Fact / Claim / Opinion Extraction in Solar Energy News Articles

Vrishmi Parikh, Mahmood Topiwala, Anurag Shukla and Tanaz Pathan

Abstract

In this project, we present a model for automatically classifying solar energy news content into facts, claims, and opinions. Given the absence of domain-specific datasets for solar news, a custom dataset of 57,000 solar-specific articles was curated from an initial corpus of 103,000 renewable energy articles. An iterative active learning framework was employed to classify sentences, leveraging LLM-based annotations and manual verification to generate high-quality labels. The distilledbert-base-uncased model was the core classifier trained using confidence-based active learning. The proposed model was benchmarked against ClaimBuster and CONCORD, demonstrating superior performance in claim detection with a precision of 0.8765 and an F1 score of 0.8637, significantly surpassing ClaimBuster's scores. We further extended our work to include URL-based article processing and semantic embeddings for improved content retrieval and classification.

1. Problem Statement

The renewable energy sector, particularly solar energy, has experienced exponential growth in media coverage, generating vast amounts of content that combines factual reporting, future projections, and expert opinions. Separating facts from claims and opinions is critical for market analysis and planning in this fast-growing sector. This mixed nature of content presents unique challenges for stakeholders across the industry. Researchers require reliable factual information to support their studies, while investors need to distinguish between verified facts and speculative claims. Policymakers must separate objective data from subjective opinions, and industry analysts need to track verifiable trends versus projected developments. Market strategists require accurate differentiation between proven results and future projections.

Manual analysis of such content is both time-consuming and impractical at scale, creating a clear need for automated classification systems. Our work addresses this challenge by presenting a framework for automated classification of statements into three distinct categories:

Facts: Objectively verifiable statements, typically containing specific numerical data or reporting completed actions.

Claims: Future-oriented or unverified statements, including predictions and projections.

Opinions: Subjective assessments and personal judgments.

2. Dataset

To address the lack of domain-specific datasets for solar energy news, we created a custom dataset by scraping over 106k renewable energy articles from seven major sources: PV Mag - Global (36.28%), Saur Energy (18.87%), Mercom India (15.33%), PV Mag - USA (11.49%), Energy World - Economic Times (8.70%), PV Mag - India (8.14%), and Economic Times (1.20%). Each article includes metadata such as publication date, title, summary (if available), body, and source URL. The dataset spans from 2014 to August 2024, with an increasing trend in daily publication volume.

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202318013 (V. Parikh); 202318030 (M. Topiwala); 202318039 (A. Shukla); 202318056 (T. Pathan)

To keep the dataset current, we developed an automated pipeline that can be manually triggered to fetch new articles. For solar-specific content, we fine-tuned a **distilledbert-base-uncased** model for binary classification of article headlines. The labeling process combined LLM-based majority voting (Gemma2-9B, Llama 8B, Mistral 7B) with manual classification of 1,000 headlines for test validation. The model was trained for 40 epochs with a learning rate of 0.00001 and a classification threshold of 0.70, achieving a precision of 0.9706, recall of 0.8730, accuracy of 0.9406, and PR curve AUC of 0.9767.

The filtering process yielded 57k solar-specific articles, from which over 900k unique sentences were extracted using spaCy's sentencizer. These sentences form the basis for our claim, fact, and opinion classification task.

3. Methodology

The proposed methodology employs an active learning framework for classifying sentences into facts, claims, and opinions, enabling iterative model refinement and efficient use of labeled data while minimizing manual annotation requirements.

The data annotation and labeling process began with LLM-generated annotations. Sentences were labeled using the Llama3.1 70B Instruct API, which produced an initial set of 1,500 annotated sentences. To ensure label quality, a subset of these annotations underwent manual verification, during which 1,200 verified labels were added to the test set, while 300 sentences were incorporated into the training set to enhance the model's learning process.

For model training, the project utilized a distilledbert-base-uncased architecture, a lightweight and efficient transformer-based model. The model was trained using key hyperparameters, including 30 epochs, a learning rate of 0.000025, and macro-averaged F1 score as the primary decider metric to guide the active learning process. To ensure optimal selection of informative training samples, a query strategy was implemented, which prioritized sentences from the bottom 10th percentile of model confidence for annotation in subsequent iterations. This strategy allowed for targeted annotation of the most uncertain predictions, significantly improving the model's classification capabilities.

