Bibliography Report.

Towards BetterProduct Quality.

Identifying Legitimate Quality Issues Using NLP & Machine Learning.

Presented by: Sandeep Nidheesh.
Sachin Saji.
Alen Sam.
Vijeesh K S.

1. Galke, L. and Scherp, A. (2022)

Author(s), Year: Galke, L. and Scherp, A. (2022)
Publication Venue: **Are We Really Making Much Progress in Text Classification? A Comparative Review**

Research Focus

This influential survey paper conducted a broad meta-analysis questioning the prevailing necessity of complex deep learning models for standard text classification tasks, particularly when compared against well-tuned, simpler alternatives.

Methodology

The authors performed a comprehensive literature review and meta-analysis, systematically comparing reported performance metrics of modern Transformer models (BERT, GPT) against traditional classifiers (SVM, Logistic Regression) across numerous established academic benchmarks.

Key Findings

The key conclusion was that in many standard, non-context-heavy scenarios, simple classical models like **Logistic Regression and trigram-based SVMs** often perform comparably, and sometimes even **outperform, deep models**, especially in low-data regimes. The linear models proved their long-term value and stability.

Relevance

This review provides powerful academic justification for your project's main conclusion and the success of the Passive-Aggressive classifier. It reinforces the principle that the most effective, scalable, and sustainable industrial solution is often the simplest and most interpretable one, validating your approach to avoiding unnecessary DL complexity.

Scope Comparison

This is a general academic review of classification theory, not specific to industrial data. However, its core message—that model simplicity often trumps complexity when efficiency is prioritized—is **directly applicable** and highly supportive of our latency-sensitive industrial environment.

2. Wei, J. and Zou, K. (2019)

Author(s), Year: Wei, J. and Zou, K. (2019)
Publication Venue: EDA: **Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks**

Research Focus

The research introduced simple, "easy-to-implement" techniques for **data augmentation** intended to boost model performance, particularly targeting scenarios characterized by low data volume and high class imbalance.

Methodology

The authors proposed the **EDA (Easy Data Augmentation)** suite of techniques, which includes methods like **Synonym Replacement**, Random Insertion, Random Swap, and Random Deletion. These methods were rigorously tested to quantify their effectiveness when training both convolutional and recurrent neural networks.

Key Findings

EDA consistently and significantly improved performance. Notably, models trained with EDA achieved the same accuracy with only **50% of the original training data** on average across five datasets. **Synonym Replacement** was particularly highlighted as an effective and scalable method for preserving semantic meaning.

Relevance

This paper is **FOUNDATIONAL** for your project. It directly inspired and validated your implementation of **Synonym Replacement** as the primary strategy to address the critical **84/16 class imbalance** in the service job data, ensuring the minority "Rejected" class was sufficiently represented for training.

Scope Comparison

The study focused on general academic text classification benchmarks, not industrial data. Our project can adapt this method, validating its utility and effectiveness specifically within the unique, highly technical, and jargon-filled domain of manufacturing maintenance reports.

3. Huang, Q. et al. (2021)

Author(s), Year:r: Huang, Q. et al. (2021)
Publication venue: **Survey on deep learning methods for imbalanced text classification.**

Research Focus

The goal was a **comprehensive review of advanced deep learning strategies** specifically designed for mitigating performance degradation caused by severe class imbalance in various Natural Language Processing (NLP) tasks.

Methodology

The authors systematically categorized imbalance mitigation techniques into three main groups: **Data-level** (sampling/augmentation), **Algorithm-level** (e.g., **cost-sensitive loss functions** like Focal Loss), and **Ensemble methods**. They summarized the strengths and weaknesses of each category based on contemporary literature.

Key Findings

The survey highlighted that Algorithm-level techniques, which explicitly **penalize the misclassification of the minority class** more severely in the loss function, often yield the best Macro F1-Scores for imbalance without introducing synthetic data noise or risk of overfitting.

Relevance

This work is highly relevant for refining your model in **Future Work**. While your current project used a successful data-level fix (EDA), this study points toward using algorithmic solutions (e.g., weighted loss or Focal Loss) to achieve deeper and more reliable improvements on the critical 16% minority class.

