National College of Ireland

Detecting Customer Dissatisfaction in Support Chats Using AI-Based Sentiment Analysis

MSc Research Project
MSc Artificial Intelligence

Diego Lemos
Student ID: x20204787

School of Computing
National College of Ireland

Supervisor: Dr. Devanshu Anand

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Diego Lemos
Student ID:	x20204787
Programme:	MSc Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Dr. Devanshu Anand
Submission Due Date:	11/08/2025
Project Title:	Detecting Customer Dissatisfaction in Support Chats Using AI-Based Sentiment Analysis
Word Count:	8793
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	25th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting Customer Dissatisfaction in Support Chats Using AI-Based Sentiment Analysis

Diego Lemos
x20204787

Abstract

Early detection of a customer's dissatisfaction during support chat conversations is essential to improve clients retention and experience. Despite of the advances of Artificial intelligence (AI), identify customer's dissatisfaction in a short and informal conversation is still a challenge, especially under weak supervision, where labelled data is scarce. This research examine whether AI-based sentiment analysis is capable of detect effectively a dissatisfaction in support chats before an escalation occurs. Weak supervision labelling throughout lexicon-based and transformer-based sentiment models were used on a public available dataset with over 3 million interaction between support agents and customers on Twitter. We have developed and evaluated three modelling pipeline, one with traditional machine learning model such as Logistic Regression, Random Forest and Support Vector Machine, they were trained on PCA-reduced BERT embeddings, second we used the same models but now they were trained on TF-IDF feature derived from raw text, and the last, a Bidirectional Long Short-Term Memory (BiLSTM) network was trained on tokenised text sequences. The results showed that BiLSTM outperformed all other traditional approaches reaching a precision of 84.45%, demonstrating the effectiveness of the sequential deep learning on informal conversational data. Even though the traditional machine learning using TF-IDF showed competitive results, their performance dropped dramatically when using embeddings with PCA reduction, highlighting the risks of dimensionality reduction in preserving semantic data. These findings emphasise the importance of sequential based models and context-awareness for dissatisfaction detection in customer support environment. Future researches should focus on real time implementations and multilingual capabilities to extend its applicability.

1 Introduction

In recent years, the customer support environment has gradually moved to the online world, making the overall amount of service-related communication in real-time and using text-based communications vast. Since customer experience (CX) has become one of the key determinants of business performance, it is necessary to identify dissatisfaction as early as possible in these discussions in order to increase retention and service quality (Huang and Rust; 2021; McKinsey & Company; 2022). A scalable approach such as AI-powered sentiment analysis provides organisations with a way of automating such manual tasks to help them become proactive to mitigate negative experiences before they get escalated.

The methods of natural language processing (NLP) have improved a lot, where traditional machine learning systems have handcrafted features dependent on pre-conceived linguistic knowledge and replaced them with deep learning models, which can understand intricate linguistic patterns. Traditional methods, like TF-IDF vectorisation accompanied by some algorithms like Support Vector Machines (SVM) and Logistic Regression, have been traditionally effective to perform sentiment classification in structured scenarios (Salton and Buckley; 1988; Araque et al.; 2017). Nevertheless, recent studies demonstrate the effectiveness of deep neural networks in representing sequential dependencies and took semantics in noisy and short-text datasets like social media, using deep neural structure like Bidirectional Long Short-Term Memory (BiLSTM) and transformer networks like BERT of deep neural architecture (Zhang et al.; 2018; Schuster and Paliwal; 1997; Devlin et al.; 2018).

In spite of these developments there are issues of applying the models to customer service situations. Previous studies have concentrated on an unstructured corpus with structured reviews, but here we have a kind of an informal conversation, language, abbreviations and incomplete sentences. Besides, a great portion of real-world datasets that are not labelled with a gold-standard annotation necessitate the application of weakly supervised methods, which, though being scale-efficient, become the source of noise in labels, which deteriorates the model quality (Zhou and Zafarani; 2020; Ratner et al.; 2017). Such a difference indicates that something justifies examining how different modelling strategies perform under that situation.

Research Objective and Question: The research focuses on the possibility of using AI models to predict the early occurrence of customer dissatisfaction in customer support chat, in particular, the question it poses is:

> *Can AI-based sentiment analysis detect customer dissatisfaction in support chats before an escalation occurs?*

Contributions: To address this research question, the current study has several significant contributions. It first evaluates classical machine learning classifications, such as the Logistic Regression, Random Forest, and SVM using the reduced BERT embeddings of the PCA and also TF-IDF features in English. Second, it constructs and tests a Bidirectional LSTM (BiLSTM) deep learning model, which has been trained to work with tokenised raw use of text sequences, which demonstrates the benefit of associating sequential dependencies with customer care data text. Third, it addresses the trade-off the dimensionality reduction introduces, which helps see how the running of PCA on BERT embeddings compromises semantics and becomes a performance issue. Finally, the pilot study also imparts some useful knowledge about how to use AI-based sentiment analysis, which can detect customer dissatisfaction within the weak supervised real world environments.

Thesis Structure: The rest of the thesis is structured as follows; Section 2 presents the literature review of related sentiment analysis and weak supervision researches, section 3 explains the data, pre-processing pipeline and modelling strategy, Section 4 gives an analysis and results of the experiment. Lastly, section 5 ends the study by talking about implications, limitations, and future research directions of the study.

2 Related Work

Support interactions sentiment analysis has become one of the focus areas of the businesses that seek to develop the level of service and predict unsatisfied customers. With the increasing amount of conversational data accessible in the real world and the advances in the domain of natural language processing (NLP), sentiment analysis has also become a widespread usage in various fields, such as customer support, e-commerce or telecom (Mashaabi et al.; 2022; Jia and SungChu; 2020). When referring to the sentiments, in this literature review, there is a consideration of the development of sentiments classification strategies, and in respect to the customer service instances, the sentiments classification strategies are of particular interest.

The papers under review propose a wide range of approaches, the classical machine learning and lexicon based approach all the way through the modern transformer based models to the large language models (LLMs) (Cambria et al.; 2017; Krugmann and Hartmann; 2024). More than that, the issue of human labelling, data imbalance, and explainability has supported the use of weak supervision systems and semi-supervised pipelines (Jia and SungChu; 2020; Salcedo-Gallo et al.; 2022). Other papers indicate that there is a growing amount of importance in the topic of domain adaptation, aspect-based sentiment classification, and sentence-conditioned reaction generation (Idrissi-Yaghir et al.; 2022; Jayakody et al.; 2024).

This review is structured so that it goes through thematic presentation of labelling strategies, comparison between machine learning and transformer-based models, differences between real-time and retrospective sentiment detection approches, and limitations present in existing strategies. It aims at synthesising knowledge and identifying gaps defining the area of focus in this piece of the research.

2.1 Labelling Strategies in Sentiment Analysis

Labeled data is an important process in developing sentiment analysis models is data acquisition which can be challenging in customer support applications where no manually labeled datasets usually exist and sentiment is subjective. In this regard, we have examined a number of studies with alternatives to traditional supervised labelling.

Weak supervision has risen as an existing way to minimize the need manual labelling. Jain Jain (2021) showed how the Snorkel framework could be used to aggregate the results of sentiment lexicon such as VADER, TextBlob, and AFINN, and domain-specific rules to course weak sentiment labels on support chat data. They have found that a RoBERTa model pre-trained on these weak labels could match the performance of commercial APIs, annotating less and achieving the same level of reliability. In the same way, Salcedo-Gallo et al. Salcedo-Gallo et al. (2022) used message-wise sentiment annotations through a rule-based labelling function along side with contextual sentiment development, which allowed them to detect customer dissatisfaction early on without resorting to complete manual annotation.

