

Development of an Artificial Intelligence Based Correction System for Spelling Errors in Product Reviews

Okan ÇİFTÇİ¹, Sumru NAYİR¹, Emre Tolga AYAN¹, Ceren ULUS², M. Fatih AKAY²

¹Department of NLP, Trendyol

²Department of Computer Engineering, Çukurova University

ABSTRACT

E-commerce has experienced rapid growth in recent years and continues to expand dynamically. In this sector, maximizing customer satisfaction and enhancing the shopping experience are recognized as important strategic initiatives. To maximize customer satisfaction, it is essential to accurately determine customer needs and provide appropriate solutions to meet demand. In this context, feedback obtained from customers holds significant importance. However, customer comments often contain spelling errors, complicating the analysis of these comments. This study aims to automatically correct spelling errors in user comments regarding products sold on the e-commerce site Trendyol.com. For this purpose, a system based on transformer architecture has been created. Various spelling error detection and correction models were subsequently developed based on this architecture. Prediction models have been developed using two separate datasets consisting of Trendyol user comments and two additional datasets, including the Turkish Spelling Check Dataset taken from the Hunspell library, and the effects of these four datasets on prediction performance have been examined. The success of the models has been evaluated using the Accuracy metric. The performance of the developed models was also compared with that of the model in the Zemberek library. As a result of the study, it has been observed that the utilization of the Turkish Spelling Check Dataset positively influences prediction performance. The developed system enhanced customer experience by correcting spelling errors in comments.

Keywords: Spelling error system, Product reviews, Transformer architecture, Turkish spelling check dataset, Zemberek

1. INTRODUCTION

The e-commerce sector has shown significant growth in recent years, driven by rapid technological advancements and digitalization [1]. Companies in this sector continuously expand their consumer base by leveraging the broad access provided by digital platforms, thus diversifying competitive dynamics. This rise in the popularity of e-commerce can be attributed to a global shift in consumer behavior toward digital channels, alongside the convenient, fast, and user-friendly shopping experiences offered by online platforms. Consequently, a steady increase is observed in the user base of e-commerce companies, which, in turn, drives significant growth in transaction volumes. This surge in transactions creates a pressing need for companies to manage high volumes efficiently and underscores the necessity for streamlined process management.

E-commerce platforms revolutionize customer interactions with products and services, enabling real-time feedback that significantly influences purchasing decisions. Customer reviews serve as critical feedback, where users share their experiences with products or services, express satisfaction levels, or raise complaints. These comments not only impact the purchasing decisions of potential customers but also provide businesses with valuable insights for improving product or service quality. Customer reviews may touch on aspects such as product features, quality, customer service, or delivery processes, and are essential for assessing brand reliability and customer satisfaction in the realms of digital marketing and e-commerce. Customer feedback and reviews are vital indicators of product reliability and quality, playing a key role in the decision-making processes of prospective buyers. This dynamic empowers consumers to make more informed decisions and enhances their overall shopping experience. However, the prevalence of spelling and grammar errors in reviews poses a significant issue. As long as comments do not contain harmful content like insults,

*Corresponding Author Email: f.cerenulus@gmail.com

Submitted: 04.11.2024 Revision Requested: 06.11.2024 Last Revision Received: 08.11.2024

Accepted: 11.11.2024 Published Online: 03.12.2024



Burdur
Mehmet Akif Ersoy
University Press



Cite this article as: Çiftçi, O., Nayir, S., Ayan, E.T., Ulus, C., Akay, M.F. (2024). Development of an Artificial Intelligence Based Correction System for Spelling Errors in Product Reviews. Scientific Journal of Mehmet Akif Ersoy University, 7(2): 99-108.

DOI: <https://doi.org/10.70030/simakeu.1577809>

<https://dergipark.org.tr/simakeu>



Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

profanity, or spam, they are visible to all users. Such errors can negatively affect other customers' perceptions and damage brand reputation. Correcting these errors is essential to prevent confusion, improve readability, and ensure that the intended message is clearly understood. Proper spelling enhances communication quality, lends a professional appearance to the text, and is therefore of great importance. [2]

Automating spelling and grammar error correction is a strategic necessity for e-commerce platforms. By facilitating decision-making based on clear, well-expressed information, platforms contribute to more favorable outcomes. Moreover, enabling users to articulate their thoughts clearly and concisely improves information flow, enhances customer satisfaction, boosts sales, and fosters brand loyalty.

