Using Information Fusion to Predict Customer Sentiment*

Mónica Martins[1], Filipe Santos[1], Alexandre Almeida[2], Susana M. Vieira[1], and João M. C. Sousa[1]

[1] IDMEC, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
{monicamartins,filipempsantos,susana.vieira,jmsousa}@tecnico.ulisboa.pt
[2] AXIANSEU II DIGITAL CONSULTING, S.A., Lisbon, Portugal
alexandre1.almeida@axians.com

Abstract. Sentiment analysis is a common approach to measuring customer satisfaction, which plays a critical role in customer retention. While traditional approaches rely solely on textual reviews to predict customer sentiment, in this paper, a novel framework is proposed to expand on the traditional approaches, by adding additional data to textual data. The developed fusion framework uses decision-level fusion architecture to combine predictions from BERT (Bidirectional Encoder Representations from Transformers) and Extreme Gradient Boosting (XGBoost). The E-Commerce Clothing Reviews dataset was chosen to evaluate the proposed fusion approach, as it is a real use-case and for having both textual and numerical data. To establish a baseline, individual assessments were carried out for each data type: BERT was trained on textual data, while XGBoost was trained on numerical data. The results demonstrate the promising performance of the decision-level fusion framework, outperforming BERT with a macro-average F1 score of 86.65%.

Keywords: Sentiment Analysis · Information Fusion · Decision-Level Fusion · Customer Feedback · Customer Reviews · Large Language Models · BERT

1 Introduction

E-commerce has fundamentally reshaped the interactions between businesses and customers, more evidently during the COVID-19 pandemic, and also on a smaller scale during the remainder of the last decade. In 2022, e-commerce represented 19% of retail sales worldwide. Moreover, it is expected to reach 23% of total global retail sales in 2027, according to [9].

One key advantage of e-commerce platforms lies in their interactive nature, allowing customers to rate purchased products or services, and to write reviews on them. Some platforms even offer extra flexibility, allowing users to attach pictures to their reviews. Data in these formats is in high demand, as both customers and companies benefit from it. Customers can make better-informed

* Supported by LAETA and FCT.

decisions on what to buy, while companies have plenty of feedback to add to their sales data, which in turn can be used to better position themselves on the market.

A critical factor for any business is how they deal with negative feedback, whether it is through poor ratings or reviews or, more drastically, in the case of customers asking for refunds or replacements. Efficient and reliable customer service plays a crucial role in having loyal customers and maintaining a good public image [7]. Although identifying dissatisfied customers is crucial to retaining them, identifying satisfied customers is also advantageous. Satisfied customers are more likely to return; therefore, designing marketing campaigns specifically targeted at them may lead to higher revenue [6].

In sum, identifying both satisfied and unsatisfied customers has notorious benefits for companies. In the past, this analysis would be done manually. Nowadays, such analysis can be done automatically using various Machine Learning (ML) techniques. This framework is called Sentiment Analysis (SA).

Several studies show how reliable ML models are at identifying customer satisfaction from text [10, 18, 3, 2, 12, 4, 11, 1]. However, the integration of additional data into this task may further enhance the performance of ML models, as each additional modality may contain complementary or completely new information, when compared to that of text [15]. How to synergize all types when building a ML model is the object of study of the Information Fusion field [20]. Integrating various data types enhances the performance capabalities of ML models. However, it adds to the already complex process of designing a ML framework: not only must the framework extract important information from each modality, but it must also effectively fuse such information.

In this paper, we propose a Multimodal SA framework to predict customer sentiment. The proposed approach expands on the traditional work in the field, by introducing both numerical and categorical data to the SA framework.

The remainder of this paper is organized as follows: Section 2 presents an overview of related works; Section 3 presents the BERT model used for SA and details the proposed architecture; Section 4 outlines the experimental setup, results, and discussion; Section 6 presents the conclusions and suggestions for future work.

