Multi-Label Classification of E-Commerce Customer Reviews via Machine Learning

Paper : Multi-Label Classification of E-Commerce Customer Reviews via Machine Learning

Team : INSIDE OUT

Team Members:

1. Kriti Madumadukala (2022101069)
2. Vysishtya Karanam (2022102044)
3. Madireddy Ananya (2022101102)

1. Problem Statement

The rapid expansion of e-commerce has resulted in an enormous volume of customer reviews, making it imperative to develop automated methods for extracting meaningful insights. Traditional multi-label classification techniques often rely on TF-IDF or Word2Vec embeddings, which may fail to capture deep semantic relationships between words. These conventional approaches struggle with contextual understanding, leading to suboptimal classification performance.

This study aims to enhance multi-label classification of e-commerce customer reviews by incorporating **contrastive learning-based embeddings (SimCSE, SBERT)** to generate richer and more discriminative feature representations. By comparing **TF-IDF, Word2Vec, and contrastive learning embeddings**, we seek to determine the most effective technique for multi-label classification. Additionally, we will assess the computational efficiency and scalability of contrastive learning methods.

2. Project Scope

This project focuses on applying and evaluating different embedding techniques for multi-label classification of e-commerce customer reviews. The scope includes:

- Data preprocessing and augmentation of e-commerce review datasets.
- Implementing and fine-tuning contrastive learning models (SimCSE, SBERT) for feature extraction.
- Benchmarking classification performance across TF-IDF, Word2Vec, and contrastive learning embeddings.
- Evaluating classification models using machine learning classifiers such as Random Forest, SVM, and Neural Networks.
- Conducting computational analysis to assess the trade-offs between accuracy and efficiency of the proposed methods.

3. Datasets

Primary Dataset: A curated dataset of over **50,000 e-commerce customer reviews** (collected from Turkish consumers). The dataset is multi-labeled, where each review may belong to multiple categories such as **fabric quality, price performance, express delivery, and product durability**.

- **Preprocessing Steps:** Tokenization, stop-word removal, lemmatization, and vectorization will be applied to clean and prepare the dataset.
- **Data Augmentation:** Techniques such as synonym replacement and back translation may be employed to increase dataset diversity.

Dataset Statistics:

- **Total Reviews**: ~50,000
- **Categories**: Electronics, Women's Wear, Home and Life
- **Labels**: Multiple labels per review (e.g., product quality, delivery speed, price performance)

4. Literature Review

4.1. Multi-Label Classification

Multi-label classification is an extension of traditional single-label classification, where each instance can be associated with multiple labels. This approach is particularly useful in text classification tasks, such as sentiment analysis, where a single review can express multiple sentiments or aspects.

Multi-Class Classification: Traditional Approaches

2.1 Machine Learning-Based Methods

Early approaches to multi-class classification relied on standard machine learning algorithms such as:

- **Naïve Bayes (NB)** (McCallum & Nigam, 1998): A probabilistic classifier based on Bayes' theorem. It assumes word independence, making it simple yet effective for text classification.
- **Support Vector Machines (SVMs)** (Joachims, 1998): Uses hyperplanes to separate classes with a **one-vs-one** or **one-vs-all** approach.
- **Decision Trees & Random Forests** (Breiman, 2001): Effective for structured data but limited in handling large vocabulary spaces.

4.2. Feature Extraction Methods

- **TF-IDF**: A frequency-based method that evaluates the importance of a word in a document relative to a corpus.
- **Word2Vec**: A prediction-based method that captures semantic relationships between words using neural networks.
- **Contrastive Learning (SimCSE, SBERT)**: Advanced techniques that generate better embeddings by contrasting positive and negative pairs of sentences.

4.3. Related Work

Several studies have focused on sentiment analysis and multi-label classification of customer reviews. However, most of these studies use traditional feature extraction methods like TF-IDF and Word2Vec. The introduction of contrastive learning for generating embeddings in multi-label classification is a novel approach that has not been extensively explored.

5. Implementation Plan

Phase 1: Data Preparation (Week 1–2)

Goal: Gather, clean, and preprocess text data for multi-label classification.

Tasks:

- **Dataset Selection**: Choose an open-source dataset (e.g., Reuters-21578, Amazon reviews, or any multi-label dataset).
- **Data Cleaning**: Convert text to lowercase, remove stopwords, punctuation, and special characters.
- **Label Encoding**: Convert multi-label outputs into a binary format.
- **Train-Test Split**: Divide data into training (80%) and test (20%) sets.
- **Basic Exploratory Data Analysis (EDA)**: Understand label distribution and dataset properties.

Deliverable: Cleaned dataset ready for embedding generation and classification.

Phase 2: Embedding Generation (Week 3–4)

Goal: Generate text embeddings using three different methods for comparison.