Scope Comparison

As a survey, it provides the theoretical framework but is not an implementation. It serves as a strong guide for our next steps in handling severe class imbalance, a challenge that our current project addressed effectively at the data-level.

4. Fromme, A. and Fromme, A. (2021)

Author(s), Year: Fromme, A. and Fromme, A. (2021)
Publication Venue: **ContextGen: Using GPT-2 to generate synthetic domain-specific training data.**

Research Focus

The study aimed to create **high-quality, synthetic, domain-specific text samples** using generative models. The goal was to alleviate issues of data scarcity and class imbalance by generating more realistic and contextually rich training data.

Methodology

The authors proposed **ContextGen**, a two-step method. First, a large generative model (GPT-2) was used to produce new text. Second, this generated text was verified and labeled using a discriminator model (BERT) before being added to the training set.

Key Findings

ContextGen was shown to be **superior to simple augmentation methods (like EDA)** because the generated data was highly contextually and grammatically rich. This resulted in a final classification model that learned a more robust decision boundary compared to models trained on simpler, less diverse augmented data.

Relevance

This paper outlines a crucial and advanced pathway for your project's **Future Work**. If the simple synonym replacement (EDA) approach is exhausted, using Generative AI (LLMs) to create more realistic and technical fault reports for the 16% minority class is the next logical step to maintain continuous improvement.

Scope Comparison

This work provides a more sophisticated, albeit computationally heavier, alternative to the simple EDA technique adopted by our project, showcasing the rapid evolution of data augmentation strategies in NLP.

5. Oprea, I. et al. (2023)

Author(s), Year:Year: Oprea, I. et al. (2023)

Publication Venue: **Investigating cost-sensitive learning strategies for highly imbalanced multi-class HR texts.**

Research Focus

The research addressed **extreme imbalance in multi-class text classification** by focusing on optimizing the penalty of misclassification rather than simply resampling or augmenting the data. The goal was to minimize the economic impact of errors.

Methodology

The methodology tested various balancing schemes, including oversampling (SMOTE) and a **cost-sensitive approach**. The latter was formulated as a numerical optimization problem, using a Differential Evolution algorithm to automatically determine the optimal misclassification penalty matrix.

Key Findings

The study confirmed that the use of **cost-sensitive classifiers**, where the penalty matrix is optimized based on the financial or operational impact of misclassification, yielded the best performance metrics (Precision and Recall) for the highly underrepresented minority classes.

Relevance

This finding strongly supports the idea that for critical minority classes (like your 16% 'Rejected' issue, which directly impacts warranty cost-saving), **cost-sensitive optimization** is more effective and strategically sound than generic oversampling methods. This is an excellent conceptual validation for your approach to prioritizing the minority class.

Scope Comparison

This research uses a distinct domain (HR job descriptions) but successfully validates the same core principle of focusing on the economic or operational cost of errors in severely imbalanced industrial data.

6. Yadav, J. et al. (2024)

Author(s), Year:: Yadav, J. et al. (2024)
Publication Venue: **Data Augmentation using back-translation and label-conditioned generation for severe class imbalance.**

Research Focus

The objective was to develop **advanced data augmentation techniques** capable of maintaining the semantic integrity and thematic coherence of the original text, specifically targeting severe class imbalance scenarios where simple methods fail.

Methodology

The approach involved combining two sophisticated techniques: **back-translation** (where text is translated from the source language to another and back) and conditional generation to create semantically diverse, high-quality synthetic samples for the minority class.

Key Findings

The authors found that back-translation augmentation significantly increased the F1-Score for minority classes. This demonstrated that generating **sentence-level paraphrasing** introduces higher diversity and less noise than rudimentary word-level synonym replacement (EDA).

Relevance

This paper provides a blueprint for a **more powerful augmentation strategy** for your project's minority class in a future iteration. Given your existing use of a translation API for preprocessing, adapting this method to create better synthetic data is a logical and high-impact next step.

