ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

CP-IMOS: A Cross-Platform Imbalance-Aware Methodology for Sentiment Classification of MOOC Reviews in IT Education

Melissa T. Guillermo[1]*, Reagan B. Ricafort[2]

[1]Dr. Filemon C. Aguilar Memorial College of Las Piñas, Las Piñas City, Philippines

[2]AMA University, Philippines

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600015

Received: 17 June 2026; Accepted: 22 June 2026; Published: 03 July 2026

ABSTRACT

MOOC review datasets offer a practical way to study course strengths, learner frustrations, and platform-level issues. However, sentiment models built from these data can give a distorted picture when positive ratings dominate the corpus. This paper presents CP-IMOS, a Cross-Platform Imbalance-Aware MOOC Sentiment Classification methodology, using public review data from Coursera, Udemy, and Stepik. The Coursera file contained 1,454,711 review rows; after cleaning and exact duplicate removal, 498,401 review records remained. Keyword filtering of course identifiers and titles then yielded 258,828 IT-related Coursera reviews from 215 course identifiers. For Udemy, a 200,000-comment Kaggle sample was processed with Course_info.csv to support second-platform validation. After cleaning, 46,377 comments from Development and IT & Software courses were retained from 15,836 course identifiers. The Stepik Russian-language corpus was included to test multilingual ingestion, duplicate removal, and Cyrillic-character validation during retrieval, but it was excluded from supervised sentiment training because it did not provide rating or sentiment labels. CP-IMOS maps rating-derived labels into five classes that correspond to ratings 1 through 5 and adds the Platform Sentiment Imbalance Index (PSII) to measure majority-class dominance before model interpretation. The classifier stage uses TF-IDF features with transparent machine-learning components. In the Coursera analysis sample, SGD Logistic Regression obtained the highest macro-F1 (0.383), while Multinomial Naive Bayes produced higher accuracy (0.759) but a weaker macro-F1 (0.250). In the Udemy validation set, SGD Logistic Regression also produced the highest macro-F1 (0.430). The findings show that cross-platform MOOC review data are strongly skewed toward positive ratings, although Udemy contains more non-positive ratings than Coursera. The main contribution is a reproducible retrieval-to-classification workflow that uses PSII, macro-F1, per-class metrics, and platform-specific interpretation before sentiment results are used in IT education analytics.

Keywords: CP-IMOS; MOOC reviews; sentiment classification; IT education; class imbalance

INTRODUCTION

Massive Open Online Courses collect a steady stream of learner feedback through star ratings and written reviews. These records can help teachers, platform teams, and researchers see which course elements are working and where learners encounter difficulty. In IT education, the comments are especially valuable because learners often discuss programming activities, installation problems, laboratory instructions, debugging support, assessment clarity, software versions, and links to career goals.

The present work uses three public MOOC-related sources: Coursera, Udemy, and Stepik. Coursera and Udemy provide the rating-based evidence for sentiment classification. Stepik, a Russian-language corpus, is used more narrowly to document the multilingual retrieval part of the workflow. This separation keeps the study's claims tied to the evidence available in each source.

Rather than arguing that a single universal model can represent all MOOC platforms, the paper develops a practical workflow for retrieving, cleaning, filtering, labeling, and evaluating review data across more than one

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

source. This matters because course metadata, rating scales, language coverage, and review-writing practices vary by platform. A method that appears adequate on one dataset may behave differently when transferred to another.

The study makes the following contributions:

1. It presents CP-IMOS as an end-to-end, cross-platform, imbalance-aware methodology for IT MOOC review sentiment classification.
2. It defines the Platform Sentiment Imbalance Index (PSII) as a simple measure of majority-class dominance before model results are interpreted.
3. It brings together platform-specific IT-course retrieval, duplicate-aware cleaning, rating-derived label harmonization, and macro-F1-based classifier selection in a reproducible workflow.
4. It tests the workflow on Coursera and Udemy, while using Stepik to examine multilingual retrieval operations such as non-English text ingestion, duplicate removal, and Cyrillic-character validation.
5. It shows why classifier outputs in positive-skewed educational review data should be read through per-class metrics and confusion matrices, not accuracy alone.

Class imbalance is a central concern in this paper. Online course reviews often contain far more positive than negative ratings. Under that condition, a classifier can look accurate simply because it predicts the majority class most of the time. For course improvement, though, the smaller groups of negative, neutral, and moderately positive reviews may carry the most useful evidence. For that reason, the study reports macro-F1, per-class results, and confusion matrices alongside accuracy.

Research Questions

RQ1. What data profile is produced after retrieving, cleaning, and filtering Coursera and Udemy MOOC review datasets for IT education?

RQ2. How imbalanced are the rating-derived sentiment classes in the Coursera and Udemy IT-related subsets?

RQ3. How do the TF-IDF-based classifier components of CP-IMOS perform in five-class sentiment classification on each platform?

RQ4. What does the Stepik Russian-language corpus contribute to the multilingual retrieval scope of CP-IMOS, especially in terms of non-English text ingestion, duplicate removal, and Cyrillic character validation, and what limitations apply to its use?

LITERATURE REVIEW

MOOC review mining and educational analytics

MOOC review mining is now a common way to examine learner satisfaction, engagement, course quality, and platform experience. Prior work has used review data to support course selection, identify satisfaction drivers, summarize learner concerns at scale, and analyze feedback from different online learning environments [4], [5], [6], [16], [18], [19], [21]. Taken together, these studies show that reviews can support course improvement when they are processed systematically and interpreted with caution.

IT and computing courses add a distinctive vocabulary to this task. Learners may refer to code examples, installation steps, operating-system problems, cloud tools, database exercises, network configuration, and project-based assignments. These details make IT review mining useful, but they also require careful filtering so that the final corpus represents computing-related courses rather than a broad mix of unrelated subjects. Other educational analytics studies likewise show that textual and tabular learning data are most useful when metadata and learner context are made explicit [23], [24].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Sentiment classification and rating-derived labels

Sentiment classification assigns text to categories such as positive, negative, or neutral. In many MOOC datasets, however, the available label is a learner's numerical rating rather than a manually annotated sentiment category. This paper therefore treats the task as rating-derived sentiment classification. The distinction is important because course reviews can contain ambiguous, indirect, or mixed reactions that a star rating may not fully capture [17], [26], [27].

Recent educational-review studies have used transformer models, aspect-based sentiment analysis, attention mechanisms, and hybrid approaches [7], [8], [9], [10], [18], [20], [25]. These methods can be powerful, but they also need more computing resources and careful validation. In this paper, TF-IDF representations and classical machine-learning classifiers are used as transparent components within CP-IMOS rather than as isolated baseline exercises. This choice keeps the workflow repeatable with standard resources and leaves room for transformer-based extensions after the dataset, labels, and platform scope are stable [17], [23].

