Machine Learning-Driven Emotional Weather Newsreading with NAO V6

by

Rebecca Holland

24/25 Critical 4: Machine Learning for Data Science and Ethical Computing

16th June 2025

Abstract

The aim is to create an emotionally expressive weather newsreader by using a sentiment classifier trained on rewritten and sentiment-labelled UK weather headlines. The NAO robot will interpret and vocalise headlines with appropriate tonal variation, LED facial changes, and gestures based on sentiment. The project explores both the conceptual and technical aspects of data wrangling, model building, evaluation, and ethical considerations in real-world ML deployment.

Contents

0.1 Concept and Scope

The NAO V6 stands out to me, mainly for its expressive capacity, compact form, and rich interaction features. The NAO robot's capabilities include text-to-speech, LED facial feedback, head and limb motion, and real-time programmability using the NAOqi SDK (Creative Computing Institute 2025). These features position NAO as ideal for performance-based social robotics, particularly for tasks that require emotional nuance and human-facing interaction.

I want to leverage these features to create an emotionally responsive weather newsreader. By feeding weather headlines scraped from The Guardian into a sentiment classifier, the system assigns each headline a sentiment label. The NAO robot then delivers the headline vocally with an appropriate pitch, facial light colour, and physical gesture. For example, negative sentiment (e.g., storms) may produce a red glow and a concerned tone, while positive sentiment (e.g., sunshine) prompts a cheerful tone and open-arm gesture.

The scope includes scraping and cleaning the dataset, training and evaluating supervised machine learning models, and integrating the trained model into a real-time robotic pipeline.

0.2 The Role and Impact of Data in Machine Learning

In this project, data is a core asset that defines the quality and credibility of the robot's output. The original dataset was sourced by scraping weather headlines from The Guardian's online platform. While this offered an authentic representation of real-world language, the headlines were often ambiguous or contained inconsistent emotional cues.

Collecting and preparing data required decisions about ethical reuse and modification. Since the headlines were publicly accessible, and their use was for educational, non-commercial purposes, the data handling complied with fair use principles. However, the headlines were not presented verbatim in the final implementation. Instead, they were rephrased and categorised into a structured form, enhancing clarity for both human interpretation and model training (Žliobaitė 2017).

In short, the impact of data in this project extends beyond training. It shapes the robot's personality, dictates the boundaries of its responses, and affects user perception. A well-structured dataset not only improves model performance but ensures that the robot's behaviours are consistent, appropriate, and engaging.

0.3 Supervised vs Unsupervised Machine Learning in Context

Machine learning can generally be divided into supervised and unsupervised paradigms. Understanding the distinctions between these approaches is critical to justifying model selection within this project.

In supervised learning, a model is trained on a labelled dataset where each input is paired with a known output. In this project, the inputs are rewritten weather headlines, and the outputs are sentiment categories: positive, neutral, or negative. The model learns to associate specific language patterns with sentiment classes, allowing it to make accurate predictions on new, unseen data. Supervised learning offers the advantage of interpretability, consistent performance, and quantitative evaluation using metrics like accuracy, precision, recall, and F1-score (Pedregosa et al. 2011).

In contrast, unsupervised learning involves models trained on unlabelled data. The system attempts to infer hidden patterns or groupings in the data without explicit guidance. In this project, early implementations using VADER and TextBlob are examples of unsupervised or lexicon-based sentiment estimation (Hutto & Gilbert 2014, Loria 2018). These tools rely on pre-compiled dictionaries of words with associated polarity values, often missing context or misclassifying emotionally ambiguous phrases.

The decision to adopt supervised learning enhanced the model's reliability, enabled precise evaluation, and supported the project's overarching goal: mapping emotional tone accurately onto a robotic interface. This approach also allowed for more effective training iterations, where specific failure modes (e.g., consistently misclassified neutral headlines) could be addressed directly through dataset refinement and feature engineering.

0.4 Areas for Improvement and Plan

I performed initial experiments using VADER and TextBlob, unsupervised sentiment tools, to classify headlines. These approaches consistently mislabelled headlines due to context blindness. For example, storm-related warnings were sometimes identified as neutral or even positive. Misclassification's like these were not acceptable for a robot delivering public-facing messages. These lexicon-based approaches struggled with contextual ambiguity and failed to correctly classify weather-specific emotional tone. Logistic regression was chosen for its significantly higher performance, achieving robust classification accuracy and providing the flexibility to iterate on feature engineering and retrain as needed.

