**Research Brief: Sentiment Analysis for Detecting Greenwashing in Corporate Social Media Communication**

**Background**

Greenwashing refers to deceptive practices where companies exaggerate or misrepresent their environmental sustainability efforts to appeal to consumers, investors, and regulators. With increasing regulatory scrutiny and public awareness, detecting greenwashing has become a critical task for ensuring corporate accountability.

Social media platforms, such as **Twitter, LinkedIn, Facebook, and TikTok**, serve as primary channels where companies communicate their sustainability efforts. However, the responses from users can provide critical insights into whether a company's claims align with public perception or whether contradictions exist between corporate messaging and consumer sentiment.

Existing greenwashing detection frameworks focus on corporate disclosures, **Net Zero transition plans, and regulatory risk assessment**. Some approaches leverage **Natural Language Processing (NLP) models** to analyze ESG reports and corporate filings. However, there is limited work on **using social media sentiment analysis** as a tool for detecting contradictions in corporate sustainability claims.

**Scope of the Study**

This research will focus on analyzing **user sentiment and contradictions in corporate social media sustainability posts**. The goal is to assess **how social media users react to corporate claims on environmental issues** and determine **if negative sentiment or contradiction patterns correlate with potential greenwashing practices**.

**Key Research Questions**

1. What are the dominant sentiments (positive, neutral, or negative) in user responses to corporate sustainability posts?
2. How do sentiment trends vary across different social media platforms?
3. Can a **contradiction score** be developed that measures the gap between corporate claims and user responses?
4. How does sentiment analysis compare with existing greenwashing detection frameworks?

**Target Companies & Platforms**

* Companies: Selected based on industries with high greenwashing risk (e.g., energy, fashion, automotive, finance).
* Platforms: **Twitter, LinkedIn, Facebook, TikTok**.

**Potential Methodologies**

**1. Data Collection and Preprocessing**

This study will focus on collecting and analyzing social media data to assess corporate sustainability claims and detect potential greenwashing. The dataset will consist of **corporate posts** (official statements from companies) and **user comments** (responses to these posts), covering multiple social media platforms (Twitter, LinkedIn, Facebook, TikTok).

**1.1 Data Sources and Extraction**

* **Corporate Sustainability Posts**:
  + Scraped from **official social media pages** of companies.
  + Filtering by keywords: *"net zero," "carbon neutral," "sustainable," "eco-friendly," "climate commitment,"* etc.
  + Storing timestamp, engagement metrics (likes, shares, retweets, views), and media (text, images, videos).
* **User Comments**:
  + Extracted from responses to corporate posts.
  + Comments will be collected along with metadata:
    - **Sentiment indicators**: Likes, reactions (e.g., "angry" reaction on Facebook).
    - **Engagement volume**: Number of replies, retweets, or shares.
    - **User credibility**: Verified accounts vs. bots (potential filtering step).
* **Historical Greenwashing Cases**:
  + Past **regulatory violations, lawsuits, and controversies** related to greenwashing.
  + Extracting historical social media data during the time period of these controversies.
  + Sources: **Regulatory reports, news articles, and corporate disclosures.**
* **External ESG Ratings**:
  + Cross-referencing sentiment analysis with existing **ESG risk scores** (e.g., Sustainalytics, MSCI ESG, GWI database​).

**1.2 Data Cleaning and Structuring**

* **Text Preprocessing**:
  + Remove **spam, advertisements, bots** using automated NLP-based filtering.
  + Tokenization, stopword removal, lemmatization.
  + Detect and **handle sarcasm or irony** using a sarcasm detection model.
* **Data Structuring**:
  + Organize into a structured dataset:
    - post\_id: Unique ID of corporate post.
    - company: Name of the corporation.
    - platform: Twitter, LinkedIn, Facebook, TikTok.
    - timestamp: Date of post.
    - content: Text of corporate post.
    - engagement\_metrics: Likes, shares, retweets.
    - comment\_id: Unique ID for comments.
    - sentiment\_score: Computed from NLP model.
    - contradiction\_score: Computed based on semantic comparison.

**2. Sentiment and Emotion Classification**

To analyze user reactions, we will employ **a combination of sentiment analysis and multi-class emotion classification**.

