Increase ML Model Visibility with NLP Monitoring
Increase ML Model Visibility with NLP Monitoring

Insurance. Healthcare. Financial Services. Retail. Regardless of which industry you’re in, natural language processing (NLP) is at the heart of your customer experience.

Domain specific experts and data scientists rely on machine learning to gain deeper understanding of industry-specific jargon—in application forms, patient/customer documentation, recommender engines, customer service tickets, email correspondence, chatbots, and more—to fine tune NLP models over time.

Enterprises heavily rely on accuracy and explainability of NLP model predictions and insights to inform NLG (natural language generation) strategies for creating equitable customer experiences at scale for deeper engagement, brand satisfaction and retention to increase customer lifetime value (CLV).

NLP Overview

Arthur empowers enterprises to directly ingest raw words and documents in our platform. Next, Arthur computes the embeddings from the original word-document ingestion to surface various visualizations and workflows in the platform UI for drift, anomaly detection, and prediction-level model explainability.

Best of all, Arthur’s NLP performance monitoring for supervised learning use cases are model and platform-agnostic, enabling teams to deploy the solution with any AI tech stack.

Improve ML Performance with Data Drift Detection & Monitoring

An NLP model, just like any other machine learning model, is trained on a training dataset that represents the world at a point in time—and as the world changes and the input data coming into your language model starts to change, your model will start to run into performance issues. Over time, your NLP model performance will degrade if you’re not proactively monitoring for and correcting data drift as it occurs.

Monitoring NLP models for data drift involves comparing the statistical similarity of new input documents to your training set documents. As input documents shift in their typical word-use, you might need to update your model to account for the linguistic patterns. Even as the NLP industry matures and technology advances, accuracy will continue to be a challenge because of the fluid nature of language. Data scientists and domain experts alike need to understand when and where drift is occurring to maintain the integrity of your NLP models over time.

Anomaly detection is a primary data drift feature of Arthur’s platform. It enables teams to identify new inference data points which deviate significantly from original reference training data. In a production environment, you often don’t have prior knowledge of what kind of distribution shifts to expect or access to labeled data.
Identify Unfair Model Bias for NLP Models

Arthur offers bias detection capabilities for NLP models, allowing data science teams to uncover differences in model outcomes across different subgroups. The Arthur platform offers performance-bias analysis for natural language and tabular models, as well as the ability to partition by multiple attributes at a time to provide you more granular insights into potential biases.

You can track the performance of your NLP models on input documents partitioned by sensitive attributes that you may know. For example, you can track whether your medical document classifier is more or less accurate for documents from medical visits of men versus women, or Black patients versus white patients. Uncovering differences in accuracy and other performance measures across different subgroups can help you identify—and fix—unfair model bias.

NLP OOD: News Headlines dataset divided by each category as the reference set. Compares cross-class anomaly accuracies (where another class is all "anomalous").
Understand Black Box NLP Model Behavior

Arthur’s NLP explainability tools provide token-level explanations for your NLP model's predictions. With Arthur, you can gain an intuitive understanding of which specific words within a document contributed most to a given prediction. This is hugely helpful for validation and debugging models as well as bringing less technical stakeholders more insight into critical ML applications.

Flexibility across common NLP stacks

Arthur supports common NLP tasks (classification, multi-label, and upcoming NER) and is agnostic to which tools and platforms you use for development. Whether you’re using HuggingFace, spaCy, TF, scikit, or any other frameworks, you can easily integrate with Arthur to track your data and model stability.

If you’re deploying NLP models into production, you need real-time optics into post-production activities. Arthur offers monitoring support for NLP models, including NLP data drift detection, token-level explainability to provide valuable insight into the key drivers in NLP classification model predictions, as well as model bias detection.