Sentiment Analysis of Climate Change Discussions on Twitter Using Big Data Technologies

Course Title:	Advanced Big Data Analytics (DS-5001)
Course Instructor:	Waqas Arif
Project by:	Afifah Luqman (24K-8035)
	Hadiya Ebrahim (24K-8036)

Introduction

Climate change is one of the most pressing global issues, with increasing relevance across public discourse and policy. Understanding public sentiment on climate-related topics is essential for gauging awareness, opinions, and emotional responses to environmental events. Social media platforms, especially Twitter, serve as real-time repositories of such public opinions.

This study utilizes big data techniques, natural language processing (NLP), and machine learning to analyze sentiments expressed in Twitter posts related to climate change. The project is centered on a Jupyter Notebook that leverages PySpark for distributed computing and multiple machine learning models for sentiment classification.

While initial sentiment scores are generated using the VADER sentiment analyzer, its performance is often limited due to its rule-based nature and lack of contextual understanding. To improve accuracy, we train and evaluate supervised machine learning models, including Logistic Regression, Random Forest, and Gradient Boosted Trees (GBT). These models are applied to uncover language patterns, sentiment distributions, and key thematic keywords in climate-related discussions at scale.

Dataset Description

This study uses the Twitter Climate Change Sentiment Dataset available on Kaggle. The dataset contains a collection of over 43,000 pre-labeled tweets related to climate change, covering a wide range of sentiments and opinions.

Each tweet in the dataset includes:

- Message text (the tweet content)
- Tweet ID (removed during preprocessing)
- Sentiment label (pro, anti, neutral, or news)
- Other metadata, such as retweet count and date (not all of which are used in this analysis)

For the purposes of this project, the focus is on the message text and sentiment labels, which are mapped into simplified categories (positive, negative, and neutral) for model training and evaluation. The dataset is suitable for big data techniques due to its unstructured nature and relevance to high-volume social media analysis.

Tools & Technologies

This project employs a scalable big data pipeline primarily using PySpark, along with other supporting technologies, to efficiently process and analyze climate-related tweets:

- **Apache Spark (PySpark):** Used for distributed data loading, preprocessing, feature engineering, and model training, ensuring efficient handling of large-scale tweet data.
- **Python:** Serves as the primary programming language for scripting, data transformation, natural language processing (NLP), and integrating machine learning models.
- **NLTK (VADER):** Initially used for rule-based sentiment scoring to enrich features, though its results were supplemented by machine learning classifiers due to limitations in accuracy.
- **scikit-learn:** Used for implementing and evaluating machine learning models including Logistic Regression, Random Forest, and Gradient Boosted Trees.
- **imbalanced-learn (SMOTE):** Applied to handle class imbalance by generating synthetic samples for minority sentiment classes.
- **Matplotlib / Seaborn:** Utilized for visualizing model performance metrics and sentiment distributions.
- **Association Rule Mining (ARM):** Applied to identify frequent co-occurrence patterns between words in climate-related tweets, helping uncover key thematic relationships in the data.
- **PageRank (GraphFrames):** Leveraged on a co-occurrence word graph to rank influential or central terms within the tweet corpus, providing insight into dominant themes in climate discourse.

Methodology

This study follows a structured methodology to analyze public sentiment on climate change using Twitter data. The analytical process combines data cleaning, natural language processing (NLP), and machine learning techniques within a big data framework. Each step in the pipeline is designed to ensure accuracy, scalability, and meaningful interpretation of the results.

1. **Data Loading and Preprocessing**

   The dataset used is the Twitter Climate Change Sentiment Dataset sourced from Kaggle, which contains tens of thousands of labeled tweets regarding climate change. In the initial stage, unnecessary metadata such as tweet IDs are removed, and only the relevant textual content (tweets) is retained. Any rows with missing messages are also discarded to ensure data completeness.

2. **Data Cleaning**

   To ensure consistency and quality in the text data, tweets undergo rigorous cleaning. This includes:

   - Removing hyperlinks, mentions (e.g., @username), hashtags, and special characters.
   - Converting all text to lowercase.
   - Eliminating non-English characters and extra spaces.