Cross-platform and multilingual issues

Cross-platform review analysis is not straightforward because the same numerical rating may carry different meanings across platforms. A four-star Coursera review and a four-star Udemy comment may differ in length, tone, learner expectation, course progress, or course context [22]. The metadata also differ: Coursera records are linked to course identifiers and course names, while Udemy comments must be joined with course information before category filtering can be applied. Because educational text analytics can draw on several data types, these platform-specific metadata decisions need to be stated before model results are compared [24].

Multilingual data create an additional validity issue because language resources, annotation availability, and label comparability affect what can be claimed [11], [12], [13], [28]. Here, the Stepik corpus is treated as a Russian-language retrieval test case rather than as supervised sentiment evidence. Its role is to check whether CP-IMOS can ingest, clean, de-duplicate, and validate a Cyrillic MOOC review corpus while keeping unlabeled data separate from the labeled Coursera and Udemy experiments.

Evaluation metrics for imbalanced classification

Accuracy is familiar and easy to report, but it is weak evidence when one class dominates. In strongly positive-skewed review data, a model can achieve high accuracy by predicting the very positive class too often. Macro-F1 is more useful for this setting because it gives equal weight to each class, including the smaller negative and neutral groups [14], [15], [29], [30]. The best model in this study is therefore selected by macro-F1 rather than by accuracy alone.

Analytical Framework

CP-IMOS is organized into eight stages. Public review datasets are first retrieved from documented sources. The files are then inspected for columns, rating fields, text fields, and course metadata. Review text is cleaned by trimming spaces, standardizing repeated whitespace, and removing exact duplicates with platform-appropriate keys. IT-related records are selected through metadata suited to each platform. Where labels are available, ratings are mapped into a common five-class sentiment scale. PSII is then computed as a majority-versus-non-majority ratio to diagnose class dominance before model results are interpreted. TF-IDF features and transparent classifier components are trained and evaluated with macro-F1, per-class results, and confusion matrices. Finally, the platform results are compared, while unlabeled multilingual corpora are used only for retrieval-stage validation.

This design is intentionally conservative. Coursera and Udemy support supervised sentiment classification because they contain both review text and ratings. Stepik supports only multilingual retrieval-stage validation in the present study because it contains Russian-language review text but no rating or sentiment field. The framework can therefore test character-set validation and non-English corpus handling without making unsupported multilingual sentiment-classification claims.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

METHODOLOGY

Research design

The study followed an empirical model-training design with cross-platform validation. Coursera served as the main large-scale dataset, Udemy served as the second-platform validation dataset, and Stepik served as the multilingual retrieval extension. For supervised modeling, the unit of analysis was one written review or comment paired with a rating-derived sentiment label.

Proposed CP-IMOS methodology

CP-IMOS, or Cross-Platform Imbalance-Aware MOOC Sentiment Classification, was developed for IT education reviews. The methodology addresses three recurring problems in MOOC review mining: differences in platform metadata, imbalance in rating-derived classes, and model selection that can be misleading when it relies on accuracy alone. It combines platform-sensitive retrieval, duplicate-aware cleaning, rating harmonization, imbalance diagnosis, transparent classifier components, macro-F1-based model selection, and second-platform validation.

Platform Sentiment Imbalance Index

To measure platform-level class dominance, CP-IMOS introduces the Platform Sentiment Imbalance Index (PSII). For a platform p with K sentiment classes, n(p,c) is the number of reviews in class c, and N(p) is the total number of labeled reviews across all K classes. The majority-class count is Nmajority(p) = max n(p,c). The non-majority total is Nnonmajority(p) = N(p) - Nmajority(p), which combines all classes except the majority class. PSII is defined as PSII(p) = Nmajority(p) / Nnonmajority(p). The denominator is therefore the sum of the non-majority classes, not the full dataset size. A higher PSII signals stronger one-class dominance and a greater risk that accuracy will overstate classifier performance.

With this definition, the Coursera PSII is 190,979 / (258,828 - 190,979) = 190,979 / 67,849 = 2.82. The Udemy PSII is 25,883 / (46,377 - 25,883) = 25,883 / 20,494 = 1.26. Reporting the calculation makes clear that the majority class is being compared with the combined non-majority classes.

Algorithm 1. CP-IMOS cross-platform imbalance-aware methodology

Input: Platform review datasets, course metadata, review text, and rating fields.
Output: Platform-specific sentiment classifiers, imbalance diagnostics, and cross-platform validation results.
Step 1. Retrieve public MOOC review datasets and document their source roles.
Step 2. Inspect text fields, rating fields, course identifiers, and course metadata.
Step 3. Normalize review text and remove exact duplicates.
Step 4. Select IT-related courses using platform-specific filtering.
Step 5. Harmonize ratings into a five-class sentiment scale.
Step 6. Compute class distribution and PSII for each labeled platform.
Step 7. Extract TF-IDF unigram and bigram features.
Step 8. Train transparent classifier components with imbalance-aware settings where applicable.
Step 9. Select the strongest component using macro-F1 as the primary criterion.
Step 10. Diagnose per-class precision, recall, F1-score, and confusion-matrix errors.
Step 11. Validate the workflow on a second platform.
Step 12. Treat unlabeled multilingual data as retrieval evidence only unless verified labels are available.

Data sources

The Coursera data were taken from the Kaggle dataset Course Reviews on Coursera, which includes Coursera course metadata and about 1.45 million review records [1]. The Udemy data came from the Kaggle dataset Udemy Courses, which provides course information and user comments [2]. A 200,000-comment Udemy sample

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

was processed with the full Course_info.csv file for second-platform validation. The Stepik data came from the Mendeley Data corpus Dataset of MOOCs Reviews from Stepik on Russian Language [3].

For Coursera, the first validity question was whether the retrieved corpus remained analytically defensible after cleaning and domain filtering. Table I therefore documents the audit trail from the raw review file to the IT-related subset, including validity checks, exact duplicate removal, metadata linkage, and keyword-based course selection.

Table I. Coursera dataset profile after cleaning and IT-course filtering

Profile item	**Value**
Original Coursera review rows	1,454,711
Rows after valid text/rating checks	1,454,708
Clean Coursera rows after exact duplicate removal	498,401
Exact duplicate rows removed	956,307
Coursera course metadata rows	623
IT-related Coursera review rows	258,828
IT-related Coursera course identifiers	215
IT filter used	keyword search on course_id, course title, and institution metadata

The drop from 1,454,711 raw Coursera rows to 498,401 clean rows shows that duplicate removal was a major validity control. If the 956,307 exact duplicates had been retained, the analysis would have overstated the amount of independent learner evidence and could have distorted the sentiment distribution. The final IT-related subset, 258,828 reviews from 215 course identifiers, is large enough for modeling, although the keyword-based filter still requires caution because it may include borderline computing-related courses and miss courses that use less explicit technical terminology.