To improve performance, a curated dataset of 137 UK weather headlines was constructed with clearly defined sentiment categories. These headlines were rewritten to better reflect real-world emotional tone and classified into positive, neutral, or negative sentiment using structured criteria. The use of supervised learning, specifically logistic regression, allowed the model to learn contextual sentiment patterns from this improved dataset.

I performed further enhancements included rebalancing the dataset to ensure class equality, conducting exploratory data analysis to understand feature distribution, and applying evaluation metrics such as confusion matrices to validate performance. These steps dramatically improved accuracy and consistency, especially in edge cases involving ambiguous language.

In future stages, the project could explore transformer-based models such as BERT for deeper language understanding and generalisation. Transfer learning could be applied using pre-trained models on broader news datasets, followed by fine-tuning on weather-specific data. Reinforcement learning could also enhance NAO's delivery by learning from user interactions and adjusting its tone and gesture dynamically based on user feedback.

0.5 Machine Learning and Data Science Foundations

0.5.1 Conceptual Framework

The project employs both foundational and applied machine learning. Initially, unsupervised approaches such as lexicon-based VADER were tested. These failed to capture the contextual nature of weather language. The project pivoted to supervised learning using manually labelled data, allowing more accurate sentiment classification.

The chosen machine learning approach was supervised learning, specifically using a logistic regression classifier. Logistic regression is well-suited for sentiment classification tasks due to its efficiency, interpretability, and ability to model probabilistic boundaries between multiple sentiment classes. It provided a balance between performance and simplicity, making it ideal for deployment within a real-time robotic system such as NAO.

A pipeline was created using CountVectorizer to convert text into feature vectors, followed by the logistic regression classifier to learn sentiment boundaries.

0.5.2 Data Science Techniques

The data science process included:

- Scraping UK weather headlines using a custom Python script
- Cleaning and rewriting for clarity and sentiment diversity
- Exploratory data analysis using Pandas and matplotlib to assess balance
- Feature engineering via Bag-of-Words vectorisation
- Model evaluation using precision, recall, F1-score, and confusion matrix

This structured approach enabled high model performance and strong alignment with the goals of reproducible and interpretable ML.

0.6 ML Pipeline and Data Flow

0.6.1 Conceptual Flow

1. Data Capture: Headlines are scraped and stored as CSV
2. Preprocessing: Cleaned and annotated
3. Analysis: Visualised for class distribution and imbalance
4. Model Training: Supervised learning using logistic regression
5. Inference: Real-time headline sentiment prediction
6. Mapping: Sentiment output controls NAO's speech, gesture, LEDs

0.6.2 Implementation Flow

In practice, the data is loaded from a CSV and preprocessed using standard Python libraries. The dataset is split into training and test sets. A pipeline is created to vectorise and classify the data. Once trained, the model is saved and used in a live script that accepts input headlines.

When run in a real-time environment, the system classifies incoming headlines and logs them to a timestamped file (robot.log). NAO then uses NAOqi modules to deliver speech and expressive feedback based on the sentiment.

The data pipeline was constructed to operate end-to-end. It began with loading CSV-based headline data, followed by preprocessing for consistency. This was passed through a machine learning pipeline comprising CountVectorizer for Bag-of-Words vectorisation and logistic regression for classification. The trained model was saved using joblib and reloaded during runtime to make predictions on user-provided headlines. Sentiment predictions were then used to drive real-time multimodal behaviours in NAO, including voice pitch, facial LEDs, and posture adjustments. Logs were maintained for each interaction, supporting future analysis.

0.7 Data from the Robot and Analysis

Although NAO V6 does not produce traditional sensor data in this implementation, it does log timestamped interactions. Each entry includes the headline, predicted sentiment, and resulting robot behaviour. These logs can be analysed to understand classifier behaviour over time.

This logging supports iterative development. Early logs revealed misclassifications (e.g., treating storm warnings as neutral), prompting improvements in dataset design and model choice. In later stages, these logs validate that the robot responds appropriately to inputs, providing critical feedback for model evaluation.