This study aims to automatically correct spelling errors in user comments regarding products sold on the e-commerce site Trendyol.com. For this purpose, various spelling error detection and correction models have been developed, and a system based on transformer architecture has been created.

This study is organized as follows: Section 2 includes relevant literature. Dataset is given in Section 3. Section 5 presents methodology. Details of the system is presented in Section 5. Section 6 presents the results of the study. Section 7 concludes the paper.

2. LITERATURE REVIEW

[3] presented a two-step, deep learning-based model for detecting misspelled words in Turkish. This model incorporates character-based, syllable-based, and Byte Pair Encoding (BPE) approaches alongside Long Short-Term Memory (LSTM) and Bi-Directional LSTM (Bi-LSTM) networks. Additionally, a false positive reduction model was integrated to minimize false positives caused by the use of foreign words and abbreviations frequently found on online platforms. The results indicated that the proposed Bi-LSTM model with the BPE tokenizer performed effectively. [4] proposed an LSTM-based classification model that utilizes pre-trained language models to correct erroneous query inputs and spelling mistakes. Model pruning techniques, such as No-Teacher Distillation, were employed to address latency issues. The results demonstrated at least a 16% increase in accuracy. [5] aimed to improve the robustness of predictions by modifying traditional autoregressive encoder-decoder models to address spelling correction challenges. Various methods were tested, and a comparative analysis was conducted. [6] proposed a solution to enhance spelling error correction in search queries within e-commerce. Various types of spelling errors were systematically analyzed and categorized. A Transformer model was employed along with synthetic data generation techniques for contextual spelling correction, showing significant improvements over multiple models on both human-labeled data and online A/B experiments. [7] introduced a new method for spelling correction in search queries or single words using a procedure for generating artificial spelling errors. This technique was used to train a generative spelling correction model based on a Transformer architecture, outperforming the conventional noise addition approach. [8] proposed a hybrid model using the Bidirectional Encoder Representations from Transformers (BERT) masked language model combined with Levenshtein Distance (LD) for identifying and correcting various spelling errors. Spelling errors were categorized, and a comprehensive dataset was created. The results showed that the system effectively identified and corrected spelling mistakes. [9] developed a modern spell checker model designed for integration into search engines on e-commerce platforms, specifically for Turkish. The results indicated that the typo corrector performed successfully, even beyond search engine contexts. [10] introduced a two-stage framework using pre-trained language models for correcting Thai spelling errors. Character Edit Distance was applied as a post-processing step to improve corrections. Experiments with two standard datasets showed that the model corrected misspelled words with a 60% success rate. [11] proposed two Seq2Seq deep learning-based automatic spelling error detectors and correctors for social media comments, analyzing input words at the phonogram level. The hybrid system demonstrated effective performance. [12] developed a spelling correction system for Indonesian using the LD algorithm, addressing complex language variations and providing more contextually appropriate corrections. The system achieved a Precision rate of 92% and an F1 Score of 90%. [13] presented a multilingual spell checker model tailored to correct user queries for specific product needs. The model was implemented for auto-completion in Adobe product searches and various applications, outperforming general-purpose spell checkers on in-domain datasets. [14] proposed a model that combines spelling correction with query expansion to enhance search engine performance. Document titles were preprocessed, with Term Frequency-