2 Related Work

2.1 State-of-the-art in Natural Language Processing

Transformer-based Large Language Models (LLMs) have been reshaping the field of Natural Language Processing (NLP) since around 2018 [16], with the introduction of models such as Generative Pretrained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT). These models achieved state-of-the-art results at the time, by being trained in two phases: firstly, models are pre-trained on large amounts of unlabeled data, to then be fine-tuned on smaller task-specific labeled datasets.

Since then, there has been a notable progression in the development of larger and more complex language models, as well as their refined iterations. For instance, BERT underwent various adaptions, such as RoBERTa, DistilBERT, and ALBERT, each optimized with specific challenges in mind. Another noteworthy example of improving on the original BERT is XLNet.

2.2 Textual Customer Sentiment Analysis

The methodologies found in the literature for customer sentiment analysis all use text as input. However, they differ in three distinct parts: the textual feature extraction, the machine learning method, and the conservation of ratings to targets for the ML model.

In regards to how features are extracted from text, previous studies have used two types of methods: those who count the frequency of tokens in a sample, such as Bag of Words [10, 18] and TF-IDF [10, 18, 19]; word embedding methods such as FastText [18, 3] and GloVe [2].

Regarding ML methods, several algorithms have been used for customer SA, such as Decision Trees [10], Random Forest [10, 19], and K-Nearest Neighbour (KNN) [19].

In addition to those, frameworks based on deep learning models have also been explored, such as Graph Neural Networks [12], and Recurrent Neural Networks with Gated Recurrent Units (GRU)[4], as well as Long Short Term Memory (LSTM) [11].

More recent studies make use of pre-trained LLMs. AraBERT, the Arabic version of BERT [1] was shown to outperform CNN-LSTM models. Additionally, the study discussed in [3] extensively compares classical machine learning and deep learning models. The main conclusions suggest that LLMs such as BERT, XLNet, DistilBERT, and RoBERTa, outperform classical methods on text-based sentiment analysis tasks.

Finally, regarding the target used, the target is usually extracted from the ratings associated with the reviews [10, 18, 2], most commonly from 1 to 5 stars, even though approaches vary from study to study. Studies agree that ratings of one and two stars correspond to negative sentiment, whereas ratings of 4 and 5 correspond to positive sentiment. However, 3-star ratings are treated differently across studies, being negative in a binary classification problem [18], neutral in a multiclass classification problem [2], or removed from the dataset to keep the problem binary [10].

2.3 Information Fusion

There are three main data fusion strategies: feature-level fusion, decision-level fusion, and model-level fusion [20]. Besides these three, some authors also consider a fourth strategy called hybrid fusion, which includes all combinations between any assortment of 2 or more distinct strategies.

Feature-level fusion involves concatenating multiple feature vectors, corresponding to the same sample, into a single one before applying a ML algorithm.

This method is mostly used for unimodal fusion, where concatenation is straightforward. The major drawback of this strategy is that it struggles with conflicting information and faces the curse of dimensionality.

A decision-level fusion strategy, in turn, consists of training a different model for all distinct sources of information available. Then, the predictions of said models go through a fusion operator, to obtain a final prediction. A common example of implementing such a strategy is the ensemble classifier [13]. One of the main advantages of this strategy is its flexibility, as it can deal with both unimodal fusion and multimodal fusion. Nevertheless, its main drawback is that not all frameworks built with this architecture can be trained end-to-end [17].

Finally, model-level fusion involves having a model with several distinct input vectors, where each input vector is initially processed individually, to then be concatenated with the remaining input vectors, within the model. Ultimately, the model generates a single prediction by weighing all input vectors. The main advantages of this strategy rely on the model being trained end-to-end while offering flexibility and potential for capturing true cross-modal relations. However, this characteristic makes it neural network exclusive. Moreover, this method is usually associated with intensive hyperparameter-tuning [14].

3 Proposed Methodology

This section presents an overview of the BERT model that was fed with the textual data, as well as the proposed information fusion architecture, that is, the decision-level fusion architecture. It is designed to receive textual, categorical, and numerical features as input. However, the conceptual framework outlined can be easily extended to include additional types.