The Udemy evidence is not treated as directly equivalent to Coursera. Table II gives the corresponding profile for the sampled Udemy comments because the Udemy component is a second-platform validation set rather than a census of all Udemy comments. Its IT-course identification also relies on platform-native categories rather than the keyword procedure used for Coursera.

TABLE II. Udemy sampled dataset profile after cleaning and IT-course filtering

Profile item	**Value**
Original sampled comment rows	200,000
Rows after valid text/rating checks	199,999
Clean sampled rows after exact duplicate removal	199,875

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Exact duplicate rows removed	124
Course metadata rows	209,734
IT-related sampled comment rows	46,377
IT-related course_id values	15,836
IT filter used	category in Development or IT & Software

The Udemy profile differs from the Coursera profile in an important way: only 124 exact duplicate comments were removed from the 200,000-comment sample. After category filtering, 46,377 IT-related comments remained. This provides a useful validation corpus, but it should not be used to make claims about the full Udemy review environment. The official Development and IT & Software categories strengthen construct validity for identifying IT-related courses. At the same time, because the Coursera and Udemy filters are structurally different, cross-platform comparisons should be interpreted as evidence of methodological transferability rather than strict platform equivalence.

The multilingual component needs separate treatment because retrieval coverage and supervised sentiment classification are different claims. Table III profiles Stepik as a Russian-language retrieval extension and records the cleaning, duplicate-removal, and Cyrillic-character checks used to see whether CP-IMOS can handle a non-English MOOC review corpus.

Table III. Stepik Russian-language corpus profile

Profile item	**Value**
Raw Stepik review rows	5,721
Rows with non-empty review text	5,719
Unique review texts after duplicate removal	5,091
Reviews containing Cyrillic characters	5,058
Use in this study	multilingual retrieval extension only; not supervised sentiment training because rating/sentiment labels were not included

The Stepik profile shows that CP-IMOS can process a non-English MOOC review source: 5,091 unique review texts remained after duplicate removal, and 5,058 contained Cyrillic characters. Its contribution is methodological rather than predictive. It tests Russian-language retrieval handling and shows why an unlabeled multilingual corpus must be kept separate from the supervised Coursera and Udemy classification tasks.

Cleaning, filtering, and label harmonization

For Coursera, the review text, rating, and course identifier fields were retained. Review text was cleaned by removing leading and trailing spaces and standardizing repeated whitespace. Exact duplicates were removed using course_id, cleaned review text, and rating. Course metadata were then joined to the review file. The IT subset was selected through computing-related keywords in the course identifier, course title, and institution metadata, producing 258,828 IT-related reviews from 215 course identifiers. Integer ratings were mapped directly to sentiment labels: 1 = very negative, 2 = negative, 3 = neutral, 4 = positive, and 5 = very positive.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

For Udemy, the comment sample was joined with course metadata through the course identifier. The IT-related subset used the platform categories Development and IT & Software. This narrower category filter was chosen instead of a broad keyword search because Udemy's course categories provide a clearer native signal. Ratings in 0.5 increments were mapped into five classes: 0.5-1.5 as very negative, more than 1.5 to 2.5 as negative, more than 2.5 to 3.5 as neutral, more than 3.5 to 4.5 as positive, and more than 4.5 as very positive.

For Stepik, the corpus was cleaned to retain non-empty and unique review text, and a Cyrillic-character check was applied as a simple script-level validation step. Because the file did not include a rating or sentiment field, Stepik was not converted into sentiment labels. It is reported only as retrieval-stage evidence for multilingual corpus handling.

CP-IMOS classifier components and model training

In CP-IMOS, TF-IDF representation and supervised classifiers serve as components of the proposed methodology rather than as stand-alone baseline experiments. The TF-IDF settings used lowercase conversion, English stop-word removal, unigram and bigram features, sublinear term-frequency scaling, minimum document frequency of 3, maximum document frequency of 0.95, and a 10,000-feature limit. These choices were made to keep the representation transparent and computationally manageable for sparse educational review text. Unigrams and bigrams preserve common single-word and short phrase cues, min_df=3 removes very rare terms, max_df=0.95 removes near-universal terms, sublinear_tf reduces the effect of repeated words, and the feature cap limits memory use while retaining a broad vocabulary. No systematic hyperparameter search or ablation study was conducted, so the results should be read as reproducible CP-IMOS component evidence rather than optimized classifier performance.

The full Coursera IT-related subset contained 258,828 reviews. Because that subset was large for the computational scope of the study, a stratified modeling sample of 99,999 reviews was drawn while preserving class proportions. This near-100,000 record cap was used to reduce memory and runtime requirements; it is a modeling sample of the full corpus, not an 80/20 split of all 258,828 records. The 99,999-review sample was then divided into 79,999 training rows and 20,000 test rows using a stratified 80/20 split. For Udemy, the full IT-related sample of 46,377 comments was used and divided into 37,101 training rows and 9,276 test rows. Both platforms used stratified 80/20 splitting and random_state = 42.

Evaluation strategy

The evaluation reports accuracy, macro precision, macro recall, macro-F1, weighted-F1, PSII, and confusion-matrix diagnostics. Macro-F1 is the main classifier-selection metric because the minority classes are important for educational analytics. For the strongest macro-F1 classifier component on each platform, the study also reports per-class results and confusion matrices. The cross-platform comparison uses held-out point estimates only. It does not include confidence intervals, bootstrap testing, or formal significance tests. Thus, the word consistent means that the same classifier component ranked highest by macro-F1 on both held-out test sets; it does not establish statistically significant superiority across platforms.

RESULTS AND DISCUSSION

Throughout this section, labels 1 through 5 correspond to: 1 = very negative, 2 = negative, 3 = neutral, 4 = positive, 5 = very positive.

Class distribution across platforms

Model results cannot be interpreted properly until the label structure is examined. Sentiment prevalence affects what accuracy, recall, and F1 scores actually mean. CP-IMOS therefore applies class distribution analysis and PSII before discussing classifier performance. From the reported counts, Coursera has PSII = 190,979 / 67,849 = 2.82, while Udemy has PSII = 25,883 / 20,494 = 1.26. Both platforms are positive-skewed, but Coursera shows

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

much stronger majority-class dominance. Table IV presents the Coursera label distribution and shows why the second and third research questions cannot be answered with accuracy alone.