The project incorporated several layers of data analysis and interpretation. Exploratory data analysis was used to assess sentiment distribution and detect imbalances in the original dataset. Inference was carried out by the trained model in response to real-time user input, and the robot's logged reactions were used to validate prediction consistency. By reviewing the robot.log output, patterns in sentiment classification were interpreted to guide dataset refinements and enhance model robustness, particularly in distinguishing neutral cases. This iterative cycle of analysis and interpretation was central to improving both the model and the user experience.

0.8 Evaluation and Results

Two models were compared:

- VADER (unsupervised): Fast but poor performance on context-dependent text
- Logistic Regression (supervised): Higher accuracy and consistency

The final classifier was evaluated using:

- Accuracy: 87
- F1-Score: Average 0.85
- Confusion Matrix: Clear sentiment class boundaries, minor confusion in neutral cases

0.9 Ethical and Critical Considerations

Ethical design was central to this project. One concern was the potential for sentiment bias, particularly when using headlines that may contain emotionally charged or ambiguous language. Although labelling mitigated some issues, the act of annotating inherently introduces subjectivity. This highlights the challenge of defining emotional categories within text. For example, one may determine 'Snowy Weather' as positive, but others may suggest that it is negative.

There is also a responsibility in presenting weather-related content with appropriate seriousness. For example, flooding or extreme storms should not be delivered with humorous or light-hearted tones. Care was taken to ensure that sentiment mappings were appropriate, and this was part of the evaluation criteria during model training.

Accessibility and inclusion were core design principles. By combining visual (LEDs), auditory (speech tone), and kinetic (gesture) feedback, the system aims to communicate clearly across a range of user needs and contexts.

This project offers potential value for individuals with learning disabilities, who may struggle to interpret the abstract implications of weather reports. By translating complex or emotionally subtle information into a multi-sensory delivery—using tone, colour, and motion—the NAO robot can help these users better understand the gravity or positivity of weather conditions. This sensory-rich approach supports cognitive accessibility and fosters greater independence in daily planning based on environmental factors.

0.10 Future Development and Expansion Plans

While the current project scope focuses on sentiment-driven weather narration, I do plan to further develop this in the future. The immediate next step would be the integration of transformer-based models such as BERT or DistilBERT (Devlin et al. 2018). These models are pre-trained on large corpora and have been shown to outperform traditional classifiers on a range of NLP tasks, especially where contextual nuance is critical. By fine-tuning such models on weather headlines, the system could achieve even greater accuracy in sentiment detection.

Another direction involves expanding the dataset. While the current model is trained on approximately 137 structured examples, scaling this to several hundred or even a few thousand examples- either through crowdsourcing or automated generation with human review- would make the model more robust. Data augmentation techniques, such as paraphrasing or translation-based expansion, could also be explored.

A longer-term aspiration is to integrate the system with live news feeds and equip NAO with the ability to vocalise them automatically. This would open up accessibility to visually impaired users by enabling real-time, spoken weather updates that reflect both factual and emotional content. Such a feature would transform the robot from a demonstration tool into a genuinely inclusive companion, bridging information gaps for those unable to engage with traditional news media.

0.11 Conclusion

This proposal presents a creative application of supervised machine learning in a robotic setting. By reinterpreting UK weather headlines as emotional cues and mapping them to robotic behaviour, the project demonstrates a real-world pipeline from data collection to interactive expression.

The switch from unsupervised to supervised approaches marked a pivotal moment, improving the accuracy and interpretability of sentiment predictions. The final system will integrate machine learning, human-computer interaction, and ethical design in a compelling way, aligning with the broader objectives of socially intelligent robotics.

Bibliography

Creative Computing Institute (2025), ‘Using nao v6’.
URL: *https://wiki.cci.arts.ac.uk/books/robotics-lab/page/using-nao-v6*

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), ‘Bert: Pre-training of deep bidirectional transformers for language understanding’, *arXiv preprint arXiv:1810.04805* .

Hutto, C. & Gilbert, E. (2014), ‘Vader: A parsimonious rule-based model for sentiment analysis of social media text’, *Eighth International Conference on Weblogs and Social Media* .

Loria, S. (2018), ‘Textblob: Simplified text processing’.
URL: *https://textblob.readthedocs.io/*

Pedregosa, F. et al. (2011), ‘Scikit-learn: Machine learning in python’.
URL: *https://scikit-learn.org*

Žliobaitė, I. (2017), ‘Measuring discrimination in algorithmic decision making’, *Data Mining and Knowledge Discovery* **31**(4), 1060–1089.