**2.1 Sentiment Analysis**

* **Lexicon-Based Models**:
  + **VADER (Valence Aware Dictionary for Sentiment Reasoning)** for Twitter.
  + **AFINN, TextBlob** for broader applications.
* **Transformer-Based Models**:
  + Fine-tune **RoBERTa or BERTweet** (for Twitter) on financial/environmental sentiment datasets.
  + Use **DistilBERT** for fast inference on large-scale comment datasets.
* **Output Classes**:
  + **Positive**: Supportive of company claims.
  + **Neutral**: No strong opinion.
  + **Negative**: Skeptical or critical of corporate sustainability efforts.

**2.2 Multi-Class Emotion Classification**

* **Why Emotion Classification?**
  + Sentiment analysis (positive/negative) is **too coarse-grained**.
  + Emotion analysis can detect **deeper public reactions** (e.g., "anger" could indicate greenwashing suspicion).
* **Model Choice**:
  + **GoEmotions (Google's 27-class emotion dataset) fine-tuned on sustainability data.**
  + **BERT-based Emotion Models (e.g., BERTweet + GoEmotions)** for social media-specific emotions.
* **Emotion Categories**:
  + **Positive Emotions**: *Trust, admiration, optimism* (corporate messaging aligns with public belief).
  + **Negative Emotions**: *Anger, disgust, disappointment* (public perceives greenwashing).
  + **Uncertainty**: *Confusion, skepticism* (signals a need for contradiction detection).

**3. Contradiction Score: Measuring Greenwashing Risk**

* **Objective**: Measure the **semantic gap** between corporate claims and user reactions.
* **Approach**:
  + Compute **semantic similarity** between corporate posts and user comments.
  + Detect **stance (agreement/disagreement)** using NLP-based **stance detection models**.

**3.1 Semantic Similarity Models**

* **Universal Sentence Encoder (USE)** for computing sentence embeddings.
* **SBERT (Sentence-BERT)** for improved context-aware comparison.
* **Metrics**:
  + Similarity Score (0-1): High similarity → User agrees with corporate claim.
  + Contradiction Score: Ratio of comments contradicting corporate claims.

**3.2 Stance Detection**

* **Why Stance Detection?**
  + Goes beyond sentiment analysis to determine **agreement/disagreement**.
* **Models**:
  + **RoBERTa-based Stance Detection** (fine-tuned on climate debate datasets).
  + **LIAR dataset models** (adapted for sustainability claims).
* **Output Labels**:
  + **Support**: User agrees with corporate claim.
  + **Neutral**: No clear stance.
  + **Oppose**: User contradicts corporate claim.

**4. Predictive Analysis and Future Risk Assessment**

**4.1 Training Models on Past Greenwashing Cases**

* Extract **historical sentiment and contradiction trends** around **past greenwashing incidents**.
* Fine-tune sentiment and contradiction models on **labeled greenwashing controversies**.
* **Data Sources**:
  + Greenwashing lawsuits and controversies​.
  + Public perception trends **before and after** greenwashing exposés.

**4.2 Predicting Future Greenwashing Risk**

* Use **historical sentiment-contradiction patterns** to train a predictive classifier.
* **Random Forest / XGBoost** to detect **companies at risk of greenwashing accusations**.
* Compare **predicted risk scores** with **real regulatory investigations**.

**5. Benchmarking Against External ESG Scores**

* Compare **sentiment-based greenwashing risk scores** with **ESG risk ratings**.
* Evaluate if **social media sentiment predicts real-world ESG controversies**.

**6. Final Deliverables**

1. **Social Media Greenwashing Index**:
   * Companies ranked by **sentiment and contradiction scores**.
   * High **contradiction + negative sentiment** = **High greenwashing risk**.
2. **Automated NLP Dashboard**:
   * Real-time tracking of corporate sustainability claims and public sentiment.
3. **Predictive Model**:
   * Identify **future greenwashing risk** based on historical patterns.

This methodology provides a **data-driven framework** for assessing corporate greenwashing through **sentiment, emotion, and contradiction analysis**. By **integrating historical data**, fine-tuning **state-of-the-art NLP models**, and leveraging **predictive analytics**, this approach aims to create **an early warning system for greenwashing risk detection**.