Table IV. Class distribution of IT-related Coursera reviews

Label	Interpretation	Count	Percentage
1	Very negative	4,382	1.690
2	Negative	4,163	1.610
3	Neutral	11,575	4.470
4	Positive	47,729	18.44
5	Very positive	190,979	73.79

The Coursera distribution is strongly positive-skewed: 73.79% of IT-related reviews are labeled very positive, while the very negative, negative, and neutral classes together represent only 7.77%. This creates a measurement risk because a classifier may appear successful by learning the dominant positive class while missing the smaller categories. In IT education analytics, those smaller categories may contain comments about debugging support, unclear programming instructions, outdated tools, assessment design, or insufficient assistance, so weak recognition of these labels would reduce the model's practical value.

Table V extends the imbalance analysis to Udemy to see whether the Coursera pattern is platform-specific or also appears in a second MOOC environment. Presenting the Udemy distribution after the Coursera profile keeps the comparison focused on platform behavior and label ecology rather than assuming that the two datasets are interchangeable.

TABLE V. Class distribution of IT-related Udemy comments

Label	Interpretation	Count	Percentage
1	Very negative	4,619	9.960
2	Negative	3,024	6.520
3	Neutral	4,885	10.53
4	Positive	7,966	17.18
5	Very positive	25,883	55.81

Udemy is also positively skewed, but less sharply than Coursera: 55.81% of comments are very positive, and the non-positive classes make up a larger share. This difference matters because Udemy provides a more demanding validation context for minority-class recognition. If model behavior is similar across the two platforms, the workflow gains descriptive support; if it diverges, that would indicate that platform-specific rating cultures and metadata structures materially affect sentiment classification.

Figure 1 places the numerical distributions from Tables IV and V side by side. The comparison makes the shared positive skew visible while also showing that the magnitude of the skew differs between Coursera and Udemy.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

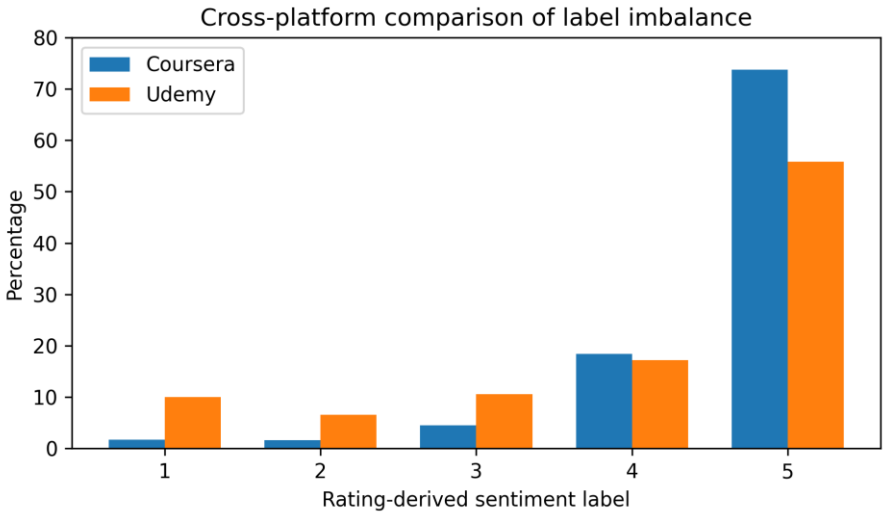

Fig. 1. Cross-platform comparison of rating-derived label imbalance.

Bars show the percentage distribution of five rating-derived sentiment labels for Coursera and Udemy IT-related subsets, where 1 = very negative, 2 = negative, 3 = neutral, 4 = positive, and 5 = very positive. The figure highlights that both platforms are positive-skewed, with Coursera showing stronger dominance of label 5 than Udemy.

Figure 1 shows that the main methodological issue is not a lack of data but uneven class representation. In Coursera, the very positive class dominates. As a result, high overall accuracy would not, by itself, demonstrate that the model can identify negative, neutral, or moderately positive feedback. This pattern supports the study's emphasis on macro-F1, per-class metrics, and confusion matrices, which are more appropriate when minority classes have educational significance.

The figure also supports the need for cross-platform validation. Udemy contains a larger proportion of non-positive ratings, so it tests whether the workflow remains informative under a less extreme but still imbalanced distribution. The comparison warns against presenting a universal MOOC sentiment model without first examining platform-specific rating practices, review-writing behavior, and metadata architecture.

Together, Tables IV and V and Fig. 1 answer the second research question: both platforms are imbalanced, but not to the same degree. This finding shapes the rest of the evaluation. A model selected only by accuracy would favor the majority class and could hide weak performance on the smaller classes that are most useful for course diagnosis and improvement.

Model performance on Coursera

Table VI reports the Coursera modeling results and compares conventional accuracy with imbalance-sensitive metrics. The comparison is central to the third research question because all four CP-IMOS classifier components used the same TF-IDF feature setting, making it possible to examine the trade-off between overall correctness and balanced class performance.

Table VI. Model performance on the Coursera IT-related test set

Model	**Accuracy**	**Macro precision**	**Macro recall**	**Macro-F1**	**Weighted-F1**
SGD Logistic Regression	0.742	0.394	0.410	0.383	0.715

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

SGD Linear SVM	0.713	0.345	0.406	0.323	0.686
Multinomial Naive Bayes	0.759	0.386	0.247	0.250	0.699
Majority-class reference	0.738	0.148	0.200	0.170	0.627

Table VI shows why accuracy is not sufficient for Coursera. Multinomial Naive Bayes has the highest accuracy at 0.759, but its macro-F1 is only 0.250. The majority-class reference already reaches 0.738 accuracy because the very positive class dominates the test set, leaving only a small accuracy gain for Multinomial Naive Bayes (0.759 versus 0.738). SGD Logistic Regression is therefore the stronger analytical choice even though its accuracy is slightly lower, because its macro-F1 of 0.383 reflects better performance across the full label structure.

Aggregate scores do not show whether the selected Coursera classifier identifies the categories most relevant to educational intervention. Table VII therefore reports performance by class, which is necessary for judging whether the model can support instructional decision-making or mainly reproduces the dominance of highly positive reviews.

Table VII. Per-class report for the best Coursera macro-F1 model

Class/average	**precision**	**recall**	**f1-score**	**support**
1	0.309	0.413	0.354	339.00
2	0.154	0.289	0.201	322.00
3	0.207	0.238	0.222	894.00
4	0.466	0.180	0.259	3,688.00
5	0.832	0.930	0.879	14,757.00
accuracy	0.742	0.742	0.742	0.742
macro avg	0.394	0.410	0.383	20,000.00
weighted avg	0.717	0.742	0.715	20,000.00

The per-class report indicates strong performance on label 5 but weak performance on labels 1, 2, 3, and 4. This is a substantive limitation. The model recognizes the dominant very positive category far better than it detects strong dissatisfaction, moderate dissatisfaction, neutral evaluation, or ordinary positive feedback. For instructors and MOOC administrators, this means the model should not be used alone for complaint detection or course-improvement prioritization without additional error analysis, resampling, calibration, or model refinement.