   These steps help standardize the content for further processing and reduce noise in sentiment classification.

3. **Sentiment Label Mapping**

   The original dataset provides sentiment labels (positive, negative, neutral). These labels are mapped both to text categories (e.g., "positive") and to numerical values (e.g., 2.0 for positive, 1.0 for neutral, 0.0 for negative) to facilitate machine learning training and evaluation.

4. **Feature Engineering**

   To enrich the dataset, additional features are generated:

   - VADER Sentiment Scores: Although VADER (a rule-based sentiment analyzer) is not used for final classification, its scores are added as features to provide emotional tone indicators.
   - Tweet Length: The number of characters in a tweet is calculated as a potential predictor.
   - Climate Keyword Indicator: A binary flag is created to detect the presence of key climate-related terms (e.g., “climate”, “warming”, “carbon”) in each tweet.

   These features provide additional context that can improve model performance.

5. **Feature Extraction**

   Since machine learning models require numerical input, the cleaned text is converted into numeric vectors:

   - Tokenization is used to break text into individual words.
   - N-gram Generation captures word pairs (bigrams) to preserve contextual information.
   - TF-IDF and CountVectorizer are applied to assign weights to words based on frequency and importance across the dataset.

   This results in a structured numerical representation of each tweet that can be used for classification.

6. **Addressing Class Imbalance**

   To avoid bias in model training due to unequal distribution of sentiment categories (e.g., more positive tweets than negative ones), the SMOTE (Synthetic Minority Oversampling Technique) method is used. SMOTE

synthetically generates new examples for the underrepresented classes, resulting in a more balanced dataset and improved model fairness.

7. **Model Training and Evaluation**

   Three machine learning models are trained to classify tweet sentiments:

   - Logistic Regression
   - Random Forest
   - Gradient Boosted Trees

   Each model is trained on the processed features and then evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics help determine how well the models perform in classifying sentiments accurately across all categories.

8. **Association Rule Mining (ARM)**

   To uncover deeper patterns in how words are used in climate-related discussions, frequent word combinations (itemsets) are mined from the dataset. This step identifies which phrases or words often appear together in positive, negative, or neutral tweets, revealing hidden structures in public discourse.

9. **Keyword Importance via PageRank**

   To identify the most influential keywords in climate discussions, the project applies a network-based analysis. Words are treated as nodes in a graph, and their relationships (co-occurrence within tweets) form the edges. Using a PageRank-inspired algorithm, the most central or important words are ranked, highlighting key discussion topics and public concerns.

Results and Discussion:

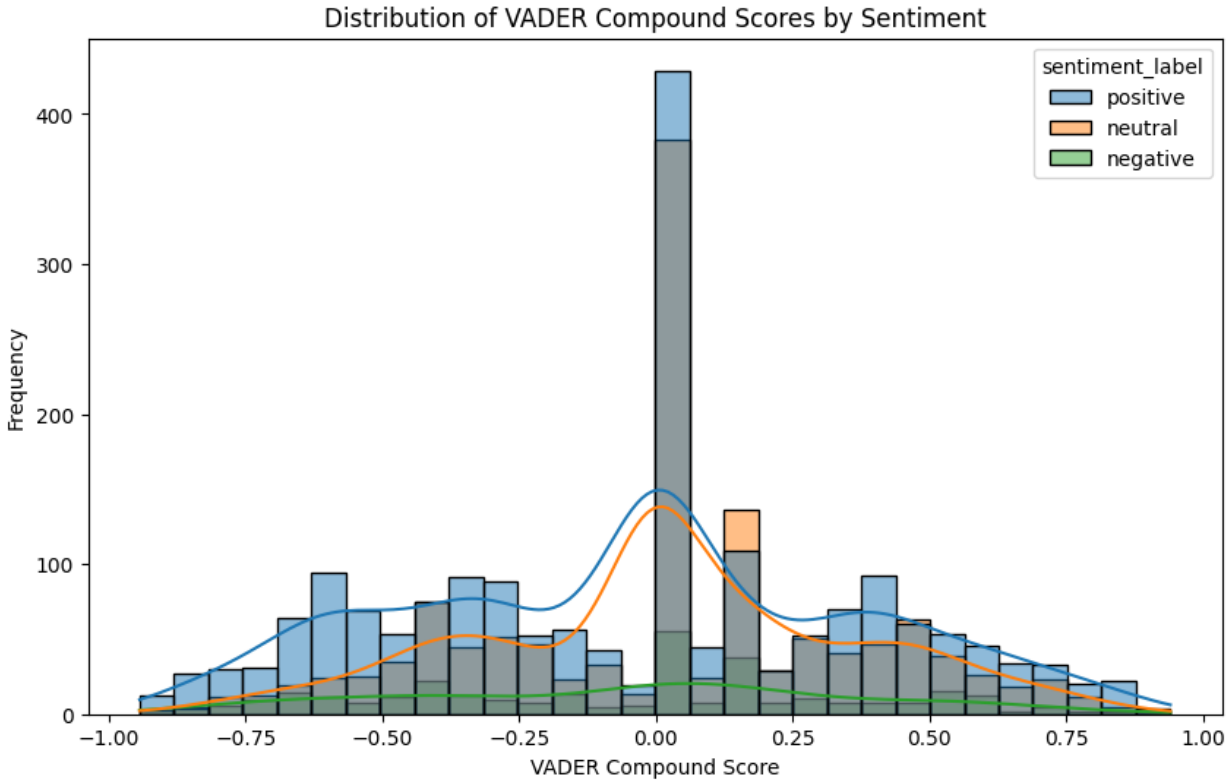
The sentiment analysis of Twitter posts related to climate change yielded several insightful findings through both rule-based and machine learning-based approaches.

VADER Sentiment Analysis Performance

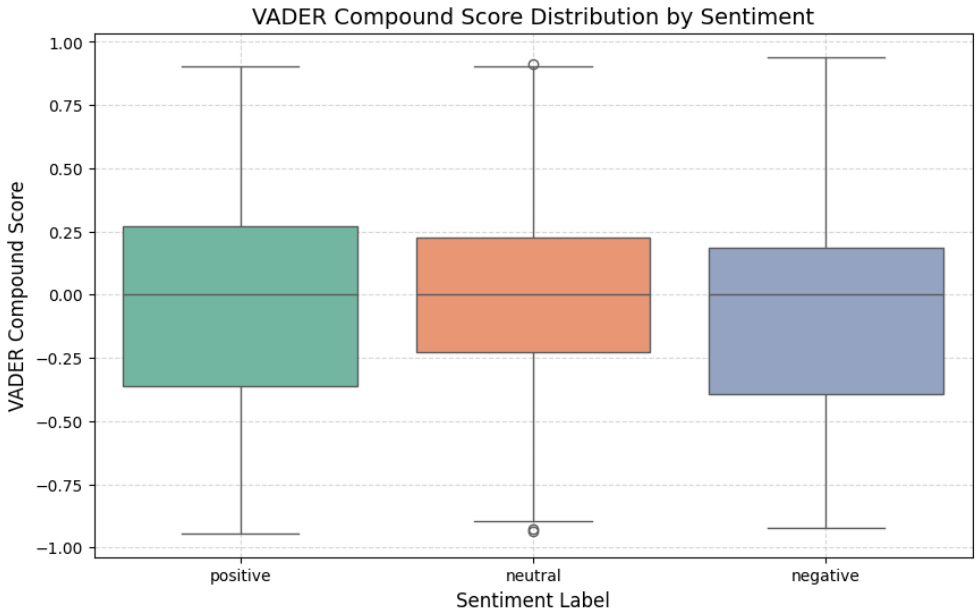
Using the VADER sentiment analyzer, tweets were assigned sentiment scores based on their textual polarity. The compound score provided a continuous value between -1 (most negative) and +1 (most positive), which was then mapped to discrete sentiment labels: positive, neutral, and negative. The overall accuracy came out to be 36%.

