Sentiment Analysis on 'HelloTalk' App Reviews Using NRC Emotion Lexicon and GoEmotions Dataset

Simay Akar, Yang Sok Kim, Mi Jin Noh

Abstract

During the post-pandemic period, the interest in foreign language learning surged, leading to increased usage of language-learning apps. With the rising demand for these apps, analyzing app reviews becomes essential, as they provide valuable insights into user experiences and suggestions for improvement. This research focuses on extracting insights into users' opinions, sentiments, and overall satisfaction from reviews of HelloTalk, one of the most renowned language-learning apps. We employed topic modeling and emotion analysis approaches to analyze reviews collected from the Google Play Store. Several experiments were conducted to evaluate the performance of sentiment classification models with different settings. In addition, we identified dominant emotions and topics within the app reviews using feature importance analysis. The experimental results show that the Random Forest model with topics and emotions outperforms other approaches in accuracy, recall, and F1 score. The findings reveal that topics emphasizing language learning and community interactions, as well as the use of language learning tools and the learning experience, are prominent. Moreover, the emotions of 'admiration' and 'annoyance' emerge as significant factors across all models. This research highlights that incorporating emotion scores into the model and utilizing a broader range of emotion labels enhances model performance.

Keywords : LDA | Sentiment Analysis | NRC Emotion Lexicon | GoEmotions

I. INTRODUCTION

In the past few years, there has been a notable transformation in the landscape of foreign language acquisition, marked by a noticeable increase in enthusiasm and involvement. This shift has gained particular prominence during the global pandemic era, as individuals around the world actively seek new opportunities for personal development and skill enhancement in the face of challenges.

Notably, language learning has emerged as a focal point for many, with language learning applications, such as Duolingo, Busuu, Babbel, and HelloTalk, serving as crucial tools in developing this interest [1]. These apps provide ease and an accessible focal point for personal development journeys in language learning. Consequently, with the continued interest and development, the global language exchange app market is anticipated to experience significant growth, particularly during the forecast period between 2024 and 2032 [2].

HelloTalk is a global language exchange app launched in 2012, boasting 20 million users today, connecting learners with native speakers in 200 countries. Since it is a conversation-based app, it facilitates

Manuscript : 2024.04.29
Revised : 2024.05.28
Confirmation of Publication:2024.06.10
Corresponding Author : Yang Sok Kim, e-mail : yangsok.kim@kmu.ac.kr

various areas, such as language learning, cultural exchange, practice, and socialization [3].

This research aims to gain valuable insights into users' experiences, opinions, and overall emotions regarding the features and usability of HelloTalk using online reviews collected from the Google Play Store. This research contributes to the research community by identifying topics related to online language learning platforms using LDA, comparing sentiment predictions with various machine learning models using topics and emotions, and analyzing factors impacting sentiment prediction through model explanation.

II. RELATED WORK

1. Sentiment Analysis

Sentiment analysis aims to examine individuals' sentiments of any topic, product, or entity by classifying and analyzing feedback, reviews, and comments related to products, hotels, online platforms, and social media posts in line with the evolving digital landscape [4]. Classifying online reviews into positive and negative sentiments using machine learning and deep learning algorithms [5] is a common sentiment analysis approach [6]. This research employs machine learning models to analyze the sentiments of the online reviews of HelloTalk.

2. LDA Topic Modeling

The Latent Dirichlet Allocation (LDA) method [7], a probabilistic generative model for corpora, is widely used to identify key themes in online reviews and group user feedback under specific topics. Many studies utilizing LDA aim to understand better how consumers address various issues and how these issues can be leveraged for better decision-making [8]. This research employs LDA to extract topics from the online reviews of HelloTalk collected from the Google Play Store. Then the topics are used as predictor variables in sentiment classification.

3. Texts and Emotions

Emotion is usually used to express an individual's feelings, such as sadness, joy, surprise, etc. Emotion is regarded as an important factor that impacts the sentiment [9]. Therefore, understanding users' emotions and addressing their needs effectively in online reviews is crucial for both consumers and e-commerce retailers [10].