Another use of domain adaptation strategies is data labelling is a way to use unlabelled domain-specific data to obtain better performance. Idrissi-Yaghir et al. Idrissi-Yaghir et al. (2022) have applied unsupervised domain adaptation to transformer-based models to unlabelled German customer feedback. Their findings suggest that even in the absence of explicit sentiment labels, pre-trained models can be well-adapted with in-domain data distributions, providing a semi-supervised alternative to complete annotation.

A second technique seen in real-world data sets is a posteriori estimation of the sentiment based on user rating or behavioural proxies.To an example, Kusal et al. Kusal et al. (2024) resorted to star ratings on which Amazon reviews that had been assigned as sentiment labels. In as much as this can be applied in a large way, it is risky to generate a label noise since there is a unique variance in the rating criteria and also the neutral scores are vary vague.

Jia and SungChu (2020) has solved the problem of labelling the customer service call through the joint utilisation of automatic speech recognition (ASR) transcriptions as well as analysis of the acoustic features. They also used a semi-supervised pipeline in labelling utterances with linguistic and prosody features integrated into categorizing sentiment into binary class. The benefit of this method is that it reflects multi modal sentiment data, but, it is sensitive to the good quality speech to text converter and intelligibility of the acoustic signal.

The research articles aid in the realisation that the weak supervision, adaptation to domains and the labelling based on heuristics can be effective alternatives to the manual labelling. Nevertheless, they are all associated with compromises in accuracy, generalisability and interpretability.

2.2 Traditional Machine Learning vs Transformer Models

A general trend in the history of sentiment analysis algorithms has been the shift toward transformer-based deep learning models from more traditional machine learning (ML) algorithms. A number of the examined studies present comparative analyses of these paradigms in terms of their effectiveness, scalability, and their appropriateness in relation to the sentiment classification of customers.

In a research by (Ashbaugh and Zhang; 2024), the extensive database examined was the Amazon reviews dataset where deep learning models, such as the CNNs and the RNNs were compared to traditional machine learning models, which included the Random Forest, the Naive Bayes, and the Logistic Regression. It has been surprisingly discovered that Random Forest and Logistic Regression have done much better than RNN and CNN models in providing nearly 99 percent accuracy in 5-category classification issues. The authors explained this performance by the fact that the simplicity of the input text and the power of feature engineering at the token level, deep models do not necessarily provide a strong benefit, unless more complex input structure or domain-specific training is performed.

In a more domain specific setting, Kusal et al. (2024) further compared transformer based models (BERT, RoBERTa, DistilBERT and DistilRoBERTa) with BiLSTM based and CNN based models on the same review data. Their findings revealed that RoBERTa performed better than any other model and got an F1-score of 0.80. On the contrary, Hossain et al. (2020) introduced SentiLSTM, a BiLSTM-based model of sentiment analysis which was trained on restaurant reviews. They obtained high accuracy and balanced precision-recall with their model, but significant manual feature engineering was necessary, and their model was not robust to different domains, a shortcoming which is being progressively addressed by transformer models, using contextual embeddings. The paper has highlighted the increased predominance of transformer architectures, especially at achieving deeper contextual meaning in text even without the necessity of task-specific fine-tuning.

Jayakody et al. (2024) also proved the validity of the fact that the aspect extraction

and sentiment classification stages should be provided by separate pipelines and proposed their version of the hybrid pipeline, InstructABSA, with DeBERTa-v3. They achieved over 91 plus accuracy on SemEval datasets, thus proving that instruction-tuned transformers were decent when it came to general-purpose and adapting tasks. They reported involvement and specialised adjustment required to acquire timely engineering in these models as well.

At the cutting edge of this change are large language models (LLMs), which (Krugmann and Hartmann; 2024) studied. They compared the zero-shot and few-shot sentiment classification features of GPT-3.5, GPT-4 and Llama 2 to enhanced transformer models such as Roberta and FinBERT. The authors raised the points of cost-effectiveness, immediate sensitivity, and what to explain as the aspects that might influence their practical application in customer services situation, though the GPT-4 was more capable to succeed in the binary sentiment analytics than the found systems.

The application of instruction-tuned large language models to few-shot sentiment categorisation was investigated by Wang et al. (2024). Their investigation showed that with a small number of labeled instances, models such FLAN-T5 and GPT can detect intent with great accuracy. The authors observed, however, that without domain adaptation or extra supervision, performance is very sensitive to rapid design and scaling to complex classification tasks can be hard to achieve.

Altogether, all these results substantiate the claim that transformer models are more advantageous than standard models in typical machine learning because they are more generalised and contextualised in terms of their results. Nevertheless, computational costs and technical requirements of deploying cutting-edge transformers or LLMs do not necessarily always align with their marginal utility in the context of the real-life situation.

2.3 Real-Time vs Retrospective Sentiment Detection

Sentiment analysis in customer support settings may have two main applications, retrospective analysis of customer feedback once it has happened, or detecting dissatisfaction in real-time. The presented literature discusses both of the paradigms, but there is a tendency towards the real-time sentiment classification to make proactive improvements of the services.

Salcedo-Gallo et al. (2022) introduced the idea of proactive detractor detection system, labeled message-by message in real time with contextual sentiment trajectories. It was shown to allow intervening early before the customer frustration level got too high, and as such has good significance to a realistic customer support system. On the same note, Huang et al. (2023) proposed a deep learning framework, which learns both semantic and affective representations across conversation windows, improving real-time sentiment change detection in a dialogue. The results of their model prove the usefulness of joint affective information and contextual embeddings when it is necessary to follow the sentiment change across utterances.

In contrast, other tasks focus on retrospective sentiment analysis, especially when working with not time-ordered datasets or those based on static resources such as reviews or social media posts. Hasan and Fattah (2024) proposed a machine learning pipeline to determine customer satisfaction by using several tweet features, such as whether hashtags and mentions were used, as well as temporal metadata. Even though their system is not real-time, it discovers useful post-hoc insights and patterns of user engagement on support-related tweets.

In a more global view, Mashaabi et al. (2022) systematically review the usage of NLP in customer service and note that the majority of sentiment systems are used in a retrospective way. They point to the fact that future systems will need to enable real-time interaction, at least in the use cases such as chatbots and conversational agents. They find that whilst the Twitter and Facebook dataset is now ubiquitous, there is little work that manages to carry retrospective knowledge into the real time.

Cambria et al. (2017) promotes the importance of sentiment analysis described as a big suitcase of sub-problems that the community should learn how to combine single frameworks of real-time, aspect-based, and multimodal sentiment analysis. Although it is a theoretical piece, it resonates with the need of comprehensive and multi-faceted systems of sentiment that can work with real-time and retrospective mode of operation.

Combined, these studies demonstrate that a trade-off existent between retrospective analysis, which puts the picture in a much deeper context and gives insights on a longer-term basis, and real-time systems that allow proactive improvements of services. A combination of the two can be the most beneficial to the contemporary customer experience platforms.

2.4 Limitations in Current Literature and Research Gap

Although significant advances have been made in the sentiment analysis-based customer support, the analysed literature shares a number of common limitations. One of them is the dependency on human-labeled data or domain-personalised proxies like review ratings (Jain; 2021; Kusal et al.; 2024). Although weak supervision and semi-supervised pipelines can save on annotation labor (Jia and SungChu; 2020; Salcedo-Gallo et al.; 2022), these approaches require well-designed labelling functions and usually cannot generalize to other domains.