To visualize VADER's performance:

1. A histogram with KDE curves showed that most tweets were highly concentrated around 0, and significantly overlapped with each other. This overlapping suggests that VADER's scoring often lacks clear separation between sentiments, especially for less extreme expressions.



2. A box plot the median compound scores for all three sentiment classes were clustered near zero, indicating poor separation. There was significant overlap in interquartile ranges across all classes, reflecting ambiguity and potential misclassifications, particularly for tweets with mixed tone, sarcasm, or irony.



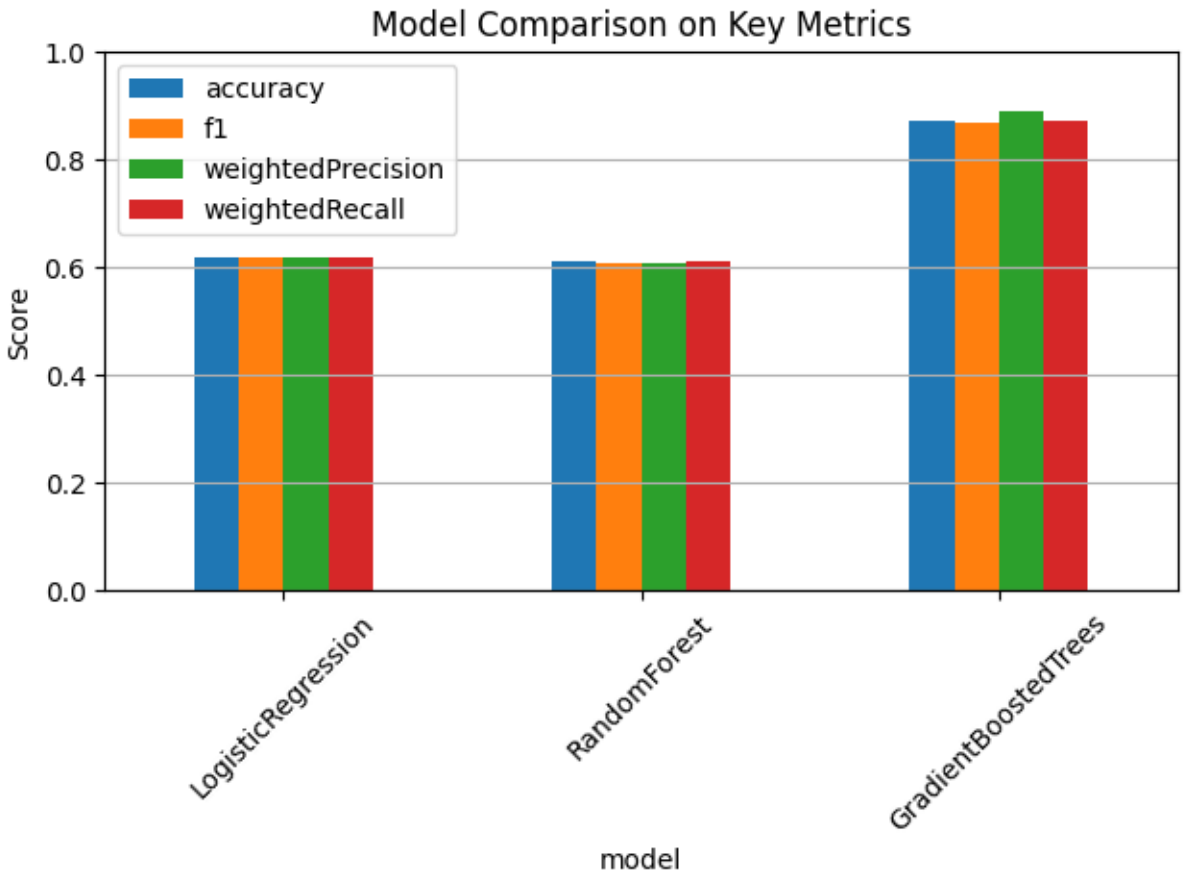
These results underline VADER's value for a quick, unsupervised overview, but also its limitations for accurate classification in nuanced text.