Several studies aim to detect the emotions from the text. The National Research Council Canada (NRC) created the comprehensive word-emotion association lexicon referring to the NRC Emotion Lexicon, through Amazon's Mechanical Turk. It consists of 14,182 entries, associated with eight basic emotions (anger, fear, expectation, trust, surprise, sadness, joy, and disgust) proposed by Ekman [11], as well as two sentiments (negative and positive) [12]. Many studies have performed sentiment analysis using the NRC Emotion Lexicon and identified factors and emotions influencing the effectiveness of reviews [13].

GoEmotions was proposed to overcome the limitations of existing emotion lexicons, which rely on small examples and emotion

classification taxonomy. It is the largest human-annotated dataset of 58,000 Reddit comments, labeled with 27 emotion categories and neutral. The taxonomy was designed by considering the scope of related works in psychology and the data coverage [14]. It draws inspiration from previous works in emotional expression understanding, notably from Picard [15] and Bostan & Klinger [16]. Several studies used GoEmotions and demonstrated its usefulness in sentiment analysis [17,18]. This study conducted sentiment analysis with the NRC Emotion Lexicon and GoEmotions dataset to analyze users' dominant emotions regarding the HelloTalk app's reviews.

III. DATA AND METHODOLOGY

Figure 1 illustrates the research process which consists of data collection, data pre-processing, data encoding with topic modeling and emotion encoding, modeling, and evaluation. Details of each step are explained in the following sections.

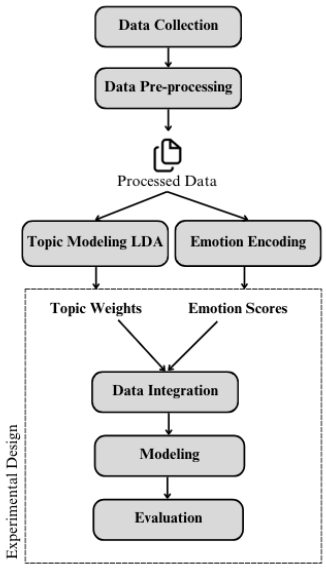

Fig. 1. Research Process

1. Data Collection

The HelloTalk review dataset, acquired through the "google-play-scraper" package from the Google Play Store, encompasses reviews from December 30, 2013, to November 6, 2023. The initial dataset comprised a total of 29,419 entries.

2. Data Pre-processing

We started the pre-processing phase by excluding 2,414 reviews with a score of 3. Negative labels were assigned to scores 1 and 2, while positive labels were assigned to scores 4 and 5. We expanded the NLTK library's word list by adding common words and then eliminated stop words. For data cleaning, we used the 'clean-text' package, involving the conversion of text to lowercase and the removal of various elements, such as Unicode characters, URLs, emails, phone numbers, numerical values, digits, currency symbols, and punctuation marks. In addition, separate functions were created for lemmatization, emoji removal, and tag discarding. A total of 238 empty reviews were identified and removed. Then, lemmatization and tokenization processes were applied sequentially. Subsequently, stop words were removed once more. As a result of this process, 118 reviews were removed due to empty elements in the processed token list, resulting in a streamlined dataset of 26,649 entries for subsequent analysis.

3. Topic Modeling Encoding

In order to derive latent factors from the text, we used a topic modeling approach. We applied TF-IDF vectorization to the

cleaned review dataset and performed topic modeling using LDA. We identified five topics based on the coherence score results, with the highest score recorded as 0.5392, as shown in Figure 2.

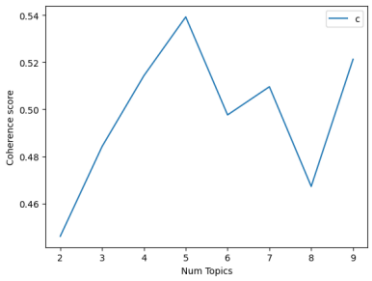

Fig. 2. Coherence Score

Then, we calculated topic weights for each review in the dataset and created a topic weights matrix. This approach served as an alternative to traditional numeric feature extraction methods, aimed at enhancing sentiment analysis performance. Traditionally, the common approach for extracting and representing numeric features from text data involves using a sparse matrix predominantly filled with zero values, along with the feature encoding method [19]. However, sparse matrices demand more computational resources and specialized algorithms, resulting in complexity in both time and space. Therefore, we opted to employ topic weights as predictor variables for sentiment analysis. Topic weights refer to the importance of a particular topic within a collection of documents. The LDA algorithm generates a probability vector for each document, distributing topics across documents and words across topics. Each value in this vector represents the weight of a particular theme in the document, offering a nuanced understanding of the document's thematic composition.