Even though the studies on RoBERTa, DeBERTa, and GPT-4 exhibit desirable outcomes, there are other weakness is the explainability of transformer based models, in Jayakody et al. (2024); Krugmann and Hartmann (2024) study, the authors also acknowledge that the predictions of large models are difficult to interpret that must be considered when one wants to use them in the high-stakes environments, namely customer service. Moreover, although the edge models with the latest versions have minor advances that match traditional ML in many benchmarks, their application has an incredible level of complexity, cost, and infrastructure requirement (Ashbaugh and Zhang; 2024).

However, a real-time detection of sentiment constitutes an open problem as well. Although systems of this kind are intended to fulfill the task of sentiment classification on a level of a customer message (Salcedo-Gallo et al.; 2022; Huang et al.; 2023), the majority, at the present stage, are retroactive in nature and are unable to take action in time. Moreover, even though domain adaptation methods promise an entry into the multilingual and low-resource world, they fail to work in zero-shot conditions when generalised to the sensitive emotional scenarios or in the chat dominated low-resource scenario rife in slang.

Finally, because sentiment analysis often draws on a range of other relevant tasks, and because of its presence in so many different applications, sentiment analysis is a so-far-unsung big suitcase in the literature (Cambria et al.; 2017). Lack of integrated and multi purpose sentiment models limits the potential application of these models to the more universal application of these models in the various customer support situations.

The existence of such limitations indicates the need of the planned research to complement the poor supervision with powerful models of the transformer type, advance

the interpretability, and enable timely detection of supportive chats in multilingual and informal contexts. The present thesis aims to close that gap by building a sentiment classification pipeline over support chat using weak labels and examining the abilities of the interpretable and transformer-based models to detect dissatisfaction early on the dialogues.

3 Methodology

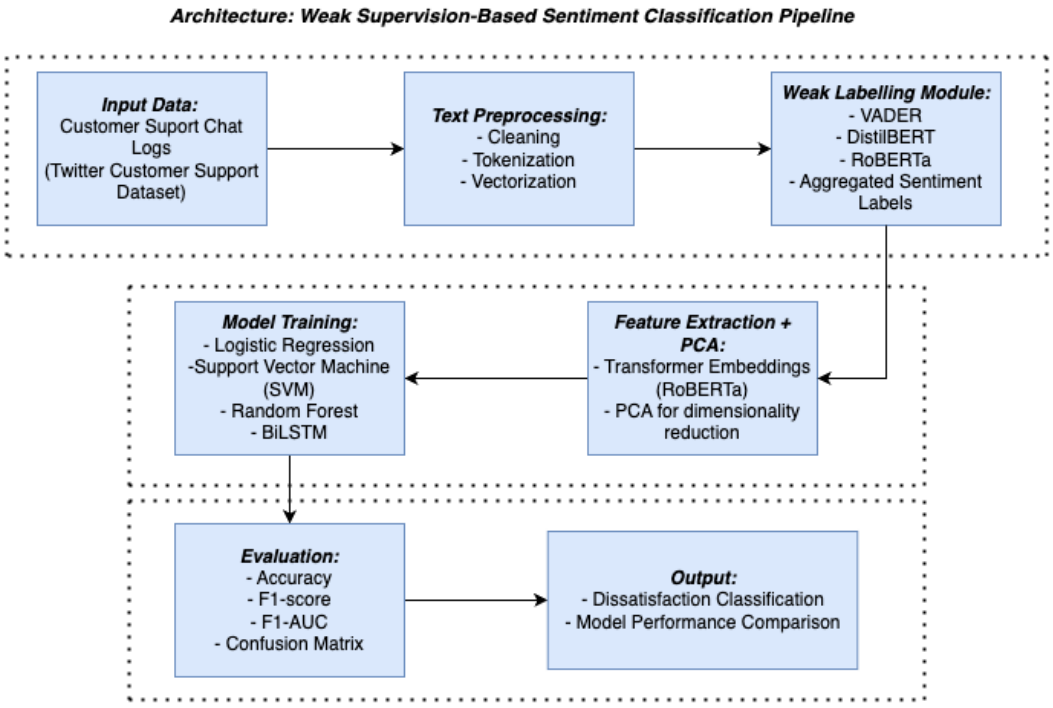

Figure 1: Architecture of the Weak Supervision-Based Sentiment Classification Pipeline

The figure 1 represents the overall sentiment classification pipeline implemented in this research paper.

3.1 Data and Preprocessing

In this study we used the publicly shared dataset Twitter Customer Support Dataset TWCS, which consists of more than two million customer support dialogues between users and large corporations on Twitter. Such discussions are a rich source of information about customer service interactions in real life and pose a complex environment in terms of sentiment classification since they have informal style, are rather short, use a large number of abbreviations, and contain multilingual information.

In order to align to the objectives of this research a rigorous preprocessing pipeline was developed. First, messages were cleaned by excluding the URL, mentions, emojis, hashtags and system-created messages, they were also normalised in lowercasing and whitespaces to maintain consistency.

Furthermore, a relative trivial language detection heuristic was applied to filter non-English tweets with the intend to keep only the interactions in the English language. The conversation was then divided into separate customer messages turning the dataset into a turn-level one where each row was a single input to the sentiment analysis.

The clean text was then tokenised with common python packages like NLTK and TensorFlow/Keras to feed the data into the neural models in the form of vectors. Lastly, a sentiment label of each message was extracted using weak supervision methods that are explained in Section 3.2.

As a result of the preprocessing pipeline, we got to a final dataset with approximately 1.35 milion individual customer messages with a weak label that could be used by the downstream experiments. The given dataset provided an augmenting and imaginable basis to train sentiment classification models in weak supervised conditions.

3.2 Labelling Strategy with Weak Supervision

This larger-scale sentiment analysis required a human intervention which is taxing of resources and time-consuming because it requires manual annotation or classification. To deal with this shortcoming, in this project we followed the idea of providing necessary sentiment labels to the dataset automatically through weak supervision. Such a plan can perform scaling to labelling by taking several heuristic or model-based signals to come up with probabilistic ground-truth labels.

This weak labellers were based on three approaches to sentiment analysis:

- **VADER (Valence Aware Dictionary and sEntiment Reasoner):**, A rule-based social media optimised tool able to identify polarity (positive, neutral, or negative) based on lexical features as well as intensity of a positive/negative statement.
- **DistilBERT:** A faster, smaller and fine-tuned version of BERT to sentiment-classification. DistilBERT is transformer-based viewpoint through contextual embeddings.
- **RoBERTa:** A robust optimised BERT model to amplify the performance of the sentences classification. RoBERTa was used by a step-by-step developed model using sentiment-marked data.

In order establish a stable target variable that all models would be trained on, we have combined the results of VADER, DistilBERT, and RoBERTa in a hybrid voting approach:

In case of unanimous similarity among the three models, the label is given to it. In case of two models voting in the affirmative and one of those models be RoBERTa, then we assume that RoBERTa is correct and abide by that term. When they do not all agree then we fallback to the majority voting.

This strategy guarantees that we can enjoy the benefit of having transformer-based models such as RoBERTa that are robust to word interchange as well as honor consensus in the presence of the same. It is between reliability and coverage, so it is good in weak supervision and training of models.