Scope Comparison

This research shows the progression of data augmentation complexity. We will try to leverage the simple and low-cost EDA method, while this work details the higher-quality, next-generation augmentation approach using translation and conditional generation.

7. Piskorski, J. et al. (2025)

Author(s), Year: Piskorski, J. et al. (2025)
Publication Venue: **SemEval-2025 Task 10: Multilingual, Hierarchical, Multi-Label Document Classification.**

Research Focus

This work addressed one of the most complex classification scenarios: **multilingual, multi-label classification** on datasets with severe category imbalance and significant domain shifts between different topics.

Methodology

The methodology primarily leveraged **mBERT/MiniLM** (multilingual Transformers) combined with advanced techniques. These included **contrastive learning** and domain adaptation methods (like Gradient Reversal Layer, GRL) to effectively enable cross-lingual knowledge transfer and handle extreme imbalance.

Key Findings

The research confirmed that multilingual models are indeed the current state-of-the-art for cross-lingual tasks. However, it also showed that even these powerful models require specialized techniques (like GRL and asymmetric loss functions) to handle the combined difficulties of extreme imbalance and multilingual domain shifts effectively.

Relevance

This paper is highly relevant to your project's **multilingual challenge (17 languages)**. It validates the robust, practical approach of your solution: unifying all text into a single high-resource language (English via Google Translate) as a necessary and effective preprocessing strategy to minimize model complexity.

Scope Comparison

This study uses deep models to *directly* handle multilingual input, which is computationally expensive. Our project uses a simpler and more robust translation-based preprocessing step that achieves the same **unification** goal with significantly lower model complexity and resource cost.

8.Conneau, A. et al. (2020)

Author(s), Year: Conneau, A. et al. (2020)
Publication Venue: **XLM-R: Training a large cross-lingual language model on 100 languages.**

Research Focus

The primary goal was the development of a single, massive **pre-trained language model** capable of achieving state-of-the-art results across over **100 different languages** without requiring specific per-language fine-tuning or training datasets.

Methodology

The authors trained a large Transformer model (**XLM-R**) on massive multilingual corpora spanning over 100 languages. The training objective used a masked language modeling approach to force the model to learn deep, shared **cross-lingual representations** across all languages simultaneously.

Key Findings

The research definitively demonstrated that **cross-lingual transfer learning** is highly effective: a model trained only on English labeled data can perform remarkably well on unseen languages (zero-shot inference). This validated the concept of English as a powerful central pivot language for classification tasks.

Relevance

This work provides the **core theoretical backing** for your project's simplified architecture. By translating all 17 languages into English, you are effectively leveraging the universal representation space learned by models like XLM-R and mBERT using a high-resource pivot language.

Scope Comparison

This is a foundational paper for all modern multilingual NLP. Our project's translation approach is a robust, practical, and **low-resource proxy** for utilizing this technology without requiring the full deployment and computational expense of a massive XLM-R model.

9.Arivazhagan, N. et al. (2021)

Author(s), Year: Arivazhagan, N. et al. (2021)
Publication Venue: **Multilingual NMT and text classification for low-resource Indian languages.**

Research Focus

The objective of this study was to improve text classification performance for **low-resource Indian languages** by using Neural Machine Translation (NMT) to pivot text to a high-resource language, effectively solving the data scarcity problem.

Methodology

The authors formally tested the hypothesis that using a **high-quality NMT** system followed by standard classification models (like SVM or BERT) in a pivot language (English) would outperform attempting direct classification within the low-resource language. This involved rigorous comparative experiments.

Key Findings

The experimental results conclusively showed that the **translation-based pivot method** consistently and significantly outperformed direct classification. This outcome emphasizes that securing access to a highly reliable translation system is a powerful and often superior preprocessing step for multilingual text analysis.

Relevance

This study **directly validates the central preprocessing decision** in your project: the method of unifying the 17 source languages into English using a quality translation API. This choice allows your ML classifier to focus solely on identifying technical keywords and patterns in a single, high-resource language.