In order to enhance comprehension and characterization of the topics, we assigned titles to each one by examining the associated keywords and app reviews. Figure 3 illustrates word cloud visualization of the topics. Topic 0 was named 'Positive user experience and satisfaction' with the keywords 'excellent' (indicating high quality), 'easy_use' (pointing out user-friendly interface), and 'interface'. Topic 1 was named 'Language learning and community interaction' with keywords like 'application', 'native_speaker' (emphasizing interaction with native speakers), and 'language_exchange' (highlighting the community aspect of the app). Topic 2 was named 'App performance and updates' with keywords, such as 'fix_bug' (addressing technical issues), 'keep_crash' (stating the problem), and 'correction' (requesting adjustments). Topic 3 was named 'Utility of language learning tools and learning experience' with keywords 'language', 'practice' (stating learning activities), and 'highly_recommend' (indicating user endorsement). Topic 4 was named 'User feedback, issues and suggestions' with keywords 'update' (suggesting the need for improvement), 'problem' (highlighting issues), and 'please_fix' (requesting resolution of identified problems).

Fig. 3. Word Cloud Visualization

4. Emotion Encoding

Emotion text coding was performed with the NRC Emotion Lexicon and GoEmotions. For NRC, each word in the lexicon is represented with binary values indicating its association with the specified emotions. Besides, each word in the lexicon is associated with a set of emotion scores, indicating the intensity or strength of the emotions it evokes. The scores range from 0 to 1, with higher values indicating a stronger association with the corresponding emotion. These emotion scores allow for a more nuanced analysis of the emotional content associated with individual words, beyond a simple binary indication of presence or absence. The eight basic emotions were added to the dataset, and emotion scores for each review were computed using NRC. As the sentiments of the reviews had already been categorized as positive or negative by evaluating scores, the inclusion of redundant sentiment values (positive and negative) in the lexicon was deemed unnecessary. To address this, the 'positive' and 'negative' tags for each word in NRC were manually removed, ensuring that each word solely retained information about the eight emotions.

GoEmotions labeled for one or more of 27 emotion(s) or neutral [14] (See Table 1). For leveraging the GoEmotions, we utilized the Hugging Face Transformers library and the Optimum package to employ a RoBERTa base GoEmotions ONNX (Open Neural Network Exchange) model for text classification. The choice of the ONNX model is motivated by its comparable accuracy and model size to the original transformers model, coupled with improved speed performance compared to the original transformers model.

Table 1. GoEmotions Labels

category	emotion
positive	admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, realization
negative	anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness
ambiguous	confusion, curiosity, surprise

5. Experimental Design

We conducted experiments using four different settings, each involving different combinations of predictor variables to form the training datasets. We calculated topic weights for each review and then transformed them into a matrix. Later, these topic weights were integrated with emotion scores in four different experiments. Experiment 1 was conducted using only these topic weights and sentiment labels. Experiment 2 was conducted using topic weights along with the emotion scores coded by the NRC Emotion Lexicon, which includes 8 emotion tags, by integrating the emotion scores into the topic weights. Experiment 3 was conducted using topic weights along with the emotion scores coded by

the GoEmotions dataset developed by Google, by integrating the emotion scores into the topic weights. This dataset comprises 28 emotion tags, making it considerably larger than the NRC Emotion Lexicon. Finally, experiment 4 was conducted using topic weights and the emotion scores coded by the NRC Emotion Lexicon and GoEmotions. For modeling, we used Decision Tree (DT), Random Forest (RF), and XGBoost (XGB). Accuracy, precision, recall, and F1 score were set as performance metrics. Finally, the best model was used to explain the features that most impact the classification.