3.1 A brief overview over BERT

BERT was introduced in 2018 [8] by Google. It is an encoder-only transformer-based LLM that can perform several NLP tasks. BERT's flexibility can be attributed to its two main components: the BERT-encoder, which is transversal across all NLP tasks, and a task-specific set of heads that can be swapped out depending on which task is being performed.

In this section, the focus will be on describing BERT's architecture, which is composed of the encoder, the pooler layer, and the classification head, focusing on how the data from the BERT encoder is fed into the classification head.

The output of the BERT encoder is the last hidden state of the encoder, where the size of the hidden state is 768 for each token in the sequence fed into BERT. In other words, upon feeding an N-token sequence to BERT, the encoder output will have a size of N×768.

Then, the encoder output is fed into the classification head. The first layer the data will go through is a pooling layer, responsible for dropping the size from N×768 to 768. From there, the data goes through a fully-connected layer with tanh as its activation function, a dropout layer, and an output layer, usually

with a softmax activation function. The architecture of BERT with a Sequence Classification Head is shown in Fig.1.

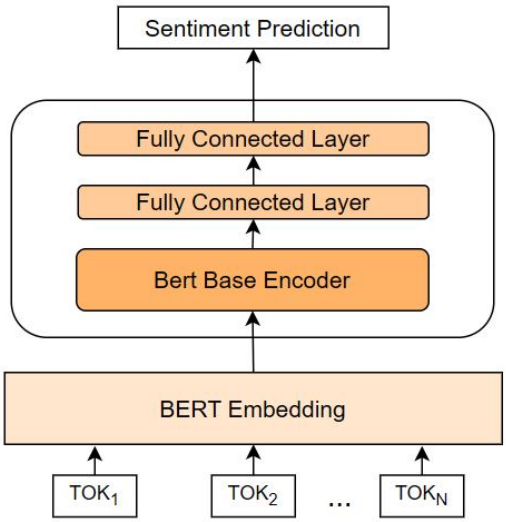

Fig. 1. Model architecture of BERT for sentiment analysis.

3.2 Decision-Level Fusion Architecture

For this architecture, two models need to be trained independently. For the numerical and categorical features, we propose using XGBoost, as it is a computationally light method, known to work well with tabular data. For textual features, fine-tuned BERT was used, as detailed in section 3.1.

Having both models trained on their specific types, the predictions of both models can be fused, using decision-level fusion. The method used to fuse both predictions was a weighted average. The architecture of the method is shown in Fig.2. The weights used on the weighted average were obtained during training.

4 Results and Discussion

This section describes the dataset used for experimental validation, the preprocessing of the data, the main results, and its discussion.

4.1 Data Description

The dataset used to test the proposed architectures was the Women's E-Commerce Clothing Reviews dataset [5], an E-commerce platform. It includes customer reviews, age information, and several categorical data that identify the purchased

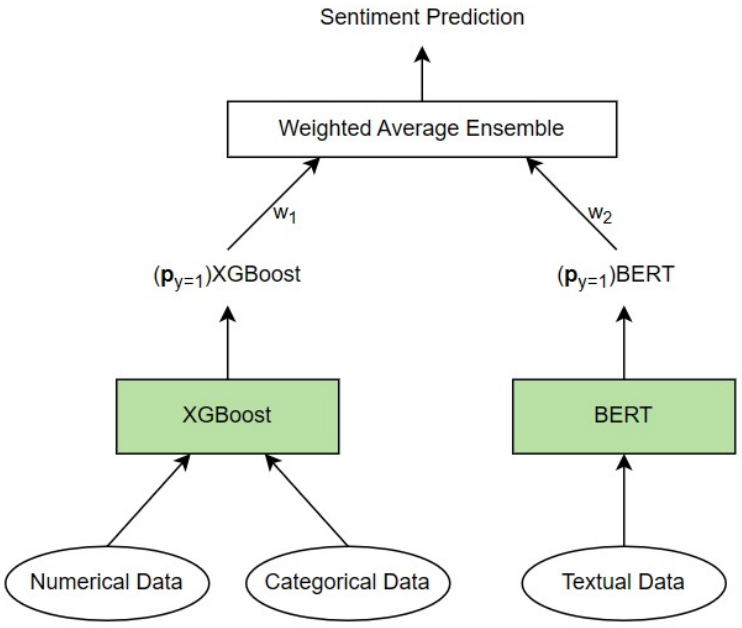

Fig. 2. Decision-level fusion with weighted average ensemble learning.