Scope Comparison

While the focus was on Indian languages, the core methodological finding is universally applicable. Our project's translation step is even more critical given the **higher number and greater linguistic diversity** of the 17 languages involved (European and Asian).

10. Zhang, D. et al. (2023)

Author(s), Year: Zhang, D. et al. (2023)
Publication Venue: **Zero-shot and few-shot cross-lingual text classification using prompt tuning on LLMs.**

Research Focus

The research aimed to develop highly **resource-efficient methods for cross-lingual classification** using minimal or zero-shot labeled data. This was achieved by leveraging the vast, pre-trained general knowledge embedded within Large Language Models (LLMs).

Methodology

The authors employed **Prompt Tuning**, a technique where a small set of continuous tokens are optimized to query a large LLM (like GPT) for the classification task, rather than incurring the expense of fine-tuning the entire model's parameters.

Key Findings

Prompt tuning achieved competitive classification performance in zero-shot cross-lingual tasks. The findings suggest that the underlying LLM already possesses the necessary cross-lingual and domain knowledge, which can be **accessed cheaply and quickly** via minimal prompt optimization.

Relevance

This paper outlines a sophisticated, advanced idea for your project's **Future Work**. If the operational costs of translation become a limiting factor, using **zero-shot prompting** on an LLM could be an alternative to directly classify the native language text (e.g., classifying a German report as "Quality Issue" without translating it first).

Scope Comparison

This study focuses on LLM efficiency and the future of low-resource tasks. Our project provides the current, **robust production foundation** (translation + ML) that these newer, more experimental LLM methods are ultimately trying to replace or enhance.

11. Zhuang, Y. et al. (2021)

Author(s), Year: Zhuang, Y. et al. (2021)
Publication Venue: **Leveraging Multilingual BERT for text classification with limited target language data.**

Research Focus

The research focused on empirically quantifying the effectiveness of **zero-shot transfer** capabilities when moving from a high-resource source language (English) to various low-resource target languages using the Multilingual BERT (mBERT) model.

Methodology

The methodology involved fine-tuning the mBERT model **only on the English portion** of a multilingual training dataset. The model's classification performance was then measured directly on data from other target languages to quantify the zero-shot transfer capability.

Key Findings

The study confirmed that mBERT's shared sub-word embeddings allow for **strong performance in zero-shot classification**. This validation reinforces the idea that having a central, high-quality, English-labeled dataset is the most efficient and scalable core training resource for a multilingual system.

Relevance

This study strongly reinforces the strategic importance of using translation to unify your training data into a high-resource language (English). This strategy maximizes the potential benefit derived from any future use of pre-trained models by focusing labeling and cleaning efforts on a single, universal data set.

Scope Comparison

This work relies on the implicit linguistic relationships learned by the mBERT model. Our project ensures reliable classification success **regardless** of mBERT's specific transfer capability by explicitly translating all input and relying on the resulting high-quality English text representation.

12.Hong, S. and Pishdad-Bozorgi, P. (2022)

Author(s), Year: Hong, S. and Pishdad-Bozorgi, P. (2022)
Publication Venue: **Automated classification of maintenance work orders (CMMS data) using SVM to predict required resources.**

Research Focus

The research aimed to **automate the classification of maintenance work orders** logged in Computerized Maintenance Management Systems (CMMS). The goal was to predict the required maintenance time and personnel resources based on the unstructured text description, enhancing planning efficiency.

Methodology

The methodology involved applying several traditional ML approaches, including **SVM**, Naïve Bayes (NB), and Logistic Regression (LR), to features derived from maintenance report text. The comparison tested their accuracy in predicting numerical resource values, making it a regression-like output task.

Key Findings

The study confirmed that traditional ML, particularly **SVM**, is highly effective for predictive modeling on short, technical maintenance reports. The model achieved reliable resource allocation predictions, proving the utility of these fast models in structured industrial decision-making.

Relevance

This research provides a powerful parallel domain example (**CMMS data**) that validates the inherent suitability of **SVM/LR/Passive-Aggressive** models for structured decision-making based on unstructured maintenance notes. This reinforces your decision to select a fast, reliable linear model.