The resultant weak labels, although not ideal, made up a good training signal to supervised stream learning experiments. The results of the models trained on these labelling are then examined and reported in Section 3.5, Perfect label quality is not assured

since this is a weakly supervised arrangement. The hybrid voting mechanism, however, offers practical tradeoffs of trust in state-of-the-art models and rule-based consensus.

3.3 Feature Extraction and Representation

In this project, several methods of encoding textual information in a numerical manner were investigated to make sentiment classification of the text possible with the approach of classical learning models and deep learning. There were two broad courses of action that were pursued, transformer-based sentence embeddings and frequency-based vectorisation techniques.

With regard to the transformer-based approach, embedding sentences could be generated by the pre-trained BERT-based model through the SentenceTransformers library. Such embeddings have proven to work quite well in capturing semantic relationship and contextual dependencies in natural language, which is why they have been successfully applied to modeling sentiment analysis and particularly short and informal language like customer support messages. In contrast with the former approaches like TF-IDF or Word2Vec, BERT does use the word order and context to make itself more robust over noisy social media text.

By memory considerations, the embedding process was carried out in chunks. This dataset has been grouped in chunk and this chunk was processed and saved as a pkl file. In this way, the embeddings could be reused in other stages and spare unnecessary calculation, being more efficient.

The BERT embedding were fixed to 768 dimension, the standard for classic BERT models, While this woks fone for the traditional machine learning models, it does not work for deep learning models sequential-based, such as the BiLSTM that requires input in tokenised levels in the form of ordered sequences of fixed-size vectors.

The BERT vectors were of small size so as to reduce on the computing resources and storage demands and thus, they were reduced according to the method of Principal Component Analysis (PCA). It has reduced the embedding to 100 dimensions/out of the 768 of original variance and it is still above 95% using IncrementalPCA. Although this dimension reduction resulted in higher speed of processing and effective training of the models, performance trade-off lacked accuracy owing to the fact that any reduction in the accuracy of the model classification occurred particularly with the SVM and Random Forest models in question that were handled most. This implies that there was a loss of semantic information during the compressing stage, this is since it resulted to a negative impact in the performance of the classifiers in regard to predicting between the classes.

In the other hand, BiLSTM pipeline made it possible to exclude PCA-reduced embeddings, which play a pivoting role in retaining the context and the sequence information.

At the same time, a frequency representation strategy would be researched and considered in terms of the Term Frequency- Inverse Document Frequency (TF-IDF) method. This conventional approach estimates the relative importance of the used words in a message relative to the whole corpus and was used to train the conventional models such as logistic regression, support vector machines and random forests on raw text.

In the deep learning model, the raw text underwent the tokenisation process to be transformed into padded integer sequences. The sequences were sent to an embedding layer in the BiLSTM model, which lets the network create task specific word representations as they are trained.

The important factor that this project lead to the evaluation of traditional and

neural models on a more detailed task through the application of the embedding-based, frequency-based and sequence-based representations. Such a comparative framework provided the information about the advantages and disadvantages of each encoding strategy, regarding the weakly labeled sentiment data.

3.4 Model Selection

This paper examined the standard and deep learning methods in weakly-supervised sentiment classification. The choice of the models represented a variety of algorithmic families, which allowed analyzing them comparatively within the frames of various learning paradigms.

Logistic Regression, Random Forest, and Support Vector Machines (SVM) were examples of models used due to their explainability, performance, and basic efficiency in doing the tasks of text classification. These models were especially fine to work with TF-IDF vectorisation and PCA-reduced BERT embeddings because it was possible to adapt those to fixed-size numerical input vectors.

BiLSTM (Bidirectional Long Short-Term Memory) on the other hand has been chosen to represent the deep learning family, so due to its ability to learn long-term dependencies in a text sequence, as well as to learn the context in both directions of a text sequence. The model was used on raw text that was in a sequential form and was tokenised instead of having the input embedded like in the traditional classifiers. The choice was in line with architectural design of BiLSTM, which takes input by token, that is a sequence of fixed-length vectors as a matter of time.

Training Strategy: All models were treated using stratified 5-fold cross-validation training strategy, which is a training technique that makes results robust by ensuring it is fair with respect to the classes distribution. Such approach avoided overfitting and made it possible to estimate the performance more reliably, especially because sentiment weakly labeled data are imbalanced in nature.

Hyperparameter tuning: All models were hyperparameter tuned by means of grid search through a parameter space of choice. This was done with independence in training folds so as to prevent leakage of data. Parameters Regularisation strength of Logistic Regression, number of estimators and depth of Random Forest, and kernel of SVM were optimised using cross-validated F1-score.

Scalability and Efficiency: The training of the classical models was relatively fast in time and less demanding of memory space especially when specialised to PCA but low embedding. On the other hand, the BiLSTM models introduced a bigger cost of computation, due to their sequential nature, and a larger model. In an attempt to accommodate such limitations experiments on deep learning were carried out on small pieces of batch and limited epochs in a way that the experiment resources would not be applied in the way of the experiment.

Such a variety of the model portfolios could allow a detailed comparison of interpretable conventional models and context-aware neural networks in the same data circumstances.

3.5 Evaluation Strategy

In order to critically examine the performance of the models chosen, this paper used multimetric assessment approach that focused on predictive performance and discrimination

of classes. Since the sentiment labels were weakly supervised and sometimes noisy, it was crucial to select the right metrics to achieve meaningful results.

Cross-validation: All the traditional machine learning models were subjected to the 5-fold cross-validation. The data set was split in to five stratified subgroups, to maintain class balance. In each iteration four folds were used for training while the last one were used for testing ensuring every fold served as a test once. However, due to significantly higher computational cost, the 5-fold cross-validation was not applied to BiLSTM model as we had already used three separated train-test splits (80/20, 75/25 and 70/30) that allowed us to evaluate its performance under different distributions.

Evaluation Metrics: Such metrics as accuracy, precision recall, F1-score (macro and weighted), and confusion matrix were recorded and later averaged across all folds. Table-1 gives a brief description of each of the metrics.

Table 1: Evaluation metrics used for model performance assessment.

Metric	**Description**
Accuracy	Proportion of correctly predicted instances over all samples.
Macro Precision	Precision calculated independently for each class, then averaged equally.
Macro Recall	Recall calculated independently for each class, then averaged equally.
Macro F1-score	Harmonic mean of Macro Precision and Macro Recall.
Weighted Precision	Precision averaged across classes, weighted by class frequency.
Weighted Recall	Recall averaged across classes, weighted by class frequency.
Weighted F1-score	F1-score averaged across classes, weighted by class frequency.
Confusion Matrix	Tabular layout of true vs. predicted classes.

Cluster of Comparison and Analysis: Following the cross-validation, means of the scores of each metric were calculated and included in a table so that they could be compared. Per-model analysis as well as per-representation analysis were carried out. Particular attention was given to the negative sentiment detection (i.e. dissatisfaction), which is of special interest in the real world application in customer service environment.

Explainability Considerations: Although explainability was not a fundamental feature of this research, sometimes qualitative knowledge would be able to be made by eyeballing cases manually where something was misclassified. Nevertheless, no automatized explainability tools (e.g., SHAP or LIME) were used. Such methods can be used in the future work to perform more insightful analysis of feature contributions in conventional and deep learning models.

This method of evaluation offered a highly detailed scheme to evaluate models with regard to their predictive abilities not just, yet also with regard to how well they could work on the weakly supervised data. The resulting measures were used to guide the further analysis which is available in Section 5 where model performance is reported and compared at length.