item. The dataset is also anonymous, with any references to names having been removed, and references to the company's name were replaced with "retailer".

In summary, this dataset is composed of three different data types, represented by the following dataset features:

- Categorical data: `Clothing ID`, `Rating`, `Recommended IND`, `Division Name`, `Department Name` and `Class Name`.
- Numerical data: `Age` and `Positive Feedback Count`.
- Textual data: `Review Text` and `Title`.

Due to the presence of both the `Recommended IND` and `Rating` features in the Women's E-Commerce Clothing Reviews dataset, two main tasks can be performed:

- Recommendation prediction: predict whether the customer recommends the product, where the target label is `Recommended IND`.
- Satisfaction prediction (also referred to as sentiment analysis): predict whether the customer is satisfied with the product, where the target label is inferred from the `Rating`.

4.2 Data Preparation and Feature Engineering

The dataset was split into training, validation, and test set, using a stratified train-test split with a ratio of (0.6/0.2/0.2).

Labels In this paper, the focus is on customer SA. Therefore, the target label was inferred from the `Rating` feature. The classification task labels were formulated as follows:

- 4-star and 5-star ratings were encoded as 1, that is, positive sentiment;
- 1-star, 2-star and 3-star ratings were encoded as 0, that is, negative sentiment.

With this formulation, the problem becomes binary. However, the classes are unbalanced (22. 95% for label 0 and 77.05% for label 1).

Textual Data Regarding textual features, the `Title` feature was not utilized for model training due to the high prevalence of missing data.

As BERT is the model used to process text, the first step is to tokenize the input text using the BERT tokenizer - in this case, the uncased version of the `BertTokenizer` was chosen. During this process, samples with fewer tokens are padded with [PAD] tokens, so that all sequences have the same amount of tokens. In this process, sequences with more tokens than the predefined sequence limit are truncated. This allows data to be fed in batches during training, increasing computational performance. However, if all sequences are far smaller than the sequence limit, the training process will become unnecessarily more computationally demanding.

The `BertTokenizer` maximum sequence length is 512. To address potential computational inefficiencies, the text was tokenized without truncation and padding, to infer the sequence size distribution across samples. The results are shown in Fig. 3. By analyzing the figure, one can conclude that most reviews have fewer than 150 tokens. Thus, a limit of 150 tokens was defined, and tokenization was repeated with both padding and truncation. It should be noted that the tokenization process adds the [CLS] (Classification), [SEP] (Separation), and [PAD] (Padding) tokens.

Numerical and Categorical Data Finally, in terms of numerical and categorical data, two numerical features were derived from the body of the reviews, namely, the number of question marks and the number of exclamation points.

Moreover, the `Recommended IND` was not excluded from the numerical features utilized for model training, as it represents potential data leakage problems.

Numerical features were normalized with the Standart Scaler, mathematically represented by Equation 1.