Table VIII adds the Coursera confusion matrix to show where the errors go. This diagnostic view is important because two models with similar aggregate scores may have very different practical implications depending on whether they confuse adjacent categories or absorb minority feedback into the dominant class.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table VIII. Confusion matrix of SGD Logistic Regression on Coursera

Actual class	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5
True 1	140	74	36	15	74
True 2	60	93	77	26	66
True 3	85	160	213	135	301
True 4	90	163	449	662	2,324
True 5	78	115	252	583	13,729

The confusion matrix shows that many true class-4 reviews are predicted as class 5, and many neutral reviews are also pushed toward more positive labels. This upward shift matters because it can make course feedback appear more favorable than the learner comments justify. In applied IT education analytics, such errors could hide concerns about programming exercises, software installation, debugging assistance, or assessment clarity.

Figure 2 summarizes the Coursera model comparison using macro-F1, the main imbalance-sensitive metric in the study. Placing the visual ranking after Tables VI to VIII reinforces the point that the preferred classifier under a balanced-class criterion is not necessarily the model with the highest raw accuracy.

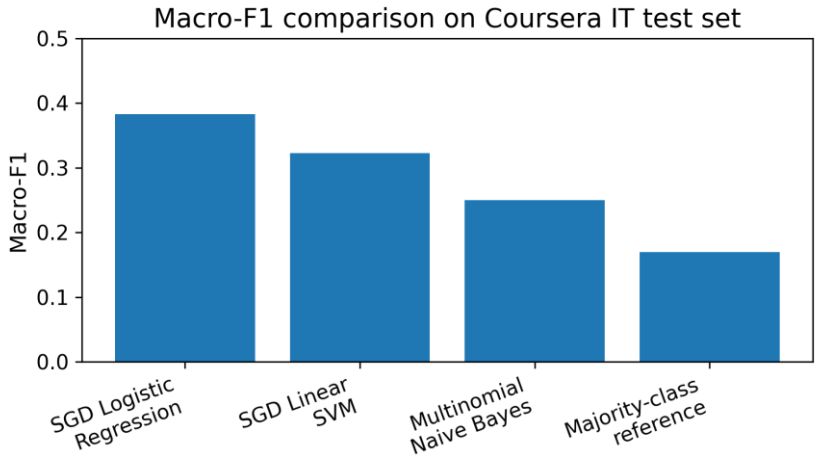

Fig. 2. Macro-F1 comparison of Coursera classifier components.

Bars show held-out macro-F1 scores for SGD Logistic Regression, SGD Linear SVM, Multinomial Naive Bayes, and the majority-class reference component on the Coursera IT test set. The figure shows that SGD Logistic Regression has the strongest macro-F1 even though Multinomial Naive Bayes has higher accuracy in Table VI.

Figure 2 supports the interpretation of Table VI. SGD Logistic Regression has the strongest Coursera macro-F1, while the majority-class reference performs poorly under a metric that gives each class equal weight. This result supports class-balanced linear models as transparent and reproducible CP-IMOS components. It should not, however, be taken as deployment readiness because several educationally important classes remain poorly detected.

Taken together, the Coursera results support CP-IMOS as an evaluation procedure, not as a fully reliable diagnostic instrument. The weak F1-scores for labels 1, 2, 3, and 4 mean that the classifier may fail in exactly the regions where intervention evidence is most valuable. Any institutional use of these outputs should therefore

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

include manual review of minority-class predictions and systematic inspection of misclassified comments before instructional decisions are made.

Overall, the Coursera results answer the third research question with caution. SGD Logistic Regression is the strongest tested Coursera component by macro-F1, but the modest minority-class performance shows that the framework exposes a central classification problem rather than solving it completely. This limitation motivates the Udemy validation before broader cross-platform claims are made.

Model performance on Udemy

Table IX applies the same modeling workflow to Udemy to examine whether the Coursera findings hold under a different platform distribution. Because Udemy has a less extreme positive skew, it serves as a validation context rather than a direct replication.

Table IX. Model performance on the Udemy IT-related test set

Model	Accuracy	Macro precision	Macro recall	Macro-F1	Weighted-F1
SGD Logistic Regression	0.625	0.447	0.431	0.430	0.595
SGD Linear SVM	0.608	0.411	0.423	0.395	0.573
Multinomial Naive Bayes	0.621	0.446	0.339	0.330	0.537
Majority-class reference	0.558	0.112	0.200	0.143	0.400

Table IX shows the same macro-F1 ranking pattern. SGD Logistic Regression again has the strongest point estimate, reaching 0.430 on Udemy, while Multinomial Naive Bayes has comparable accuracy but a weaker macro-F1. This repeated ranking supports the usefulness of macro-F1-based interpretation for imbalanced MOOC sentiment classification. It should still be read descriptively, not as a formal statistical test of superiority. The absolute values also remain modest, so the results are best understood as reproducible CP-IMOS component evidence rather than as the upper limit of possible sentiment-model performance.

Table X breaks down the best Udemy model by class to see whether the less extreme label distribution reduces the weaknesses observed in Coursera. Since Udemy contains more non-positive ratings, the per-class report tests whether stronger minority representation improves recognition of dissatisfaction and neutrality.

Table X. Per-class report for the best Udemy macro-F1 model

Class/average	precision	recall	f1-score	support
1	0.497	0.505	0.501	924.00
2	0.284	0.279	0.281	605.00
3	0.346	0.324	0.335	977.00
4	0.358	0.158	0.219	1,593.00
5	0.750	0.887	0.813	5,177.00
accuracy	0.625	0.625	0.625	0.625
macro avg	0.447	0.431	0.430	9,276.00
weighted avg	0.585	0.625	0.595	9,276.00

Table X shows better recognition of very negative Udemy comments, with label 1 reaching an F1-score of 0.501. This is an improvement, but it is still not strong enough to support reliable automated detection without further validation and error analysis. Label 4 remains weak, suggesting that ordinary positive comments are difficult to

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

separate from very positive comments when the representation is limited to TF-IDF features. This appears to be an ordinal and semantic boundary problem, not simply a data-volume problem.

Table XI reports the Udemy confusion matrix to identify the error patterns behind the per-class scores. This step is necessary because validation should do more than confirm the strongest model; it should also show where the model remains unreliable.