Tasks:

- **Baseline Models:**
  - Implement **TF-IDF** (Scikit-learn).
  - Implement **Word2Vec** (Gensim).
- **Contrastive Learning Model (Main Contribution):**
  - Use **pre-trained SBERT** instead of training from scratch (to save time and computation).
  - Extract sentence/document-level embeddings using SBERT.

Deliverable: Stored embeddings from TF-IDF, Word2Vec, and SBERT for downstream classification.

Phase 3: Multi-Label Classification (Week 5–6)

Goal: Train and compare classification models using different embeddings.

Tasks:

- **Model Selection:** Use a simple classifier (e.g., Logistic Regression, Random Forest, or a small Neural Network).
- **Train the models using:**
    - TF-IDF embeddings
    - Word2Vec embeddings
    - SBERT embeddings
- **Compare performance across embeddings.**

Deliverable: Trained classifiers with performance metrics for comparison.

Phase 4: Evaluation & Comparison (Week 7–8)

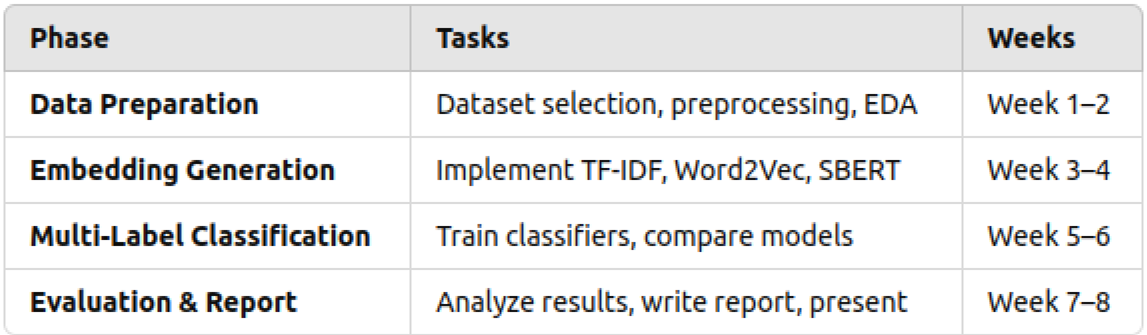
Goal: Analyze the results and draw conclusions on the effectiveness of contrastive learning embeddings.

Tasks:

- **Performance Metrics:**
    - Accuracy, Precision, Recall, F1-score (Micro & Macro), and Hamming Loss.
- **Comparison of Embeddings:**
    - Which embedding method performs best?
    - How much improvement does SBERT (contrastive learning) provide over TF-IDF and Word2Vec?
- **Computational Cost Analysis:**
    - Compare training time and memory usage.
- **Documentation & Final Report:**
    - Summarize findings, methodology, and results.

Deliverable: Final project report and presentation with key findings.

6. Tentative Timeline



Phase	Tasks	Weeks
Data Preparation	Dataset selection, preprocessing, EDA	Week 1–2
Embedding Generation	Implement TF-IDF, Word2Vec, SBERT	Week 3–4
Multi-Label Classification	Train classifiers, compare models	Week 5–6
Evaluation & Report	Analyze results, write report, present	Week 7–8

7. Novelty: Contrastive Learning for Multi-Label Classification

7.1. Introduction to Contrastive Learning

Contrastive learning is a self-supervised learning technique that aims to learn representations by contrasting positive pairs (similar instances) against negative pairs (dissimilar instances). Contrastive learning represents a significant shift from traditional embedding methods such as TF-IDF and Word2Vec. While TF-IDF captures simple word frequency patterns and Word2Vec relies on co-occurrence statistics to learn word-level embeddings, these methods do not directly optimize for the specific task of distinguishing between multiple, often overlapping, labels in a document. In our project, we propose using advanced contrastive learning techniques—employing models like SimCSE or SBERT—to generate context-aware embeddings that are particularly well-suited for multi-label classification.

7.2. Expected Outcomes

- **Improved Embeddings**: Contrastive learning is expected to generate more semantically meaningful embeddings compared to traditional methods.
- **Better Classification Performance**: The use of advanced embeddings is expected to improve the accuracy and robustness of multi-label classification models.
- **Generalizability to Other Domains:** The contrastive learning approach can be applied to other NLP tasks like topic modeling, sentiment analysis, and named

entity recognition.

8. Conclusion

This project aims to develop a multi-label classification model for e-commerce customer reviews, with a focus on comparing traditional feature extraction methods with advanced contrastive learning techniques. By introducing contrastive learning, we aim to generate better embeddings and improve the performance of multi-label classification models. The project will provide valuable insights into the effectiveness of different feature extraction methods and classification algorithms, contributing to the field of sentiment analysis and customer review classification.

9. References

1. Zhang, M., & Zhou, Z. (2020). A review of multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 32(3), 567-584.
2. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
3. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP*, 3982-3992.
4. Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 898-910.
5. Yang, P., Fang, H., & Lin, J. (2017). Anserini: Enabling the use of Lucene for information retrieval research. *Proceedings of SIGIR*, 1253-1256.