$$z = \frac{(x - u)}{s}\,, \tag{1}$$

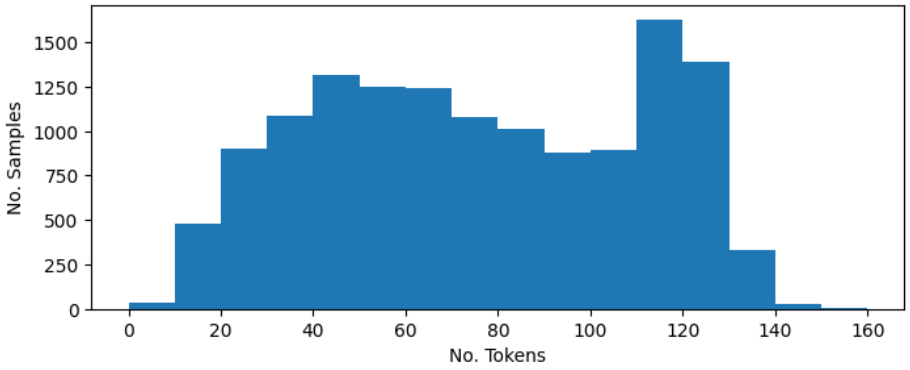

Fig. 3. Distribution of the number of tokens per review in the Women's E-Commerce Clothing Reviews dataset.

where x represents the sample, u represents the mean of the training samples, and s is the standard deviation of the training samples.

Finally, categorical features were encoded using one-hot encoding.

4.3 Results

The proposed fusion framework was trained 5 times. In parallel, to draw baselines, a neural network and the XGBoost model were trained on only numerical and categorical data; BERT was trained on textual following the same procedure. All these tests were run on an NVIDIA GeForce RTX 3050 graphics card. The obtained results are presented in Table 1.

The main metric that was taken into account is the macro-average F1-score since the goal is to obtain a good balance between true positives and true negatives. Nevertheless, other metrics were also considered, such as Balanced Accuracy, Matthew's Coefficient, Cohen's Kappa, Area Under the ROC Curve (AUC-ROC), Area Under the Precision-Recall Curve (AUC-PR), Sensitivity, Specificity, and Accuracy, as these offer complementary information on the model predictions.

From the inspection of Table 1, it becomes evident that among the individual models, BERT consistently outperforms XGBoost, indicating that the textual features encapsulate more sentiment than the other features. In other words, this demonstrates that categorical and numerical features do not convey as much sentiment as textual features.

Regarding the proposed decision-level fusion architecture, it can be stated that it achieved promising results, outperforming both the other models. Although XGBoost by itself had a far weaker performance than BERT, within the fusion framework, it rectified some misclassifications made by BERT, while not hindering its general performance. This was accomplished by assigning a minimal weight (specifically, 0.3) to XGBoost's predictions, thereby influencing the

Metric	BERT	XGBoost	D.-L. Fusion
Macro F1	86.64±0.59	53.65±0.61	86.65±0.65
Accuracy	90.38±0.43	56.61±0.64	90.39±0.47
Balanced Acc.	87.76±1.06	61.14±0.93	87.76±1.08
Matthew's Coef.	73.45±1.21	18.65±1.53	73.45±1.33
Cohen's Kappa	73.30±1.19	15.22±1.18	73.31±1.31
AUC-ROC	95.47±0.33	65.52±0.65	92.57±1.07
AUC-PR	98.58±0.10	85.95±0.53	97.02±0.50
Sensitivity	92.55±0.89	52.88±0.83	92.56±0.87
Specificity	82.98±2.76	69.40±1.91	82.96±2.73

Table 1. Mean and Standard Deviation for each metric.

final prediction only when BERT's predictions approached the decision threshold of 0.5. The empirical optimization of weights assigned to each unimodal model's output facilitated precise adjustments to their contributions within the ensemble.

5 Conclusion

This study aimed to effectively combine different data modalities to predict customer sentiment. This paper proposes a decision-level fusion framework for such task. Within it, BERT is used to make predictions on customer sentiment based on textual data, while XGBoost generates predictions based on numerical features. The framework then weighs both of them to generate an ultimate prediction.

This framework was tested on the Women's E-Commerce Clothing Reviews dataset as it contains real-world data. The results show that BERT by itself performs substantially better than XGBoost, as BERT was trained on actual product reviews, while XGBoost was trained on numeric data, which are mainly descriptive of which article was being reviewed. Nevertheless, the proposed framework showed promising results, outperforming BERT. By weighing XGBoost predictions into the framework decision, the framework was able to adjust BERT predictions, and consequently improve them, whenever these were close to the decision threshold.