Scope Comparison

The study focuses on predicting *resources* (a numerical/multi-class value) rather than *fault diagnosis* (your binary classification). However, the unstructured source data's technical nature and domain alignment are highly similar to our project's challenges.

13. Wu, Y. et al. (2020)

Author(s), Year:: Wu, Y. et al. (2020)
Publication Venue: **Using deep learning (LSTM) to value free-form text data for predictive maintenance and duration prediction.**

Research Focus

The objective was to extract actionable intelligence from free-form maintenance logs to predict critical time-based metrics, such as **machine downtime duration** and **time-to-repair**, which are essential for predictive maintenance strategies.

Methodology

The authors employed a deep learning method using a **Long Short-Term Memory (LSTM)** recurrent neural network. The LSTM was applied to text features to explicitly model the sequential and contextual nature of maintenance logs, allowing for the prediction of numerical outputs (a regression task).

Key Findings

The study demonstrated that LSTM models are highly effective at capturing the temporal and contextual flow of text, leading to **accurate predictions of maintenance duration**. This validated the use of deep, sequential models for turning unstructured text into valuable predictive metrics.

Relevance

This represents a relevant Deep Learning (DL) approach for maintenance data, justifying the need for your project to test and compare models like LSTM/MLP in its architecture comparison. It highlights a high-value follow-up task (duration prediction) that could be built upon your core classification system.

Scope Comparison

The core focus is on **predictive regression** (duration) rather than **classification** (fault type). While the maintenance domain is the same, the complexity of the output variable is different from our simpler binary task.

14. Marocco, D. and Garofolo, M. (2021)

Author(s), Year: Marocco, D. and Garofolo, M. (2021)
Publication Venue: **Text mining of maintenance reports for knowledge extraction and summarization.**

Research Focus

The research objective was to utilize text mining and Natural Language Processing (NLP) to extract structured knowledge and underlying patterns from maintenance reports. The key driver was to assist human quality analysts by focusing on **interpretability** and summarization.

Methodology

The methodology primarily consisted of **unsupervised and rule-based text mining methods**, complemented by visualization tools. These tools were used to discover common noun phrases, defect-component relationships, and recurring event sequences within the data.

Key Findings

The study highlighted that the biggest barrier to analysis is **inconsistent, jargon-filled text**, proving that strong preprocessing (like your custom dictionary) is critical. It concluded that highly **interpretable models** and visualization tools are necessary for successful adoption by human analysts and engineers.

Relevance

This work strongly supports your project's emphasis on high-quality preprocessing and provides context for *why* an interpretable model (like the Passive-Aggressive classifier, which uses transparent feature weights) is often preferred over a black-box Transformer for quality control purposes.

Scope Comparison

This work focuses on **knowledge extraction** (unsupervised pattern discovery), whereas our project focuses on **automated decision-making** (supervised classification). The two goals are highly complementary within an operational quality control system.

15. Ohata, E. F. et al. (2022)

Author(s), Year:Ohata, E. F. et al. (2022)
Publication Venue: **Text classification methodology to assist a large technical support system using multiple ML models.**

Research Focus

The objective was to develop a robust **classification pipeline** to assist technical support agents by automatically recommending standard resolutions based on the short, technical text messages submitted by customers or field agents.

Methodology

The study conducted a comprehensive evaluation of an entire pipeline, testing multiple feature extractors (TF-IDF, Word2Vec) and numerous classifiers (RF, SVM, NB, etc.). The experiments were run on real-world, short-text technical support data, simulating an industrial triage environment.

Key Findings

The experimental results consistently showed that simple **ensemble methods and SVM performed most reliably**. The methodology successfully handled the core industrial complexities of short technical wording and initial class imbalance with robust, deployable performance.

Relevance

This paper is **highly relevant** as its core experimental comparison and mission closely mirror your project's. It provides a strong, external baseline showing that classic ML often provides the most robust, reliable, and deployable solution for this specific domain.