Inverse Document Frequency weighting applied to terms. The LD algorithm corrected typos, and query expansion improved search results. The system was tested on a dataset of 2,045 entries, achieving an average Recall of 95.91%, Precision of 63.82%, and Non-Interpolated Average Precision of 86.29%. [15] used a Masked Language Model and Edit Distance to correct spelling errors, selecting candidates based on Recall (correction rate) metrics. Results showed significant improvement over previous studies. [16] introduced QazSpell, a Kazakh spell-check and auto-correction system. A large web crawl of noisy data was used to understand misspellings, with a substring alignment model applied to detect symmetric and asymmetric patterns in word-error pairs. The model performed well in generating correction suggestions. [17] introduced a new algorithm to improve spelling error detection in Indonesian. The algorithm has been begun by gathering and combining correct and incorrect sentences, employing Bi-LSTM networks and Multi-Head Attention mechanisms to capture sequence complexity. The model achieved an Accuracy rate of 92%. [18] described the development of a model for detecting misspelled Assamese words in digital content. This model considered the context of the expression when determining word spelling. It highlighted that certain Assamese words may have multiple meanings, and even if spelled correctly according to the dictionary, they may be unsuitable for the intended context of a particular expression. Two machine learning techniques, LSTM and BiLSTM, were employed, with the BiLSTM model achieving the highest accuracy rate of 89.52%. The results indicated that the proposed approach significantly enhanced spelling error detection, outperforming previous research on the Assamese language. [19] presented a spelling checker model for detecting misspelled Urdu words. The model utilized dictionary search and edit distance techniques to generate correct spelling suggestions. A hybrid approach incorporating ranking methods such as Soundex, Shapex, LCS, and N-gram was developed to identify the best candidate word. The system demonstrated high accuracy, achieving an F1 score of 94.02%. [20] proposed a model combining a neural network-based language model with an N-gram model to detect and correct specific Vietnamese spelling errors. In experiments, this model achieved an F1 score 1% to 14% higher than other neural network-based language models and was further compared with various N-gram models.

3. DATASET

The dataset has been created by analyzing spelling errors in at least 1 million comments on the Trendyol platform. The Turkish Spelling Dataset taken from the Hunspell library [21] has also been incorporated into the main dataset. Table 1 presents the main dataset, while Table 2 shows the dataset with the added Turkish Spelling Check Dataset.

Table 1. The main dataset

Dataset Name	Trendyol Reviews
Train Dataset	45000
Test Dataset	2500
Validation Dataset	2500

Table 2. The dataset with the added Turkish spelling check dataset

Dataset Name	Trendyol Reviews & Turkish Spelling Check Dataset
Train Dataset	75000
Test Dataset	3000
Validation Dataset	3000

To expand the spelling error dataset, neighboring letters on the Turkish Q keyboard have been identified and replaced with their respective main letters. The sample representation of turkish q keyboard characters and neighbors is given in Figure 1.

```

Self.ascii_char_neighbor_mapping =
    {'q': ['w', 'a', 's'],
     'w': ['q', 'e', 's', 'd'],
     'e': ['w', 'r', 'd', 's', 'f', '3'],
     'r': ['e', 't', 'f', 'd', 'g'],
     't': ['r', 'y', 'g', 'f', 'h'],
     'y': ['t', 'u', 'h', 'g', 'j'],
     'u': ['y', 'i', 'j', 'h', 'k'],
     'i': ['u', 'o', 'k', 'j', 'l'],
     'o': ['i', 'p', 'l', 'k'],
     'p': ['o', 'l', 'i'],
     'a': ['s', 'q', 'z', 'w', 'x'],
     's': ['a', 'd', 'w', 'x', 'e', 'q', 'z', 'c'],
     'd': ['s', 'f', 'e', 'c', 'r', 'w', 'x', 'v'],
     'f': ['d', 'g', 'v', 'r', 't', 'e', 'c', 'b'],
     'g': ['f', 'h', 't', 'b', 'y', 'r', 'v', 'n'],
     'h': ['g', 'j', 'y', 'n', 'u', 't', 'b', 'm'],
     'j': ['h', 'k', 'u', 'm', 'i', 'y', 'n'],
     'k': ['j', 'l', 'i', 'o', 'u', 'm'],
     'l': ['k', 'o', 'p', 'i'],
     'z': ['x', 'a', 's'],
     'x': ['z', 'c', 's', 'd', 'a'],
     'c': ['x', 'v', 'd', 'f', 's'],
     'v': ['c', 'b', 'f', 'g', 'd'],
     'b': ['v', 'n', 'g', 'h', 'f'],
     'n': ['b', 'm', 'h', 'j', 'g'],
     'm': ['n', 'j', 'k', 'h'],
     ' ': ['c', 'v', 'b', 'n', 'm']}
    
```

Fig 1. The sample representation of Turkish q keyboard characters and neighbors.

During dataset preparation, various error types, including omission, insertion, combination, substitution, transposition, splitting, and identity errors, have been considered. The types of errors considered are described below.