3.6 Tools and Frameworks

The project used Python (v3.9) as the sole programming language, which is commonly used in machine learning and general natural language processing (NLP) applications

thanks to a large number of libraries and the ability to integrate with them well. All development and experimentation was done in Jupyter Notebooks and with the code run in a local environment managed with conda to make it reproducible.

Preprocessing and manipulation of data: These data manipulation process included text cleansing, language filtering, and tokenisation were carried out with the assistance of such tools as pandas, NumPy, and NLTK. These libraries provided good frameworks in the manipulation of tabular and text data which was of big scale.

Feature representation: It was performed by TF-IDF using the scikit-learn module, for example, TfidfVectoriser, and sentence-derived representations were obtained, by using the sentence-transformers module with a pre-trained BERT model, such as a pre-trained BERT sentence representation. IncrementalPCA of scikit-learn was used in providing a dimensionality reduction.

Model Implementation: There were implemented three traditional machine leaning models, Logistic Regression, Random Forest and Support Vector Machine utilising scikit-learn library, as for deep leaning models we have selected a LSTM Bidirectional (BiLSTM) network that was implemented using TensorFlow and Keras, leveraging GPU acceleration where available to speed up the training. We have also utilised Keras preprocessing utilities to prepare text data through tokenisation and sequence padding.

Evaluation and Visualisation: The models performance were evaluated through accuracy, macro F1-score, weighted F1-score and confusion matrices, all of them computed with scikit-learn metric module. For visualisation such as confusion matrices and metric comparison plots we have used Matplotlib and Seaborn to help interpreting the results.

Version Control and Reproducibility: Git was used to version artifacts in code and experiments. Midway results as an example embeddings and other model outputs were written to serial files for example, pkl, json so that they can be re-used to avoid unnecessary re-computation and improve modularity.

The project dependencies and version are outlined in the attached requirements.txt file that will allow full re-replication of the experiment in frames compatible with Python.

4 Methodology Summary

In this section, the methodological framework for developing and evaluating sentiment classification models utilizing weak supervision was described, the data was preprocessed, weakly labelled using sentiment models predictors and also represented through feature extraction methods such as transformer-based embeddings and TF-IDF, both classical machine learning models and the deep learning model, BiLSTM, were implemented and trained using the appropiate validations strategies, also evaluated with multiple performance metrics. In the next section we bring and discuss the results of these experiments.

5 Evaluation

This section explains the flow used in the experiment to analyse different approaches to both model and feature representation in the context of weakly supervised sentiment classification. The overall goal of the experiments was to evaluate the performance of the variety of encoding methods and architectures on short, informal, and noisy customer support dialogues.

The first method entailed producing high-density sentence-level representations through a pre-trained 6- version of BERT based on a SentenceTransformers package. Because of their very large dimensions, 768 features, these embeddings were then downsized through Principal Component Analysis (PCA) in order to achieve faster training with classical machine learning models. Such PCA reduced embeddings were then used to train Logistic Regression, Random Forest with Support Vector Machine (SVM). However, even under several splits of data such as 80/20, 75/25, and 70/30, there were poor results in this particular configuration and it is possible that the semantic information was lost in the dimensional reduction procedure. Such consistency over the various splits can be seen in Figure 2 in which BiLSTM consistently has good performance metrics irrespective of the train/test ratio.

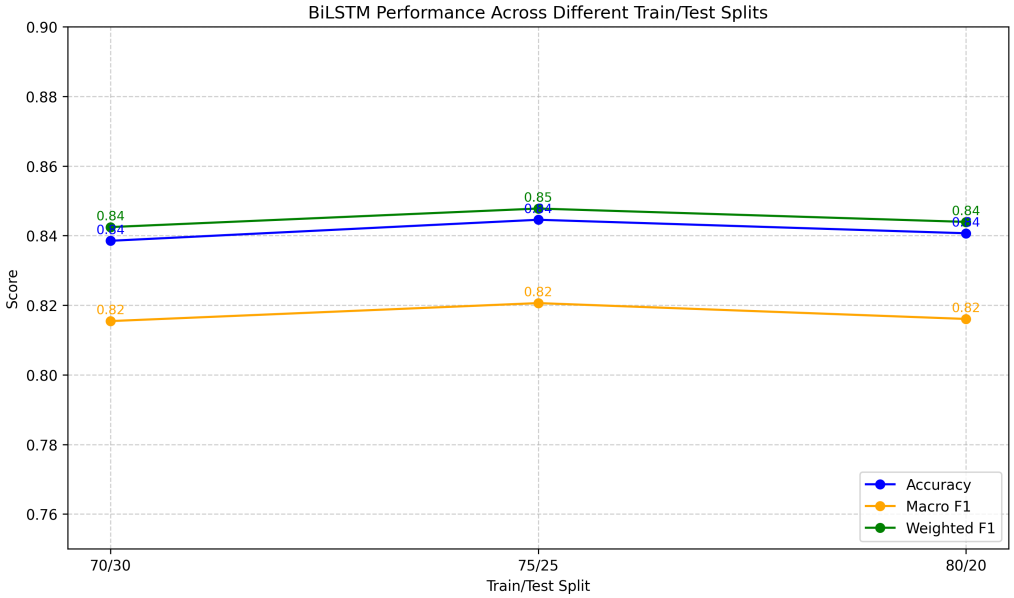

Figure 2: BiLSTM Performance Across Different Train/Test Splits. Scores remain consistent across varying splits, showcasing model stability.

In order to achieve better results, the next method employed used raw text data encoded with Term Frequency Inverse Document Frequency (TF-IDF). This allowed the traditional classifiers to run with raw sparse frequency-based vectors and maintain interpretability over the word scale without undesired compression effects of embedding. This resulted in a improved performance, especially when identifying negative sentiment, and it shows that PCA is unstable method to transform the embeddings generated by a transformer in this context.

Parallel to that, a Bidirectional Long Short-Term Memory (BiLSTM) network was trained on raw text sequences which were tokenised and padded. In contrast with the classical models, the BiLSTM model had sequences of token-level input and the word representations are learned inwardly during training. The model had the advantage of not performing dimensionality reduction, providing improved contextual dependencies of the data.

The comparison on all the models was done through a strict 5-fold stratified cross-validation system where the results were stored based on several measures, accuracy, F1-score (macro and weighted), and confusion matrices. Such a setting of the evaluation made the comparison of models and representation approaches rather fair.

Results of all configurations are shown in detail in the following subsections, and the summaries and discussion of key findings conclude them.

5.1 Traditional Models on PCA-Reduced Embeddings

The initial set of experiments had the objective to understand the performance of conventional machine learning classifiers trained on sentence-level embeddings of the BERT model to be compressed using Principal Component Analysis (PCA). The rationale in adopting this method was to take advantage of the semantic nature of transformer based embeddings whilst minimising computational reduced by dimensionally harming the data.

The sentences were first encoded to 768 dimensional vectors with the help of a pre-trained BERT model included in the SentenceTransformers package. In order to reduce the dimensionality so that this representation is more tractable to classical models, an Incremental PCA approach was used to reduce the dimensionality to a 100 component representation that maximised the number of components that had to be retained in order to meet the criteria. Input features are the lower dimensional vectors which were then trained on Logistic Regression, Support Vector Machine (SVM) and Random Forest classifiers.

In spite of the theoretical merits of deploying embeddings, the models depicted a uniformity of poor performance tests across train test scheme of 80/20, 75/25 and 70/30 split. The average F1-scores were low especially in the minority class (negative sentiment), such that there is perhaps a chance that during the PCA transformation, key semantic dimensions that aid in proper sentiment discrimination could as well have been removed.