The active learning process followed an iterative structure. Each iteration involved sampling 200 random sentences from the pool of low-confidence predictions. These sentences were then manually annotated and added to the training set, enabling the model to learn from its most uncertain decisions. By continuously refining the model in this way, the system progressively improved its ability to classify facts, claims, and opinions with greater precision and recall.

4. Evaluation

The performance of the model was compared with two benchmarks: ClaimBuster (used for claim detection) and CONCORD (a dataset of COVID-19 numerical claims).

Solar Test Dataset Comparison:

- **Our Model**: Precision = 0.8765, Recall = 0.8513, F1 = 0.8637
- **ClaimBuster**: Precision = 0.7631, Recall = 0.6954, F1 = 0.7275

CONCORD Comparison:

- **Our Model**: Precision = 0.5712, Recall = 0.5650, F1 = 0.5680
- **ClaimBuster**: Precision = 0.6710, Recall = 0.6570, F1 = 0.6639

The evaluation demonstrates that the project's model outperforms ClaimBuster in claim detection but falls short on CONCORD, suggesting the need for better generalization to diverse datasets.

5. Key Challenges and Learnings

The dataset creation challenge included building a domain-specific dataset for solar energy, which required scraping over 100k articles and filtering them to 57k solar-specific articles using a fine-tuned classification model. While developing and implementing our model, we encountered some challenges. Though LLM annotations provided a strong starting point, they required careful verification to ensure data quality and accuracy. Defining clear guidelines to distinguish between facts, claims, and opinions was critical in maintaining consistency across the dataset. This highlighted the critical importance of expert oversight throughout the annotation process. We faced computational resource constraints that necessitated efficient model design and careful batch size management.

From a technical perspective, we learned that the dual-branch architecture combining BOW and transformer features provided better classification results. Implementing the active learning loop revealed the importance of effective query strategies to select low-confidence samples for annotation, while cross-domain testing highlighted the need for enhanced generalization capabilities. Working with FAISS and semantic embeddings (Google PaLM) for content retrieval really opened some doors to how we can actually make these advanced tools align with our project goals.

Building the Streamlit app really emphasized the importance of having interfaces that are both intuitive and interactive. Why? Because without such user-friendly designs, the practicality of applying models in real-world scenarios diminishes significantly.

6. Extension of the Project

Building upon our initial model, we expanded our project's capabilities to enhance its practical applications. We developed functionality to directly process article URLs, enabling dynamic extraction and classification of content from web sources. This extension leverages Google PaLM's semantic embeddings for improved content retrieval and classification accuracy. We integrated FAISS (Facebook AI Similarity Search) to efficiently search and retrieve articles or specific content types (facts, claims, or opinions) from our database.

The enhanced system supports both the classification of new content and contextual question-answering through advanced retrieval mechanisms. To facilitate user interaction, we developed a web interface where users can input multiple article URLs simultaneously, process their content in real time, and receive classified outputs. Additionally, users can query the processed content to extract specific types of information or ask questions about the analyzed articles, making the tool more versatile for practical applications in renewable energy market analysis. [1] [2] [3] [4]

References

[1] D. Shah, K. Shah, M. Jagani, A. Shah, B. Chaudhury, CONCORD: Numerical Claims Extracted from the COVID-19 Literature using a Weak Supervision Approach, SSRN, 2023.

[2] A. Shah, A. Hiray, P. Shah, A. Banerjee, A. Singh, D. Eidnani, S. Chava, B. Chaudhury, S. Chava, Numerical Claim Detection in Finance: A New Financial Dataset, Weak-Supervision Model, and Market Analysis, arxiv, 2024.

[3] N. Hassan, A. Nayak, V. Sable, C. Li, M. Tremayne, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, Claimbuster: the first-ever end-to-end fact-checking system, Proceedings of the VLDB Endowmen 10 (2017) 1945–1948.

[4] H. Luo, W. Tan, N. Nguyen, L. Du, Re-weighting tokens: A simple and effective active learning strategy for named entity recognition, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 12725–12734.