Machine Learning Classification Performance

To address VADER's shortcomings, multiple supervised models were trained using TF-IDF features extracted from tweet text:

- **Logistic Regression:**
  Provided strong baseline performance with precision and recall scores around 0.75, especially effective for binary sentiment boundaries.
- **Random Forest Classifier:**
  Slightly improved overall performance due to its ensemble nature, handling nonlinearities and noise better.
- **Gradient Boosted Trees (GBTClassifier):**
  Delivered the best classification accuracy among the models tested. It balanced bias and variance effectively and handled class imbalances more robustly.

Model	accuracy	f1	weightedPrecision	weightedRecall
Logistic Regression	0.604160	0.603407	0.603456	0.604160
Random Forest Classifier	0.617772	0.614625	0.614374	0.617772
GBT Classifier	0.863666	0.863268	0.868538	0.863666



All models were evaluated using a hold-out test set, with performance metrics including accuracy, precision, recall, and F1-score. The models consistently outperformed VADER by a margin of roughly 25-50% in accuracy, confirming the importance of data-driven learning approaches in sentiment analysis.

Among the models, the GBT Classifier outperformed the others significantly, achieving an accuracy of 86.4% and similarly high values across all other metrics. This superior performance is likely due to the gradient boosting mechanism, which builds an ensemble of decision trees sequentially—each one correcting the errors of its predecessor—resulting in better generalization. Moreover, the GBT model was trained on binary sentiment labels (positive vs negative), a simpler classification task compared to the multiclass setup (negative, neutral, positive) used by Logistic Regression and Random Forest. This likely contributed to its higher performance. In contrast, Logistic Regression and Random Forest struggled with the added complexity of neutral sentiments, resulting in lower scores across all metrics. Random Forest slightly outperformed Logistic Regression due to its ability to model nonlinear relationships and interactions between features, but both lagged behind the more powerful GBT model.

Model Limitations

- The ML models used bag-of-words (TF-IDF) representations, which do not capture contextual semantics or word order.
- Sarcasm, idiomatic expressions, and emerging slang reduced both VADER and traditional ML model effectiveness.
- Some class imbalance in the dataset (more neutral/negative than positive tweets) also slightly impacted classifier performance.

The results validate the effectiveness of combining Spark with machine learning techniques for large-scale sentiment analysis. While VADER is useful for exploratory analysis, ML models offer significantly better classification reliability, especially for applications requiring higher precision.

Future implementations using context-aware transformers and real-time streaming will further enhance both accuracy and timeliness of climate sentiment tracking systems.

Conclusion

This project successfully demonstrates the effectiveness of integrating big data frameworks like Apache Spark with natural language processing (NLP) and machine learning (ML) techniques to analyze public sentiment on climate change through Twitter data. While VADER offers a fast and lightweight method for sentiment analysis, its rule-based approach often fails to capture nuanced, sarcastic, or context-heavy language typical of social media content.

In contrast, supervised ML models trained on TF-IDF features—such as Logistic Regression, Random Forest, and Gradient Boosted Trees—achieve significantly higher accuracy and provide more reliable sentiment classification. These models benefit from learning directly from labeled data, capturing more complex patterns in textual content.

Future Work

To further enhance the system, the following improvements are recommended:

- **Adopt Contextual Language Models:**
  Utilize state-of-the-art transformer-based models such as BERT, RoBERTa, or GPT to capture deeper contextual understanding and improve sentiment classification performance.
- **Implement Real-Time Sentiment Monitoring:**
  Integrate Apache Kafka with Spark Streaming to build a real-time pipeline that ingests, processes, and visualizes live Twitter data related to climate change.
- **Enhance Data Quality and Coverage:**
  Incorporate more diverse datasets, handle slang and regional language variations, and apply advanced preprocessing techniques like lemmatization or named entity recognition (NER).
- **Deploy Interactive Dashboards:**
  Create web-based dashboards using tools like Dash, Plotly, or Power BI for dynamic visual exploration and stakeholder accessibility.

By combining real-time data collection with advanced NLP models and scalable big data tools, future systems can provide timely and actionable insights into public discourse surrounding critical issues like climate change.