The overall performance standard measures after training the classifier on each of the 20 classes using 5-fold cross validation with PCA-reduced embeddings are summarised in the table below (Table 2). In all set-ups, models found it difficult to classify sentiment classes particularly in the area of identifying dissatisfaction which is a central focus of this study.

Table 2: Final Performance Comparison Across All Models and Feature Strategies

Model	**Features**	**Split**	**Accuracy**	**Macro F1**	**Macro Precision**	**Macro Recall**	**Weighted F1**	**Weighted Precision**	**Weighted Recall**
Logistic Regression	BERT + PCA	80/20	0.5400	0.2833	0.3367	0.3342	0.4324	0.4215	0.5400
Logistic Regression	TF-IDF	80/20	0.7507	0.7321	0.7345	0.7566	0.7587	0.7844	0.7507
Random Forest	BERT + PCA	80/20	0.5697	0.2424	0.3718	0.3334	0.4140	0.4455	0.5697
Random Forest	TF-IDF	80/20	0.7442	0.7105	0.7261	0.6989	0.7493	0.7480	0.7541
SVM	BERT + PCA	80/20	0.4862	0.3257	0.3395	0.3375	0.4218	0.4233	0.4862
SVM	TF-IDF	80/20	0.7512	0.7094	0.7162	0.7481	0.7463	0.7463	0.7512
BiLSTM	Tokenized Sequences	80/20	0.8417	0.8206	0.8037	0.8323	0.8478	0.8544	0.8455

To continue on the examples of model failures, in Figure 3 we can see the confusion matrix of the model which displayed the best performance in this configuration. The matrix reveals a bias in most of the predictions to the majority (neutral) class, and a large number of negative sentiment cases were misclassified as an issues of class imbalance which was compounded by PCA compression.

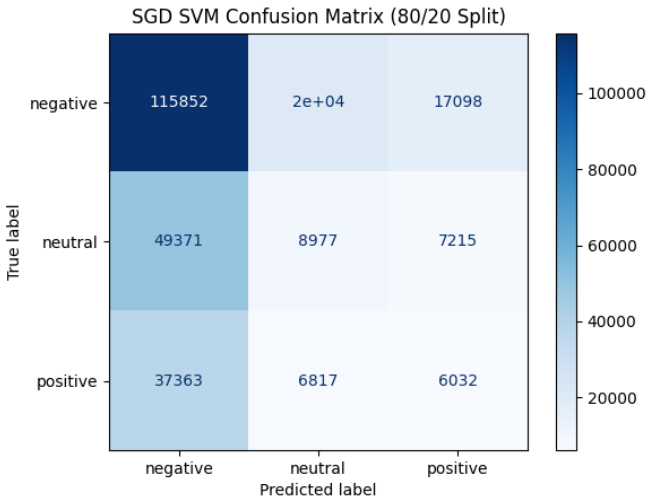

Figure 3: Example confusion matrix – SVM on PCA-reduced embeddings (5-fold CV)

These initial findings indicate that although BERT embeddings have the strength in holding contextual sense, when PCA was implemented, a bottleneck has been integrated that undermined model discrimination. Therefore, the next experiments were aimed at avoiding dimensionality reduction and testing raw text-based strategies of feature extraction.

5.2 Classical Models on Raw Text (TF-IDF)

Having identifying that PCA had its limitations when applied to the compressed versions of BERT embeddings, we tried a traditional method of text representation, the Term Frequency-Inverse Document Frequency (TF-IDF). It represents a simplified and more interpretable representation of text since it emphasises on the relative weight of words in individual messages as opposed to the weight of these words against the message corpus. In contrast to dense embeddings, TF-IDF maintains the lexical granularity and spaciousness of raw texts, it does not use out-of-the-box pretrained models and assumes no sequence geometry.

Here, during this part of the experiments, there were three identical traditional classifiers (Logistic Regression, Support Vector Machine (SVM) and Random Forest) trained over identical form of feature representations (TF-IDF representation) with use of same dataset splits. The intended goal was to see how those models could be further boosted using the lexical-level signatures that are embedded every raw text, and whether this approach to them could work even better than the previous embedding-based approaches.

The results showed that the gains registered in all the models were high. To that end, the classifier using SVM in combination with TF-IDF achieved an accuracy of 75.12%, which was nearly 15 percent higher than the one using PCA. Similarly, the Logistic Regression and the Random Forest classifiers have also exhibited a stable progression in the F1-macro and weighted scores, in the sense that they have increased the number of negative sentiment minority groups that can be detected. These gains are indicative of the fact that under the conditions of short and informal text with weak labels simpler lexical representations can outperform more semantically rich embeddings, in particular when such embeddings are dimensionally compressed.

The comparison of the PCA-based and TF-IDF-based outcomes was already provided in Table 2. The confusion matrix of the best performing conventional model in this set is that of the SVM using TF-IDF which showed much saner classification over the three classes of sentiment classes, as depicted in figure 4.

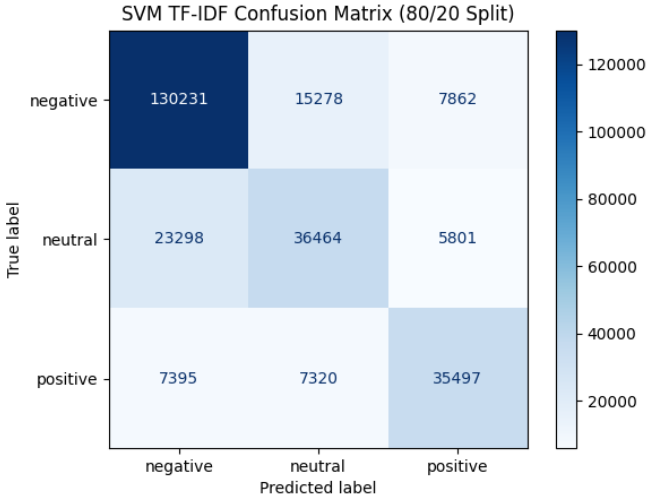

Figure 4: Confusion Matrix – SVM with TF-IDF

5.3 BiLSTM on Tokenized Sequences

The last experimental setup explored the use of a deep learning model, the Bidirectional Long Short-Term Memory (BiLSTM) applied to tokenised sequences of raw text. The BiLSTM network in contrast to the classical models using TF-IDF based features, was created to learn the internal representations of word sequences and also learn to keep the past and the future context. Sequential modelling is a beneficial feature in cases of sentiment analysis as the order of words and nuances in word phrasing can have a significant effect on meaning of it.

At this stage, to feed the model, all messages of customers were initially tokenised and represented in the form of an integer list. Then these sequences were padded to make the same length in order to facilitate the batch training. The neural architecture had an embedding layer to learn dense representations of words at training time, so the network could adjust the space of words according to the classification problem itself. This is contrary to the BERT embeddings in the previous experiments which is stationary and pre-trained.

BiLSTM proved to be the best overall model with all the tried settings. It recorded an accuracy of 84.45% and it outranked all other models in all measures of the macro and weighted F1-score. Notably, it came out as the best in detecting the instances of negative sentiments, the tent of most interest in this project. This implies that the sequential and contextual data retrieved using the BiLSTM played an important role in disambiguating short, informal, and often noisy messages used in customer support.

In Figure 5 we can find the confusion matrix of BiLSTM model which is more balanced than in the previous model in terms of the distribution of the predictions on all classes, and an obvious boost in recall on the minority class.