IV. RESULTS

1. Performance Evaluation

Table 2 shows the performance results of three models according to four different experiment settings. In experiment 1, XGB exhibited the highest performance. The notable metrics achieved by experiment 1 with XGB were accuracy (0.7407), recall (0.9977), and F1 score (0.8510). However, a higher precision value (0.7463) was observed with DT. In experiment 2, XGB demonstrated superior performance with the highest values in three metrics: accuracy (0.8043), recall (0.9338), and F1 score (0.8759) compared to other models. On the other hand, RF had the highest precision value (0.8262). In experiment 3, RF achieved the highest values in three metrics: accuracy (0.8583), recall (0.9270), and F1 score (0.9065). Meanwhile, XGB attained the highest precision score (0.8872). Lastly, in experiment 4, the RF achieved the highest values in three metrics: accuracy (0.8666), precision (0.8956), and F1 score (0.9115), while the XGB attained the highest recall score (0.9286). Among the four experiments, experiment 4 demonstrated the highest performance, particularly with RF, showing the highest metrics. Therefore, we conducted a feature importance analysis to explore experiment 4 further and identify which emotion labels and topics have emerged within the application reviews.

Table 2. Model Performance Results

	Model	Accuracy	Precision	Recall	F1 Score
Exp 1	DT	0.6396	**0.7463**	0.7790	0.7623
	RF	0.7364	0.7418	0.9889	0.8477
	XGB	**0.7407**	0.7418	**0.9977**	**0.8510**
Exp 2	DT	0.7266	0.8105	0.8226	0.8165
	RF	0.8009	**0.8262**	0.9254	0.8730
	XGB	**0.8043**	0.8247	**0.9338**	**0.8759**
Exp 3	DT	0.8024	0.8643	0.8698	0.8670
	RF	**0.8583**	0.8868	**0.9270**	**0.9065**
	XGB	0.8570	**0.8872**	0.9245	0.9055
Exp 4	DT	0.8006	0.8596	0.8733	0.8664
	RF	**0.8666**	**0.8956**	0.9281	**0.9115**
	XGB	0.8642	0.8924	**0.9286**	0.9101

2. Feature Importance Assessment

Feature importance metrics were used to quantify the significance of each feature in understanding the emotional content and topics discussed within the app reviews. Table 3 illustrates the results of the feature importance analysis conducted in Experiment 4, where the importance of the features is categorized and explained by topics, NRC emotion labels, and GoEmotions labels.

In DT, Topic 3 (2.11%) emerged as a significant factor. The emotion of 'trust' (1.34%) stood out among the NRC emotions, while within the GoEmotions set, 'admiration' (34.76%) was notably

prominent, followed by 'annoyance' (9.56%) and 'pride' (3.76%). In RF, Topic 1 (1.55%) garnered more attention among topics. Similarly, 'trust' (1.42%) stood out among the NRC emotions, while within the GoEmotions set, 'admiration' (11.13%) was the most important emotion, closely followed by 'annoyance' (6.36%) and 'gratitude' (6.04%). In XGB, Topic 1 (0.81%) was found to be an important topic. Among the NRC emotions, 'trust' (2.54%) stood out, whereas within the GoEmotions set, 'admiration' (36.53%) emerged as the most predominant emotion, followed by 'annoyance' (10.45%) and 'disapproval' (4.29%).

Despite minor variations, the emotions 'admiration' and 'annoyance' from the GoEmotions set consistently stood out within all models and across all features, highlighting the importance of these emotions in sentiment analysis. Admiration, an emotion of respect and approval, emerged as the most important emotion among HelloTalk users, indicating their appreciation of the application. We concluded that the prominence of admiration suggests users generally approve of the app and believe it contributes positively to their language-learning process. Annoyance, signifying irritation or boredom, emerged as the second most prominent emotion. We believe that the emotion of annoyance stems from the issues mentioned in the topic titled 'User feedback, issues, and suggestions'. This indicates that despite 'admiration' being the most significant emotion, the presence of 'annoyance' suggests users still have concerns about the application's features, indicating mixed feelings about its overall usability. It is essential to address these issues to improve the app in line with the user feedback in language learning applications.

