METCS777 Term Paper Proposal

Scalable Emotion Classification for Affective Computing Using Apache Spark

Devin Dyson, Madhur Deep Jain

Abstract

As the volume of text-based communication continues to grow across digital platforms, analyzing emotional tone from large corpora has become essential for applications in affective computing. This project explores scalable emotion classification using Apache Spark, focusing on distributed data processing for emotion detection in text at scale. The motivation arises from the increasing need to handle large, unstructured datasets efficiently while capturing the nuanced emotional signals embedded in language.

Our research will develop and compare two emotion classification approaches: (1) a lexicon-based model using the NRC Emotion Lexicon (EmoLex) for rule-based emotion mapping, and (2) a machine learning model trained on the GoEmotions dataset using Spark MLlib with TF-IDF features and logistic regression. The framework will leverage Spark's distributed computing capabilities to process large volumes of textual data efficiently.

For the demo, we plan to showcase sample classification outputs from both models and illustrate the model training process using Spark. These demonstrations will highlight the pipeline's ability to process and classify emotions in large datasets, bridging large data analytics with affective computing. The expected outcome is a modular and interpretable prototype that demonstrates how distributed computing frameworks like Spark can advance emotion-aware applications in social analytics, mental health informatics, and human-computer interaction.

Planned Research and Methodology

1. **Data Ingestion and Preprocessing:**

    - Load large-scale textual datasets using Spark DataFrames.
    - Perform tokenization, stopword removal, and vectorization with Spark NLP and MLlib.
2. **Model Implementation:**
    - **Lexicon-based Model:** Apply the NRC Emotion Lexicon for rule-based emotion categorization.
    - **Machine Learning Model:** Train and evaluate supervised classifiers (e.g., logistic regression, Naïve Bayes) using the GoEmotions dataset.
3. **Evaluation:**
    - Compare classification accuracy, precision, recall, and F1-score between the lexicon and ML approaches.
    - Assess interpretability and robustness across different text sources.
4. **Demo Plan:**
    - Display representative examples of model predictions on unseen text samples.
    - Demonstrate the model training and inference process in Spark, highlighting the distributed workflow.

Datasets

- **GoEmotions** – Google's fine-grained emotion dataset with 27 categories. https://github.com/google-research/google-research/tree/master/goemotions
- **AffectNet** – Large-scale dataset of facial and textual emotion annotations (pending professor request for access). https://www.mohammadmahoor.com/pages/databases/affectnet/
- **EmoLex (NRC Emotion Lexicon)** – Lexicon mapping words to emotions and polarity. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

References

- Picard, R. W. (1997). *Affective Computing*. MIT Press.
- Mohammad, S., & Turney, P. (2013). NRC Emotion Lexicon.
- Demszky, D., et al. (2020). *GoEmotions: A Dataset of Fine-Grained Emotions.*
- Zaharia, M. et al. (2016). *Apache Spark: A Unified Engine for Big Data Processing. Communications of the* ACM, 59(11), 56–65.
- Feldman, R. (2013). *Techniques and Applications for Sentiment Analysis. Communications of the* ACM, 56(4), 82–89.