Table XI. Confusion matrix of SGD Logistic Regression on Udemy

Actual class	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5
True 1	467	153	103	19	182
True 2	209	169	106	21	100
True 3	131	162	317	120	247
True 4	60	61	218	251	1,003
True 5	72	51	171	290	4,593

The Udemy confusion matrix shows that the largest remaining error is the movement of true class-4 comments into class 5. This compression within the positive ratings suggests that adjacent favorable classes may not contain sufficiently distinct lexical patterns for a bag-of-words representation to separate them reliably. Future work should therefore consider ordinal classification, probability calibration, aspect-level labeling, or contextual transformer representations to improve discrimination between neighboring sentiment intensities.

Figure 3 summarizes the Udemy validation results by macro-F1 after the tables have established the platform context and error structure. The figure provides a compact visual summary of model ranking while preserving the argument that macro-F1, not accuracy, is the more appropriate selection metric in this imbalanced setting.

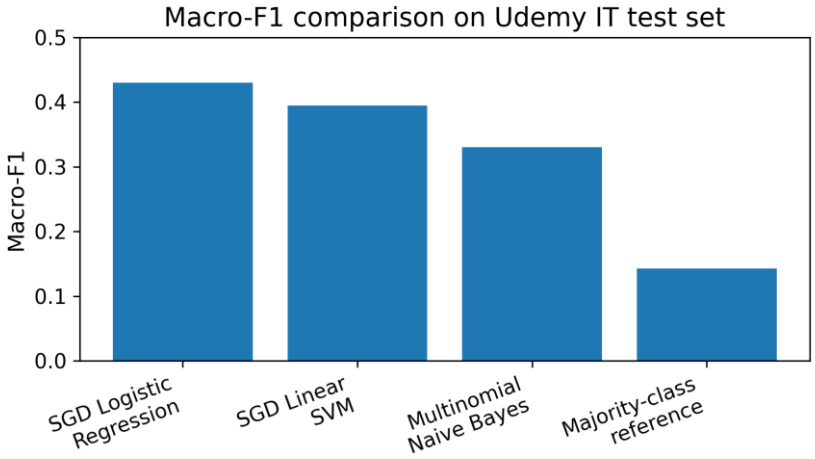

Fig. 3. Macro-F1 comparison of Udemy classifier components.

Bars show held-out macro-F1 scores for SGD Logistic Regression, SGD Linear SVM, Multinomial Naive Bayes, and the majority-class reference component on the Udemy IT test set. The figure shows the same macro-F1 ranking observed in Coursera, while the values remain too modest for fully automated deployment.

Figure 3 shows that the Udemy macro-F1 point-estimate hierarchy matches the Coursera pattern: SGD Logistic Regression leads, followed by SGD Linear SVM, Multinomial Naive Bayes, and the majority-class reference. This repetition strengthens the descriptive validation of the retrieval-to-classification workflow, but it does not replace statistical uncertainty estimation because no confidence intervals or significance tests were computed.

The Udemy results refine the cross-platform interpretation. Reduced imbalance improves recognition of strongly negative feedback, but it does not resolve confusion between adjacent positive categories. The educational

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

implication is twofold: negative-review detection may be more feasible on platforms with more non-positive ratings, while separating positive from very positive feedback likely requires richer semantic, ordinal, or aspect-based features than those used in the present TF-IDF component.

Together, Tables IX to XI and Fig. 3 answer the third research question. The strongest tested classifier component has the same macro-F1 point-estimate ranking across two platforms, but this finding is limited to transparent TF-IDF-based component modeling and does not include formal statistical testing. The evidence supports CP-IMOS as a reproducible evaluation process; it does not support a claim that the current system is a high-accuracy or deployment-ready sentiment engine.

Stepik as multilingual retrieval extension

The Stepik analysis addresses the multilingual retrieval part of the study. The corpus contained 5,721 raw rows and 5,719 non-empty review texts. After exact duplicate removal, 5,091 unique review texts remained, and 5,058 contained Cyrillic characters. These checks show that CP-IMOS can ingest a Russian-language MOOC review file, remove duplicates, and apply a script-level validation check before deciding whether a corpus is eligible for supervised modeling.

Because Stepik does not include rating or sentiment labels, it is not merged with Coursera and Udemy for supervised classification. Its contribution is limited but useful: it tests the multilingual retrieval gate of CP-IMOS through non-English text ingestion, uniqueness checking, Cyrillic-character validation, and the exclusion of unlabeled data from rating-derived sentiment training. Future Stepik data with manual annotations or rating links would be needed before multilingual models such as XLM-RoBERTa could be evaluated.

Cross-platform interpretation

The cross-platform findings point to three connected conclusions. First, both Coursera and Udemy are positively skewed, although the degree of skew differs. Second, accuracy rewards majority-class prediction and can conceal weak performance on negative, neutral, and moderately positive reviews. Third, SGD Logistic Regression with class balancing produced the strongest macro-F1 point estimate on both held-out test sets. Since no confidence intervals, bootstrap estimates, or significance tests were computed, this finding should be read as descriptive cross-platform consistency rather than formal statistical superiority.

For IT education, the practical lesson is to match the review-mining system to the decision it is meant to support. For broad satisfaction monitoring, accuracy and weighted-F1 may be useful supplementary indicators alongside macro-F1. For detecting problems in programming tasks, laboratory instructions, software installation, debugging support, or assessment design, minority-class recall, macro-F1, and confusion-matrix review are more important. A high aggregate score should not be treated as proof that the system can identify the comments most useful for course improvement.

IMPLICATIONS

Implications for MOOC sentiment classification

Cross-platform review classification should begin with data profiling because the profile determines what the model can validly claim. Duplicate removal, metadata inspection, language separation, and label-distribution analysis are not minor preprocessing steps; they affect construct validity, platform comparability, and the interpretation of every metric reported in the results.

Implications for IT education analytics

For IT education analytics, negative and neutral feedback should not be allowed to disappear inside majority-positive summaries. These smaller groups may contain the most actionable evidence about unclear coding instructions, outdated software requirements, weak debugging support, assessment problems, or missing

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

laboratory guidance. Course teams should therefore use the model as a screening aid and combine automated predictions with manual review of minority-class and misclassified comments.

Implications for future model development

Future model development should extend CP-IMOS beyond TF-IDF and classical machine-learning components by evaluating contextual transformer models such as BERT, RoBERTa, DistilBERT, and multilingual models such as XLM-RoBERTa [7], [8], [11], [13], [23], [28]. These models should be tested with the same cleaning, splitting, PSII, macro-F1, per-class reporting, and confusion-matrix logic used in the present study so that any performance gain can be compared fairly with the transparent baseline. Transformer models are especially relevant because they can capture word order, phrase context, and semantic nuance that sparse TF-IDF features may miss, including mixed reviews, indirect dissatisfaction, and adjacent positive-versus-very-positive sentiment boundaries.