Table 3. Feature Importance Results

	Factors	DT	RF	XGB
Topics	Topic 0	1.98%	1.39%	0.78%
	Topic 1	1.82%	1.55%	0.81%
	Topic 2	2.10%	1.44%	0.68%
	Topic 3	2.11%	1.39%	0.78%
	Topic 4	1.56%	1.43%	0.80%
NRC Lexicon	joy	0.98%	0.89%	2.69%
	anticipation	0.90%	0.77%	1.29%
	anger	0.91%	0.89%	1.64%
	sadness	0.73%	0.48%	1.21%
	trust	1.34%	1.42%	2.54%
	surprise	0.63%	0.99%	2.13%
	disgust	1.32%	1.27%	1.86%
	fear	0.44%	0.81%	0.99%
GoEmotions	admiration	**34.76%**	**11.13%**	**36.53%**
	amusement	0.97%	1.53%	1.01%
	approval	1.70%	3.65%	1.23%
	caring	1.09%	1.63%	1.17%
	desire	1.32%	1.38%	0.87%
	excitement	1.45%	4.24%	1.22%
	gratitude	2.13%	6.04%	1.28%
	joy	1.48%	3.54%	0.90%
	love	2.24%	2.89%	1.30%
	optimism	1.60%	2.40%	1.06%
	pride	3.76%	5.88%	3.08%
	relief	1.26%	1.90%	1.18%
	sadness	1.13%	1.59%	0.92%
	fear	0.93%	1.58%	1.14%
	embarrassment	1.21%	2.56%	1.26%
	disapproval	1.40%	4.11%	4.29%
	disappointment	1.31%	3.07%	1.20%
	annoyance	9.56%	6.36%	10.45%
	anger	2.03%	3.21%	1.82%
	nervousness	1.34%	1.49%	1.04%
	remorse	0.98%	1.50%	1.01%
	grief	1.31%	1.29%	0.89%
	disgust	1.25%	3.22%	1.10%
	realization	1.34%	1.45%	0.77%
	surprise	1.36%	1.42%	0.90%
	curiosity	1.07%	1.38%	0.85%
	confusion	1.26%	1.47%	0.93%
	neutral	1.96%	3.34%	2.40%

V. CONCLUSION

In this research, we utilized LDA to analyze user reviews of the HelloTalk

language learning application and performed sentiment analysis with topics and emotions. We conducted comparison experiments to observe how emotions affect sentiment analysis performance and subsequently assessed these experiments using classification models.

When evaluating performance, models with NRC outperformed those with only topic weights. Furthermore, models incorporating GoEmotions and topic weights outperformed others, highlighting the enhancement in analysis with a broader range of emotions. Combining both emotion schemas and topic weights resulted in the best performance among the four models. This research contributes by showing incorporating emotion scores improves model performance. Furthermore, as the emotion model incorporates a broader range of labels, classification performance also improves. This research also identified 'admiration' and 'annoyance' as dominant emotions and revealed that language learning and community interaction, as well as the utility of language learning tools and experiences, are important topics for users.

This study has several limitations. First, instead of using a vectorized sparse matrix for sentiment analysis, we employed topic weights, potentially impacting the results. Second, only two emotion schemas were employed to conduct sentiment analysis on the HelloTalk app review data and then identify prominent emotion labels. This may lead to different emotion labels when other emotion schemas are used. Third, while the NRC Emotion lexicon comprises eight emotion labels, GoEmotions consists of 28 labels, which may be seen as challenging to compare the results. Lastly, emojis were not incorporated into the research, despite their effectiveness in enhancing sentiment analysis accuracy.

Future studies could consider emojis and employ diverse emotion schemas to improve sentiment analysis. Additionally, researchers could analyze user reviews from other language learning apps besides HelloTalk to provide a broader understanding of user emotions and experiences.

REFERENCES

[1] Forbes(2020).https://www.forbes.com/sites/mergermarket/2020/04/15/language-learning-apps-are-seeing-a-surge-in-interest-during-the-covid-19-pandemic/?sh=1d1a9eb048f4(accessed Feb., 5, 2024).