Future works should encompass the inclusion of other datasets in which numerical features are more meaningful for the task at hand. It is anticipated that the proposed framework will perform vastly better under such a scenario. In terms of modeling, two primary strategies should be tested: hierarchical fusion, involving the initial integration of bimodal features followed by the incorporation of all modalities, and the refinement of decision-level fusion. The latter can be achieved through the exploration of alternative fusion operators, such as max-fusion, or thorough the utilization of a metalearner.

Acknowledgements

The authors acknowledge Fundação para a Ciência e a Tecnologia (FCT) financial support via the projects LAETA Base Funding (DOI: 10.54499/UIDB/50022/2020) and LAETA Programatic Funding (DOI: 10.54499/UIDP/50022/2020) and Filipe Santos' work was supported by the Ph.D. Scholarship 2022.12077.BDANA from FCT.

References

1. Aftan, S., Shah, H.: Using the arabert model for customer satisfaction classification of telecom sectors in saudi arabia. Brain Sciences **13**(1), 147 (2023)
2. Agarap, A.F.: Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn). arXiv preprint arXiv:1805.03687 (2018)
3. Alantari, H.J., Currim, I.S., Deng, Y., Singh, S.: An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. International Journal of Research in Marketing **39**(1), 1–19 (2022)
4. Alshamari, M.A.: Evaluating user satisfaction using deep-learning-based sentiment analysis for social media data in saudi arabia's telecommunication sector. Computers **12**(9), 170 (2023)
5. Brooks, N.: Women's e-commerce clothing reviews, https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews, last accessed 19 February 2023
6. Camilleri, M.A.: Travel marketing, tourism economics and the airline product: An introduction to theory and practice. Springer (2018)
7. Costumer Care Management & Consulting (CCMC): 2020 national customer rage study (2020), https://customercaremc.com/insights/national-customer-rage-study/2020-national-customer-rage-study/, last accessed on 19 September 2023
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Graph: E-commerce as percentage of total retail sales worldwide from 2015 to 2027 (2023), https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/, last accessed on 19 September 2023
10. Haque, T.U., Saber, N.N., Shah, F.M.: Sentiment analysis on large scale amazon product reviews. In: 2018 IEEE international conference on innovative research and development (ICIRD). pp. 1–6. IEEE (2018)
11. Iqbal, A., Amin, R., Iqbal, J., Alroobaea, R., Binmahfoudh, A., Hussain, M.: Sentiment analysis of consumer reviews using deep learning. Sustainability **14**(17), 10844 (2022)
12. Kanchinadam, T., Meng, Z., Bockhorst, J., Singh, V., Fung, G.: Graph neural networks to predict customer satisfaction following interactions with a corporate call center. arXiv preprint arXiv:2102.00420 (2021)
13. Kuncheva, L.I.: Combining pattern classifiers: methods and algorithms. John Wiley & Sons (2014)
14. Li, W., Peng, Y., Zhang, M., Ding, L., Hu, H., Shen, L.: Deep model fusion: A survey. arXiv preprint arXiv:2309.15698 (2023)

15. Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., Poria, S.: Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based systems **161**, 124–133 (2018)
16. Patwardhan, N., Marrone, S., Sansone, C.: Transformers in the real world: A survey on nlp applications. Information **14**(4), 242 (2023)
17. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine **34**(6), 96–108 (2017)
18. Shah, A.: Sentiment analysis of product reviews using supervised learning. Reliability: Theory & Applications **16**(SI 1 (60)), 243–253 (2021)
19. Turdjai, A.A., Mutijarsa, K.: Simulation of marketplace customer satisfaction analysis based on machine learning algorithms. In: 2016 International Seminar on Application for Technology of Information and Communication (ISemantic). pp. 157–162. IEEE (2016)
20. Wu, C.H., Lin, J.C., Wei, W.L.: Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA transactions on signal and information processing **3**, e12 (2014)