Innovations in Testing 2024

ATP GLOBAL

BETTER TOGETHER
Embrace change. Share solutions.

Integrating Machine Learning and NLP in Test Development: A Case Study

Anh (Andrew) Tran, Frank Martino and Benjamin Andrews

Background

- Natural Language Processing (NLP) and Machine Learning (ML) enable analysis and measurement of text similarity, alongside capabilities such as automated scoring and item creation, thereby enhancing operational efficiency and streamlining workflows.
- Managing enemy items (EI), defined as pairs of items too related to one another to appear on the same test form, challenges exam integrity by affecting scoring accuracy. Additionally, finding and properly flagging enemy items often is laborious for subject matter experts (SMEs).
- Other studies investigated how NLP techniques could be used for enemy item detection (Weir, 2019; Peng, 2020; Becker & Kao, 2022).

Purpose

- The project assesses automated detection of EI in medical sonography-related item banks using NLP and ML classification methods.
- The use of these methods could potentially identify EI more efficiently, reduce the amount of time needed for review, and provide insights into characteristics of items that are most prone to enemy relationships.

Methods

- Data:
  1. Five exam banks (4,481 multiple-choice items | 2,294,931 item pairs).
  2. Stem + key pairs, item attributes (content outline, media Y/N, item length, keywords, difficulty parameter IRT_b), and enemy status of each item pair.
- Analytical procedure:
  1. Text Processing: Text cleaning, tokenization, Document-Term Matrix, extraction of keywords, and topic modeling.
  2. Similarity Index Calculations: Local Dependency (LD), Vector Space Model (VSM), Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA).
  3. Synthetic Minority Over-Sampling (SMOTE) for Imbalanced Data.
  4. Enemy Item Pair Classification: Logistic Regression (LR) and Random Forest (RF) with different thresholds.
  5. SMEs review. Then update enemy status and retrain the classification models.

Results

- Sample Characteristics:
  - There were 2016 (0.09%) enemy item pairs originally.
  - Majority of enemy item pairs have overlapping content outlines (69%), media Y/N statuses (85%), and topics (66%).

Similarity Indexes between Enemy and Non-Enemy Pairs

	Average Item Length	Keyword Jaccard Index*	LD Yen's Q3†	VSM Cosine Index‡	LSA Cosine Index	LDA Jensen–Shannon Divergence§
Enemy	18.890	0.337	0.322	0.628	0.793	0.339
Non Enemy	17.759	0.011	0.214	0.182	0.083	0.737

*Note: *Higher means more similar; †Higher means more correlated; ‡Higher means more similar; §Lower means more similar*

- Initial Results:
  - LR revealed similarity indexes, item length, and attribute overlap as key predictors of enemy status ($p<0.01$), with different predictors leading in various models (Cosine for VSM/LSA, JSD for LDA).
  - Models demonstrated high accuracy (>0.9), with LDA showing superior performance; however, accuracy's reliability is questioned due to class imbalance. SMOTE was applied to mitigate this issue.
  - Trade-off observed between precision and recall, with many models showing high recall (0.33-0.87) but low precision (0.04-0.25), indicating a tendency to classify non-enemy pairs as enemy.
  - The F1 Score, a balance of precision and recall, identified the LDA_RF model with a 0.8 threshold as having the most balanced performance among the tested models.
- SME Review – First Round:
  - A sample of 82 item pairs with different EI statuses marked by the bank and the models were selected for four SMEs' reviews.
  - The SMEs agreed on 24 (29%) potential enemy pairs identified by the models.

Discussion

- The project focuses on creating an ML and NLP tool for the automatic detection of enemy pairs, moving away from manual processes to improve the fairness and validity of assessments.
- Initial results revealed potential enemy pairs not previously identified in the current banks, highlighting the model's ability to uncover new insights.
- Encountered limitations include incomplete enemy capture in the existing bank, lack of standardized enemy detection procedures among SMEs, and difficulties in processing items with images.
- Planned next steps involve SME review of more potential enemy pairs for validation, followed by using the new data to refine and retrain the model.

References

Becker, K. A., & Kao, S. C. (2022). Identifying Enemy Item Pairs using Natural Language Processing. Journal of Applied Testing Technology, 41-52.

Peng, F. (2020). Automatic enemy item detection using natural language processing. (Unpublished doctoral dissertation). The University of Illinois at Chicago, Chicago, IL.

Weir, J. B. (2019). Enemy item detection using data mining methods. (Unpublished doctoral dissertation). The University of North Carolina at Greensboro,Greensboro, NC.

For further information, contact
Anh (Andrew) Tran
andrew.tran@inteleos.org