[2] Language Exchange App Market - Growth, Trends and Forecast (2024-2032)(2024). https://www.linkedin.com/pulse/language-exchange-app-market-growth-trends-39fvf/ (accessed Mar., 21, 2024).

[3] A. V. Rivera, "HelloTalk," *CALICO Journal,* vol. 34, no. 3, pp. 384- 92, 2017.

[4] A. Yadav, and D. K. Vishwakarma, "Sentiment Analysis Using Deep Learning Architectures: A Review," *Artificial Intelligence Review,* vol. 53, no. 6, pp. 4335- 4385, 2020.

[5] A. Iqbal, R. Amin, J. Iqbal, R. Alroobaea, A. Binmahfoudh, and M. Hussain, "Sentiment Analysis of Consumer Reviews Using Deep Learning," *Sustainability,* vol. 14, no. 17: 10844, Aug. 2022.

[6] R. Das, and T. D. Singh, "Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges," *ACM Computing Surveys,* vol. 55, no. 13s, pp. 1- 38, Jul. 2023.

[7] D. M. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* 3, pp. 993−1022, 2003.
[8] B. C. Lee, and D. S. Kim, "Deriving the Determinants of Hotel Service Quality using OTA Reviews − LDA Topic Modeling −," *Journal of Hotel & Resort,* vol. 19, no. 4, pp. 41−58, Aug. 2020.
[9] P. Nandwani, & R. Verma, "A review on sentiment analysis and emotion detection from text, " *Social Network Analysis and Mining*, vol. 11, no. 1, 81, Dec. 2021.
[10] M. S. I. Malik, and A. Hussain, "Helpfulness of Product Reviews as a Function of Discrete Positive and Negative Emotions," *Computers in Human Behavior,* vol. 73, pp. 290- 302, Mar. 2017.
[11] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion,* vol. 6, no. 3, pp. 169- 200, 1992.
[12] S. M. Mohammad, and P. D Turney, "NRC Emotion Lexicon," National Research Council Canada, Nov. 2013.
[13] Y. L. S. Krishna, P. Paramesh, Y. T. Kumar and A. Gopi, "Sentiment Analysis of Product Reviews by using Naive Bayes and Vader Models," *ICSSIT*, Tirunelveli, India, pp. 1−6, 2023.
[14] D. Demszky, D. Movshovitz−Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine−Grained Emotions," *arXiv*, 2005.00547, Jun. 2020.
[15] R. W. Picard, "Affective Computing," MIT Press, 1997.
[16] L. A. M. Bostan and R. Klinger, "An analysis of annotated corpora for emotion classification in text," *ICCL*, pp. 2104- 2119, 2018.
[17] H. B. Jung, S. J. Jang, and B.C. Bae, "A Text−Based Emotion Analysis of the Audiences in a Webtoon−Based TV Drama: Focusing on <The Uncanny Counter>," *Journal of Digital Contents Society,* vol. 24, no. 4, pp. 755- 64, Apr. 2023.
[18] T. Sosea, C. Pham, A. Tekle, C. Caragea, and J. J. Li, "Emotion Analysis and Detection during COVID−19," arXiv, Jul. 2022.
[19] Kantorov, Vadim, and Ivan Laptev, "Efficient Feature Extraction, Encoding, and Classification for Action Recognition," *CVPR*, pp. 2593- 2600, 2014.

Authors

Simay Akar

She received her B.S. in Computer Engineering from Yeditepe University, Turkiye in 2020. She is currently pursuing her Master studies at the Department of Management Information Systems in Keimyung University, Korea. Her research interests are Big Data and Text Mining.

Yang Sok Kim

He has been serving as an Associate Professor at the Department of Management Information Systems, Keimyung University, Korea. He received his Ph.D. from University of Tasmania, Australia. His research interests are Machine Learning, Text Mining, Social Network, and Recommender Systems.

Mi Jin Noh

She received her M.S. and Ph.D. in Management Information Systems from Kyungpook National University, Korea in 2001 and 2006, respectively. Since 2022, she has been an assistant professor in Department of Business Big Data, Keimyung University, Korea. Her research interests are Big Data Analysis, Text Mining and Mobile Services.