**Reddit Insights Chatbot with RAG  
MSDS692 – Data Science Practicum 1**

**Jonish Bishwakarma**

**Progress Report for Week5**

**Project Details**This project aims to build a research tool that uses Reddit discussions to surface industry-specific insights, particularly about commonly used software and related pain points in law firms, construction, and tech. Posts and comments are scraped from selected subreddits, cleaned, and analyzed for software/tool mentions. A Retrieval-Augmented Generation (RAG) chatbot will later be developed so users can query insights conversationally. The project also compares RAG performance with an LLM-only baseline.

**Project Timeline:**

Week 1 – Project definition and submit proposal (DONE)

Week 2 – Data collection (DONE)

Week 3 – Data cleaning, preprocessing, and initial keyword filtering (DONE)

Week 4 – Build software dictionary and extract mentions (DONE)

Week 5 – Sentiment analysis and identification of pain points (DONE)

Week 6 – Develop and test RAG + LLM pipelines (In Progress)

Week 7 – Evaluation of RAG vs LLM and refinement (Pending)

Week 8 – Final analysis, presentation preparation, and submission (Pending)

**Planned Work for the Week:**  
This week I planned to run sentiment analysis on the cleaned dataset and focus on identifying pain points associated with frequently mentioned tools. The goal was to distinguish between positive and negative mentions and extract common issues from negative discussions.

**Progress for the Week:**

* Implemented sentiment analysis using VADER to label posts/comments as positive, negative, or neutral.
* Tokenized text, removed stopwords, and prepared cleaned tokens for pain point analysis.
* Extracted top negative unigrams and bigrams to highlight recurring issues.
* Generated visualizations for sentiment distribution and most frequent pain-point terms.
* Observed that Linux, Windows, firewalls, and connectivity issues (e.g., “invalid user”, “connection closed”) are frequent sources of negative sentiment.

**Roadblocks/Issues:**

* Dataset size after filtering is relatively small (~700 mentions), which limits analysis of less common tools.
* Some noise still appears in tokens (e.g., “nan” or vague terms), though stopword cleanup reduced this significantly.
* Pain points often require context (e.g., “doesn’t work” could mean different things depending on the tool). Further contextual grouping may be needed.

**Plan for Next Week:**

* Begin developing the RAG + LLM pipeline, including chunking, embeddings, and retrieval setup.
* Prepare a small evaluation dataset of questions/answers for testing retrieval accuracy.
* Start preliminary comparisons of RAG vs direct LLM responses.

**Resources for the Week:**  
No new external resources beyond existing libraries (NLTK, VADER, pandas, seaborn). Output included clean sentiment labels, pain point frequency tables, and visualization plots.