A second extension should develop a manually annotated MOOC sentiment benchmark alongside the existing rating-derived labels. Human annotators could code review sentiment using clear guidelines for very negative, negative, neutral, positive, and very positive text, with agreement checks such as Cohen's kappa or Krippendorff's alpha before model training. Comparing manual labels with star-rating-derived labels would show whether the model is learning textual sentiment or mainly reproducing platform rating behavior. This benchmark would also make it possible to evaluate difficult cases such as mixed praise and criticism, comments with technical complaints but high ratings, and short reviews whose star ratings do not fully match the written text.

A third extension should test imbalance-handling strategies as part of the CP-IMOS model stage. Candidate methods include random oversampling, SMOTE or other synthetic oversampling approaches, class-weight tuning, cost-sensitive learning, focal loss for neural models, calibrated decision thresholds, and ensemble methods. These methods should be evaluated primarily through macro-F1, macro recall, minority-class F1, and confusion-matrix movement rather than accuracy alone. The objective would not simply be to increase overall performance, but to improve detection of very negative, negative, neutral, and ordinary positive reviews that are most useful for course improvement.

LIMITATIONS

First, the Udemy comments were processed as a 200,000-row sample because the full comments file exceeded the computational scope of the study. The sample supports cross-platform validation, but it should not be described as the complete Udemy comments corpus. Second, the Coursera and Udemy labels were rating-derived rather than manually annotated sentiment labels. A learner's star rating may not fully match the sentiment or topic expressed in the text.

Third, the IT filtering approach differed by platform because the metadata differed. Coursera used keyword filtering of course identifiers and titles, while Udemy used official platform categories. These choices were transparent and reproducible, but they may still include some borderline computing-related courses or miss some computing courses. Fourth, the Stepik corpus supported multilingual retrieval only. It tested non-English ingestion, duplicate removal, and Cyrillic-character validation, but it did not support supervised sentiment classification because it lacked ratings or sentiment labels.

Finally, the models were classical machine-learning components using TF-IDF features. This design improves reproducibility and avoids overstating results, but it limits the ability to capture contextual meaning, multilingual semantics, and subtle sentiment boundaries. The TF-IDF hyperparameters were not chosen through systematic tuning or ablation analysis, and the model comparisons did not include confidence intervals or significance tests. The results should therefore be read as transparent component evidence for CP-IMOS, not as the final performance ceiling for MOOC sentiment classification. Future versions should combine transformer-based modeling, manually annotated benchmarks, and imbalance-aware training to address these limitations more directly.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

CONCLUSION

This paper presents a cross-platform MOOC review framework supported by Coursera, Udemy, and Stepik data. Coursera and Udemy provide the supervised rating-derived sentiment classification evidence, while Stepik is used as a Russian-language retrieval extension. This structure supports cross-platform analysis without making unsupported multilingual accuracy claims.

The analysis found that both Coursera and Udemy IT-related review subsets were positively skewed, with Coursera showing stronger imbalance according to PSII. On both platforms, SGD Logistic Regression produced the strongest macro-F1 point estimate among the tested CP-IMOS components. Multinomial Naive Bayes achieved higher accuracy on Coursera but a much weaker macro-F1, indicating weaker handling of minority classes. These results confirm that accuracy alone is not a reliable measure for MOOC review sentiment classification when the goal is educational improvement. They also show that future work should add uncertainty estimation before making stronger cross-platform performance claims.

The study contributes a transparent retrieval-to-classification workflow for IT education MOOC reviews and a caution about interpretation. The tables and figures show that the workflow can profile datasets, expose imbalance, compare CP-IMOS classifier components, and identify error patterns. They also show that classical TF-IDF models still struggle with minority and adjacent sentiment classes. Future research should therefore extend CP-IMOS through larger cross-platform samples, manually annotated sentiment labels, transformer-based models, multilingual evaluation with verified labels, and imbalance-handling techniques such as synthetic oversampling, cost-sensitive learning, and ensemble modeling. For Stepik, once rating-linked or manually labeled sentiment data become available, an XLM-RoBERTa-based approach is recommended as a strong multilingual classifier component for supervised evaluation.

ACKNOWLEDGMENTS AND DECLARATIONS

Acknowledgments. The authors acknowledge the public dataset sources used in the empirical analysis and the scholarly works that informed the study's framing.

Funding. No external funding was received for this study.

Conflict of Interest. The authors declare no conflict of interest.

Data Availability. The analysis used public datasets from Kaggle and Mendeley Data: Course Reviews on Coursera, Udemy Courses, and Dataset of MOOCs Reviews from Stepik on Russian Language. The processed results can be reproduced from the cleaning, filtering, sampling, and modeling decisions reported in the manuscript.

Ethical Statement. This study used public secondary data and did not collect new private learner records, contact human participants, bypass access restrictions, or redistribute personal information. Because the analysis used public datasets, no formal ethics approval was required for this manuscript. Future direct scraping or platform-level data collection should undergo institutional ethics review when applicable.

Use of Python. Python was used for dataset inspection, cleaning, duplicate checks, IT-course filtering, TF-IDF feature extraction, model training, metric computation, and figure generation. The main libraries were pandas, scikit-learn, and matplotlib.

Use of Generative Artificial Intelligence Tools. Generative AI assistance was used for language refinement, organization, and formatting support. Dataset counts, model results, tables, and figures were computed from the public datasets using the procedures described in the methodology. No AI tool was credited as an author. The authors remain responsible for checking accuracy, originality, citation alignment, and ethical compliance before submission.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

