a real-world problem. This competition gives us such an opportunity where we cannot just learn the latest tools by solving a real-world problem but also it challenges us to raise our bars to implement optimal solutions to meet the baseline scores. We are extremely excited to learn, participate and give our best performance.

In our current professional engagements, both of us are part of AI teams respectively where machine learning tools & techniques are used for problem solving. Neural networks being the most advanced & extensively used methodologies in our respective enterprises, we both have got a chance to work and closely observe best practices & implementation styles using it. We have high level understanding and hands on experience of ANN, RNN, CNN & LSTM architecture styles and its realization using TensorFlow Keras & PyTorch based implementation. For text related problems we have awareness on usage of word embeddings viz. Word2Vec & GloVe. With the latest developments in NLP, we are aware of the release of newer advanced tools like BERT, OpenAI GPT1/2/3, HuggingFace Transformers which we have not yet explored. These tools are also in our consideration list for approaching the problem solution.

Wenxi was engaged in building RNN models for contract terms cleaning and classification, and in building the Seq2Seq model for document cleaning and document state mapping.

Abhishek has experience of working closely with data science teams which utilizes neural networks-based framework implementation to solve several problems like task sequence prediction, call conversation compliance, sales forecasting, customer look-a-likes and segmentation in customer relationship management domain.

For this project, we will explore the sequential text features, and learn the associations between words in the short text. We will plan to explore the models that combine the text ranking algorithms and machine learning models (e.g. TF-IDF and SVM). Then, we will plan to build models with Recurrent Neural Networks, or Convolutional Neural Networks or a combination of both and learn to improve the model performance with different architectures. We also plan to explore attention-based models and leverage attention models/hierarchical attention models to explore the performance impact. We will compare the different models and select the model with the best performance.