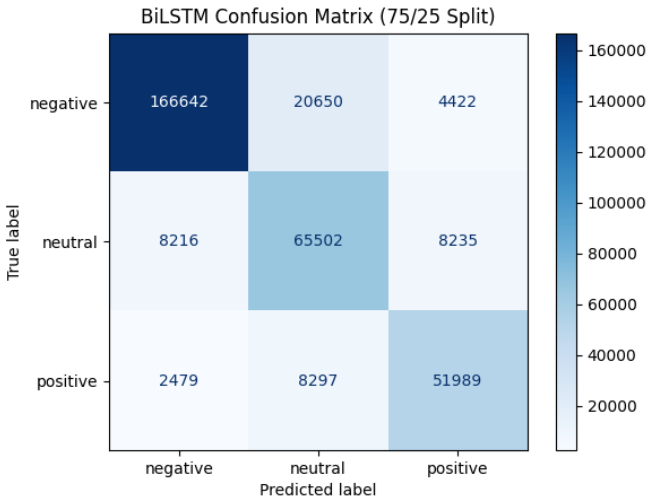

Figure 5: Confusion Matrix – BiLSTM on Tokenized Sequences.

5.4 Final Model Comparison

The BiLSTM model, which was trained on tokenised sequences, reported the best overall performance when it came to the accuracy score, macro F1 score, and the weighted F1 score as compared to all the other models. These findings validate the principle that deep contextual embeddings are strong places to get sentiment subtlety in the support chat data.

Figure 6 offers a side-by-side comparison of BiLSTM and all the top performing traditional machine learning algorithms in terms of key evaluation metrics. BiLSTM model consistently outranks all 3 metrics, which is evidence that it's a good model for this task.

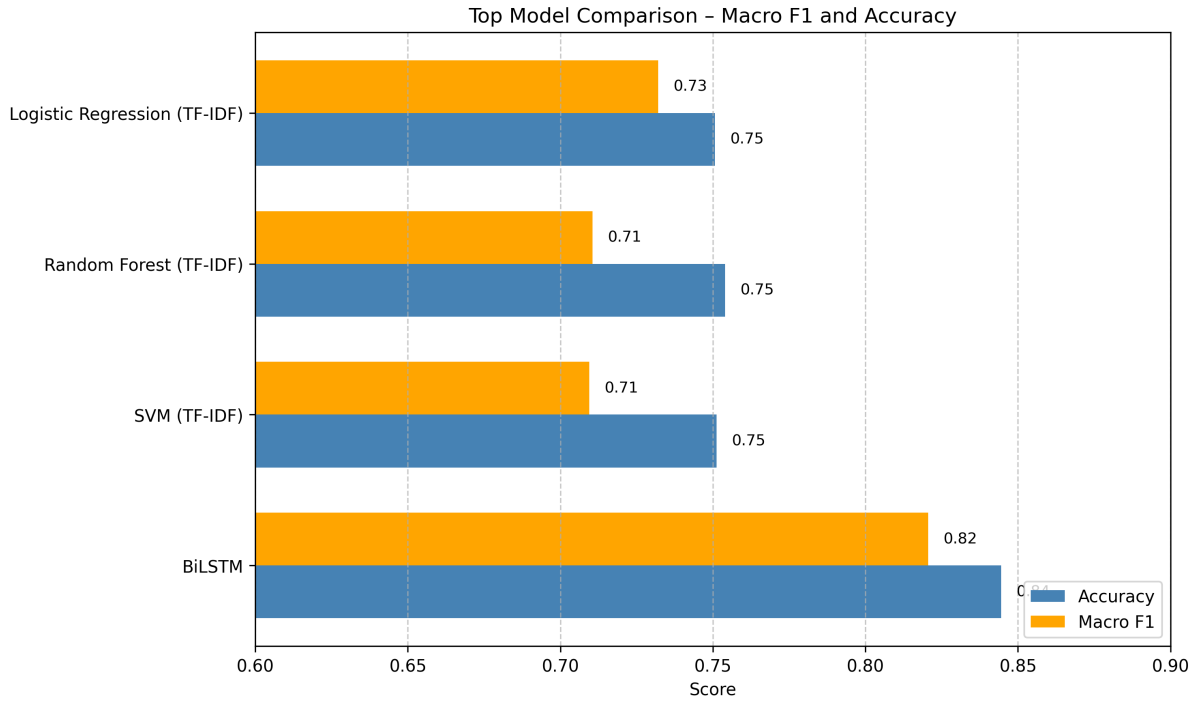

Figure 6: Comparison of Model Performance Across Evaluation Metrics. BiLSTM significantly outperformed the baseline ML models on all metrics.

6 Conclusion and Future Work

The question to be answered in this study was the following: Can AI-based sentiment analysis be used to detect customer dissatisfaction in support chats before an escalation occurs? This findings proves that AI-based models indeed could identify the dissatisfaction sign at the start of customer support dialogues to the extent that it runs its work in the context of work with weakly supervised sentiment classifications and short informal texts.

The best scores indicated the BiLSTM using raw tokenised text sequences to attain accuracy of 84.45% and macro and weighted F1- scores. Such results confirm the effectiveness of deep learning neural network design with access to sequential and contextual data when compared to the classical machine learning encodings which have access to either TF-IDF or PCA-reduced embeddings.

It will also be noted in the findings that it can be simply stated that the semantic richness of transformer-based embeddings, as is, with dimensionality reduction the performance is severely compromised and this just goes on to emphasised the importance of not losing much contextual information here. TF-IDF on the other hand gave consistent outputs on classical models but was not capable of maintaining the contextual depth that would come out on par with BiLSTM.

All in all, the results demonstrate that sentiment analysis with the help of AI can actually reveal when a person is dissatisfied early on in a support conversation in real-life customer services contexts and give a great basis to implement proactive escalation prevention measures.

6.1 Limitations

Although it has provided contributions, this study has also several limitations. First, weakly supervised labelling strategy can have possibly added noise to the training task, affecting the absolute performance of the model. Also, the work aimed at a monolingual environment and partly contacts with customers only in English, which restricts the setting where we can apply the work. It also limited tuning of model hyperparameters because of computational restrictions and did not carry out experiments with real-time or production level data streams. Lastly, although this study have focused on the models performance, it have not touch on the model interpretability or fairness, which can be very significant factors with regard to real-world applications.

6.2 Future Work

This work can be extended in a wide range of directions on the basis of the obtained findings. Significant incremental improvements can also be obtained by application of more sophisticated transformer-based models, like RoBERTa or GPT- architecture models without dimensionality ridge the data when trained end-to-end on specific domain data.

As the last sources to reach the higher results in detection of the minority classes, the dissatisfied customers and, consequently, the earlier intervention of the presentations of the customer support cases can be utilised as well, that is why it is also possible to utilize more strongly those methods and techniques, which allow to overcome the class imbalance, including, the synthetic minority oversampling of the data or cost-sensitive learning. Next to that, a prospective researcher shall desire to employ additional such model explanation tools like SHAP or LIME in order to fulfill the role of interpretation as well as rational credence estimates at an individualistic prediction level and noticing the utilisation of the tools in question.

At last, the practical utility of the model concerning both the predictive warning of a dissatisfied customer, as well as the optimisation services, can be characterised by defining how the model is put into practice with regards to a real customer support environment. They would also deepen the association between the outcomes of the experiments and the implications of the outcomes on the operations and would shaft into a practical customer experience management provided by AI.