REFERENCES

1. Muhammad, I. (n.d.). Course Reviews on Coursera [Dataset]. Kaggle. Retrieved June 11, 2026, from https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera
2. Hossain, G. H. (n.d.). Udemy Courses [Dataset]. Kaggle. Retrieved June 11, 2026, from https://www.kaggle.com/datasets/hossaingh/udemy-courses
3. Dyulicheva, Y. (2022). Dataset of MOOCs Reviews from Stepik on Russian Language [Dataset]. Mendeley Data, Version 1. https://doi.org/10.17632/8rwpvrw4hw.1
4. Chen, X., Wang, F., Cheng, G., Chow, M.-K., & Xie, H. (2022). Understanding learners' perception of MOOCs based on review data analysis using deep learning and sentiment analysis. Future Internet, 14(8), 218. https://doi.org/10.3390/fi14080218
5. Gomez, M. J., Calderón, M., Sánchez, V., Clemente, F. J. G., & Ruipérez-Valiente, J. (2022). Large scale analysis of open MOOC reviews to support learners' course selection. Expert Systems with Applications, 204, 118400. https://doi.org/10.1016/j.eswa.2022.118400
6. Jatain, D., Singh, V., & Dahiya, N. (2023). An exploration of the causal factors making an online course content popular and engaging. International Journal of Information Management: Data Insights, 3(2), 100194. https://doi.org/10.1016/j.jjimei.2023.100194
7. Chen, X., Zou, D., Xie, H., Cheng, G., Li, Z., & Wang, F. L. (2025). Automatic classification of online learner reviews via fine-tuned BERTs. International Review of Research in Open and Distributed Learning, 26(1). https://doi.org/10.19173/irrodl.v26i1.8068
8. Chen, X., Xie, H., Zou, D., Cheng, G., Tao, X., & Wang, F. (2025). Perceived MOOC satisfaction: A review mining approach using machine learning and fine-tuned BERTs. Computers and Education: Artificial Intelligence, 8, 100366. https://doi.org/10.1016/j.caeai.2025.100366
9. Dissanayake, A. D. M. B. B., & Fernando, P. (2025). Aspect-based sentiment analysis of student reviews on MOOC platforms. In 2025 5th International Conference on Advanced Research in Computing (ICARC). IEEE, Piscataway, NJ, USA. https://doi.org/10.1109/ICARC64760.2025.10963217
10. Awadh, W. A., Sulaiman, R. B., & Mahmoud, M. A. (2025). Aspect-based sentiment analysis in MOOCs: A systematic literature review introducing the MASC-MEF framework. Journal of King Saud University Computer and Information Sciences, 37, Article 2. https://doi.org/10.1007/s44443-025-00018-1
11. Aryal, S. K., Prioleau, H., Washington, G. J., & Burge, L. (2023). Evaluating ensembled transformers for multilingual code-switched sentiment analysis. In 2023 International Conference on Computational Science and Computational Intelligence. IEEE, Piscataway, NJ, USA. https://doi.org/10.1109/CSCI62032.2023.00032
12. Frias, R. R., Medina, R. P., & Sison, A. S. (2023). Attention-based bilateral LSTM-CNN for the sentiment analysis of code-mixed Filipino-English social media texts. In 2023 International Conference on Digital Applications, Transformation & Economy. IEEE, Piscataway, NJ, USA. https://doi.org/10.1109/ICDATE58146.2023.10248926
13. Kananta, G. S. C., Saputro, D. R. S., & Sutanto, S. (2026). Analysis of multilingual opinion polarization with cross-lingual language model-robustly optimized bidirectional encoder representations from transformers approach (XLM-RoBERTa). BAREKENG: Jurnal Ilmu Matematika dan Terapan, 20(2), 1709-1718. https://doi.org/10.30598/barekengvol20iss2pp1709-1718
14. Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Transactions of the Association for Computational Linguistics, 12, 820–836. https://doi.org/10.1162/tacl_a_00675
15. Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, F1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. Journal of Big Data, 12(1), Article 268. https://doi.org/10.1186/s40537-025-01313-4
16. Pant, H. V., Lohani, M. C., & Pande, J. (2023). Thematic and sentiment analysis of learners' feedback in MOOCs. Journal of Learning for Development, 10(1), 38–54. https://doi.org/10.56059/jl4d.v10i1.740
17. Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2022). Sentiment analysis and opinion mining on educational data: A survey. Natural Language Processing Journal, 2, 100003. https://doi.org/10.1016/j.nlp.2022.100003

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

18. Yan, C., Liu, J., Liu, W., & Liu, X. (2023). Sentiment analysis and topic mining using a novel deep attention-based parallel dual-channel model for online course reviews. Cognitive Computation, 15(1), 304–322. https://doi.org/10.1007/s12559-022-10083-7
19. Du, B. (2023). Research on the factors influencing the learner satisfaction of MOOCs. Education and Information Technologies, 28(2), 1935–1955. https://doi.org/10.1007/s10639-022-11269-0
20. Koufakou, A. (2024). Deep learning for opinion mining and topic classification of course reviews. Education and Information Technologies, 29, 2973–2997. https://doi.org/10.1007/s10639-023-11736-2
21. Marfani, H., Hina, S., & Tabassum, H. (2022). Analysis of learners' sentiments on MOOC forums using natural language processing techniques. In 2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS) (pp. 1-8). IEEE. https://doi.org/10.1109/ICONICS56716.2022.10100401
22. Pan, X., Hu, B., Zhou, Z., & Feng, X. (2023). Are students happier the more they learn? Research on the influence of course progress on academic emotion in online learning. Interactive Learning Environments, 31(10), 6869–6889. https://doi.org/10.1080/10494820.2022.2052110
23. Liu, Z., Kong, X., Chen, H., Liu, S., & Yang, Z. (2023). MOOC-BERT: Automatically identifying learner cognitive presence from MOOC discussion data. IEEE Transactions on Learning Technologies, 16(4), 528–542. https://doi.org/10.1109/TLT.2023.3240715
24. Qu, Y., Li, F., Li, L., Dou, X., & Wang, H. (2022). Can we predict student performance based on tabular and textual data? IEEE Access, 10, 86008–86019. https://doi.org/10.1109/ACCESS.2022.3198682
25. Su, B., & Peng, J. (2023). Sentiment analysis of comment texts on online courses based on hierarchical attention mechanism. Applied Sciences, 13(7), 4204. https://doi.org/10.3390/app13074204
26. Prasad, S. B. A., & Nakka, R. P. K. (2023). Supervised sentiment analysis of indirect qualitative student feedback for unbiased opinion mining. Engineering Proceedings, 59(1), 15. https://doi.org/10.3390/engproc2023059015
27. Alzaid, M., & Fkih, F. (2023). Sentiment analysis of students' feedback on e-learning using a hybrid fuzzy model. Applied Sciences, 13(23), 12956. https://doi.org/10.3390/app132312956
28. Mabokela, K. R., Celik, T., & Raborife, M. (2023). Multilingual sentiment analysis for under-resourced languages: A systematic review of the landscape. IEEE Access, 11, 15996–16020. https://doi.org/10.1109/ACCESS.2022.3224136
29. Ganganwar, V., Rajalakshmi, R., & Gayathri, S. (2024). Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification. Journal of Information and Telecommunication, 8(2), 167–188. https://doi.org/10.1080/24751839.2023.2270824
30. Suhaeni, C., Yong, H. S., & Ryu, H. (2024). Enhancing imbalanced sentiment analysis: A GPT-3-based sentence-by-sentence generation approach. Applied Sciences, 14(2), 622. https://doi.org/10.3390/app14020622