Scope Comparison

The study is closely aligned in terms of technical domain and task. However, our project introduces the significant, compounding challenge of **multilingual input** and the absolute necessity of a **custom dictionary** to clean industrial jargon before classification.

16. Zayzay, A. et al. (2024)

Author(s), Year: Zayzay, A. et al. (2024)
Publication Venue: **Utilizing Text-Based Association Rule Mining and LLMs for predictive maintenance on unlabeled data.**

Research Focus

The research focused on the discovery of **frequent failure patterns and correlations** between components and defects within large, unstructured maintenance logs, specifically targeting scenarios where manual labeling is scarce or non-existent (unlabeled data).

Methodology

The authors used **Association Rule Mining (ARM)** applied to text features to find common co-occurring technical terms (e.g., "motor" and "overheating"). This was coupled with LLMs used for summarizing and interpreting the discovered findings for human review.

Key Findings

The ARM technique proved highly effective, identifying **hidden failure modes and correlations** that were often missed or overlooked by standard human review processes. This demonstrated the power of unsupervised text analysis in validating defect patterns.

Relevance

This method is valuable for your project's **Future Work** or parallel analysis, as it can help refine the ground truth data used in your supervised learning model by identifying subtle, recurring patterns that validate the labels of the service reports.

Scope Comparison

The focus here is on **unsupervised learning** (pattern discovery) as opposed to our **supervised learning** (classification). However, the analysis targets the same highly valuable unstructured maintenance log data.

17. Wang, Z. et al. (2023)

Author(s), Year:: Wang, Z. et al. (2023)
Publication Venue: **Hierarchical text classification of industrial safety reports using BERT and domain knowledge graphs.**

Research Focus

The project aimed for a complex classification goal: **hierarchical classification** of industrial safety incidents (e.g., Accident Type $\rightarrow$ Root Cause $\rightarrow$ Severity). This required using complex deep learning models enhanced by structured domain knowledge.

Methodology

The methodology employed **BERT embeddings** combined with a **Knowledge Graph** to explicitly inject structured domain relationships into the classification process. This combination was used to improve accuracy, particularly at the deeper, more specific hierarchical levels.

Key Findings

The research found that the incorporation of an external knowledge graph significantly **enhanced BERT's ability to classify reports hierarchically**. This validated the concept that combining contextual semantic features with structured, human-curated data is highly effective in complex industrial tasks.

Relevance

This work is relevant for your project's **Future Work**. Your current binary problem could eventually be expanded to hierarchical classification (e.g., Quality Issue $\rightarrow$ Motor Fault $\rightarrow$ Wiring Error). The use of component codes in your current feature set is a simple analog to this knowledge graph idea.

Scope Comparison

This is a more complex classification task (hierarchical) but reinforces that **BERT** remains the benchmark for accuracy in complex, structured tasks when computational resources are available.

18. Hong, S. and Pishdad-Bozorgi, P. (2023)

Author(s), Year: Hong, S. and Pishdad-Bozorgi, P. (2023)
Publication Venue: **Comparative study of machine learning algorithms for classifying facility maintenance work order priorities.**

Research Focus

The research objective was to accurately classify the **priority** of facility maintenance work orders (e.g., High, Medium, Low) based entirely on the unstructured text description provided in the request.

Methodology

The study compared various traditional ML algorithms (Decision Trees, **SVM**, Random Forest) on features derived from text data to predict the severity and urgency of the required task. This involved multi-class classification based on severity levels.

Key Findings

The study confirmed that **ensemble methods like Random Forest** often outperform single models for priority-based classification, which involves a multi-class outcome. This validated that ML algorithms are entirely suitable for severity/priority determination based on text input alone.

Relevance

This paper provides direct validation that ML algorithms are appropriate for severity/priority determination, a core task functionally similar to your binary Quality vs. Non-Quality classification. It reinforces the stability and suitability of traditional ML in facility management.

Scope Comparison

This targets a multi-class priority task instead of a binary fault/non-fault task. However, the text type and overall methodology alignment are very strong, confirming the robustness of our chosen field of study.