6.3 Closing Remarks

In conclusion, the paper demonstrates that with the sequential token representation, deep learning models such as BiLSTM can outperform traditional machine learning sentiment analysis on customer support conversations, and it further exemplifies the significance of how much the data representation and model architecture selection are. Furthermore, the current study contributes to the theme that could be used in future research aimed at developing more realistic and effective AI-based tools to be applied to customer services environment.

To provide the visual support to the idea that the BiLSTM model will be better than the traditional methods of the machine learning, the comparative bar graph is provided in Figure 7. Trained with tokenised sequences, the BiLSTM model continued to outperform the other models in terms of all evaluation metrics such as accuracy, macro F1 and

weighted F1, and also underlines the strength of its capacity to capture sentiment in customer support chats.

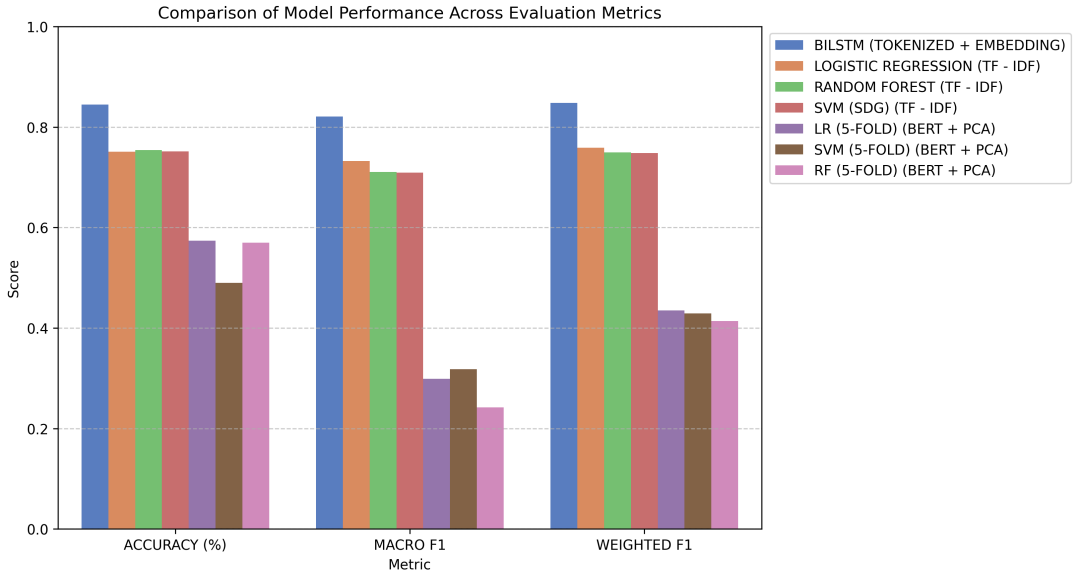

Figure 7: Bar chart comparison of best performing models across key evaluation metrics (Accuracy, Macro F1, Weighted F1).

References

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F. and Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications, *Expert Systems with Applications* **77**: 236–246.

Ashbaugh, L. and Zhang, Y. (2024). A comparative study of sentiment analysis on customer reviews using machine learning and deep learning, *Computers* **13**(12): 340.
URL: *https://www.mdpi.com/2073-431X/13/12/340*

Cambria, E., Poria, S., Gelbukh, A. and Thelwall, M. (2017). Sentiment analysis is a big suitcase, *IEEE Intelligent Systems* **32**(6): 74–80.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.
URL: *https://arxiv.org/abs/1810.04805*

Hasan, M. M. and Fattah, S. A. (2024). A machine learning approach to detect customer satisfaction from multiple tweet parameters, *arXiv preprint arXiv:2402.15992* .
URL: *https://arxiv.org/abs/2402.15992*

Hossain, E., Sharif, O., Hoque, M. M. and Sarker, I. H. (2020). Sentilstm: A deep learning approach for sentiment analysis of restaurant reviews, *arXiv preprint arXiv:2011.09684* .
URL: *https://arxiv.org/abs/2011.09684*

Huang, B., Tang, J., Li, Y. and Song, L. (2023). Customer sentiment recognition in conversation based on contextual semantic and affective interaction information, *Applied Sciences* **13**(12): 7807.

Huang, M.-H. and Rust, R. T. (2021). Artificial intelligence in service, *Journal of Service Research* **24**(1): 3–21.

Idrissi-Yaghir, A., Schäfer, H., Bauer, N. and Friedrich, C. M. (2022). Domain adaptation of transformer-based models using unlabeled data for relevance and polarity classification of german customer feedback, *arXiv preprint arXiv:2212.05764* .
URL: *https://arxiv.org/abs/2212.05764*

Jain, N. (2021). Customer sentiment analysis using weak supervision for customer-agent chat, *arXiv preprint arXiv:2111.14282* .
URL: *https://arxiv.org/abs/2111.14282*

Jayakody, D., Ayesha, B., Gunasinghe, I., Sandaruwan, S. and Suraweera, C. (2024). Instruct-deberta: A hybrid approach for aspect-based sentiment analysis on textual reviews, *arXiv preprint arXiv:2408.13202* .
URL: *https://arxiv.org/abs/2408.13202*

Jia, Y. and SungChu, S. (2020). A deep learning system for sentiment analysis of service calls, *arXiv preprint arXiv:2004.10320* .
URL: *https://arxiv.org/abs/2004.10320*

Krugmann, J.-O. and Hartmann, J. (2024). Sentiment analysis in the age of generative ai, *International Journal of Data Science and Analytics* .
URL: *https://link.springer.com/article/10.1007/s40547-024-00143-4*

Kusal, S., Patil, S., Gupta, A., Saple, H., Jaiswal, D., Deshpande, V. and Kotecha, K. (2024). Sentiment analysis of product reviews using deep learning and transformer models: A comparative study, *Artificial Intelligence: Theory and Applications*, Vol. 843 of *Lecture Notes in Networks and Systems*, Springer, pp. 183–203.

Mashaabi, M., Alotaibi, A., Qudaih, H., Alnashwan, R. and Al-Khalifa, H. (2022). Natural language processing in customer service: A systematic review, *arXiv preprint arXiv:2212.09523* .
URL: *https://arxiv.org/abs/2212.09523*

McKinsey & Company (2022). The state of customer care: Ai and analytics in customer service.
URL: *https://www.mckinsey.com/business-functions/operations/our-insights/the-state-of-customer-care-in-2022*

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S. and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision, *Proceedings of the VLDB Endowment* **11**(3): 269–282.

Salcedo-Gallo, J., Solano, J., García, H., Zarruk-Valencia, D. and Correa-Bahnsen, A. (2022). Proactive detractor detection framework based on message-wise sentiment analysis over customer support interactions, *arXiv preprint arXiv:2211.03923* .
URL: *https://arxiv.org/abs/2211.03923*

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval, *Information Processing & Management* **24**(5): 513–523.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks, *IEEE Transactions on Signal Processing* **45**(11): 2673–2681.

Wang, C., Li, Q., Zhao, T., Liu, Y., Liu, S., Zou, W. and Zhang, X. (2024). Instruction-tuned large language models as few-shot intent classifiers, *arXiv preprint arXiv:2403.17536* .
URL: *https://arxiv.org/abs/2403.17536*

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4): e1253.

Zhou, Z.-H. and Zafarani, R. (2020). A survey of weak supervision for learning with noisy labels, *ACM Computing Surveys (CSUR)* **53**(3): 1–35.