1. Omission: A letter is randomly removed from the word.
2. Addition: A letter is randomly added to the word.
3. Combination: A new word is created by combining two existing words.
4. Substitution: A randomly selected letter is replaced with one of its neighboring letters.
5. Transposition: The positions of two letters within the word are swapped.
6. Division: The word is randomly split into two parts.
7. Identity: The word is left unchanged.

As a result of the operations performed, four different datasets have become ready. The high-quality and diverse spelling error dataset is presented in Table 3.

Table 3. The high quality and diverse spelling error dataset

Error Text	Dataset	Error Distribution	Dataset Size	Output Size
A1	trendyol_reviews.csv	"identity": 0.10, "insertion": 0.16, "omission": 0.16, "substitution": 0.08, "transposition": 0.16, "combination": 0.16, "split": 0.13, "english": 0.02, "shuffle": 0.0, "minedit": 0.03, "phonetic": 0.0	45000	135000
A2	trendyol_reviews.csv	"identity": 0.10, "insertion": 0.10, "omission": 0.10, "substitution": 0.14, "transposition": 0.19, "combination": 0.19, "split": 0.13, "english": 0.02, "shuffle": 0.0, "minedit": 0.03, "phonetic": 0.0	45000	135000
A3	trendyol_reviews_data_and_h_unspell.csv	"identity": 0.10, "insertion": 0.16, "omission": 0.16, "substitution": 0.08, "transposition": 0.16, "combination": 0.16, "split": 0.13, "english": 0.02, "shuffle": 0.0, "minedit": 0.03, "phonetic": 0.0	75000	300000
A4	trendyol_reviews_data_and_h_unspell.csv	"identity": 0.10, "insertion": 0.10, "omission": 0.10, "substitution": 0.14, "transposition": 0.19, "combination": 0.19, "split": 0.13, "english": 0.02, "shuffle": 0.0, "minedit": 0.03, "phonetic": 0.0	75000	300000

4. METHODOLOGY

4.1 Transformers Architecture

Before the development of transformers, machine translation and sequential data processing tasks commonly used encoder-decoder architectures, typically employing Recurrent Neural Network (RNN), LSTM, or Gated Recurrent Unit cells. In these architectures, often referred to as Seq2Seq models, input tokens are encoded into hidden states through multiple RNN cells. These encoded hidden states are then combined before being passed to the decoder. Decoders can leverage all encoded data to predict output tokens using these hidden states. However, although the last hidden state from the encoder contains substantial meaning and context, this information may not be fully captured by the Seq2Seq decoder. This issue, known as the 'bottleneck problem,' implies that as the input sequence grows longer and more complex, the hidden state becomes less effective. The attention mechanism addresses this limitation by allowing simultaneous processing of multiple inputs. During this process, it generates weight matrices for each hidden state and computes a weighted sum of all previous encoder states, enabling the decoder to identify which hidden states to "pay more attention" to. Since Seq2Seq architecture processes input sequentially, the information from the previous hidden state ($t-1$) is required to calculate the token at the current time (t). This sequential processing may hinder the correction of final tokens, particularly in tasks with numerous erroneous tokens, such as spelling correction. The limitations of these earlier systems are overcome by the transformer architecture and its attention mechanism. The transformer, based on the encoder-decoder structure, incorporates fully connected layers and multi-head self-attention mechanisms. This structure minimizes performance degradation from long dependencies and enables parallel computation. Positional embeddings and multi-head self-attention encode additional information about token positions and inter-token relationships [22].

4.2 Zemberek Library

Zemberek's current goal is to offer a general (Natural Language Processing – NLP) framework for other, largely ignored Turkic Languages in addition to Turkish. At the moment, the framework offers fundamental NLP functions like spell checking, morphological parsing, stemming, word construction, word suggestion, word conversion from words written exclusively in American Standard Code for Information Interchange characters (also known as "deasciifier"), and syllable extraction. To facilitate the work of language developers, Zemberek externalizes some language data to text-based configuration files. However, externalization typically has a negative impact on performance and flexibility. As a result, certain details are retained in the code, such as special cases and the suffix production mechanism [23].

5. DETAILS OF THE SYSTEM

The Seq2Seq architecture based on the encoder-decoder structure has been designed for training the obtained dataset. The variations of the created dataset have been diversified using different options. The constructed architecture has been trained in various configurations with different hyperparameters. Figure 2 illustrates the transformer architecture.

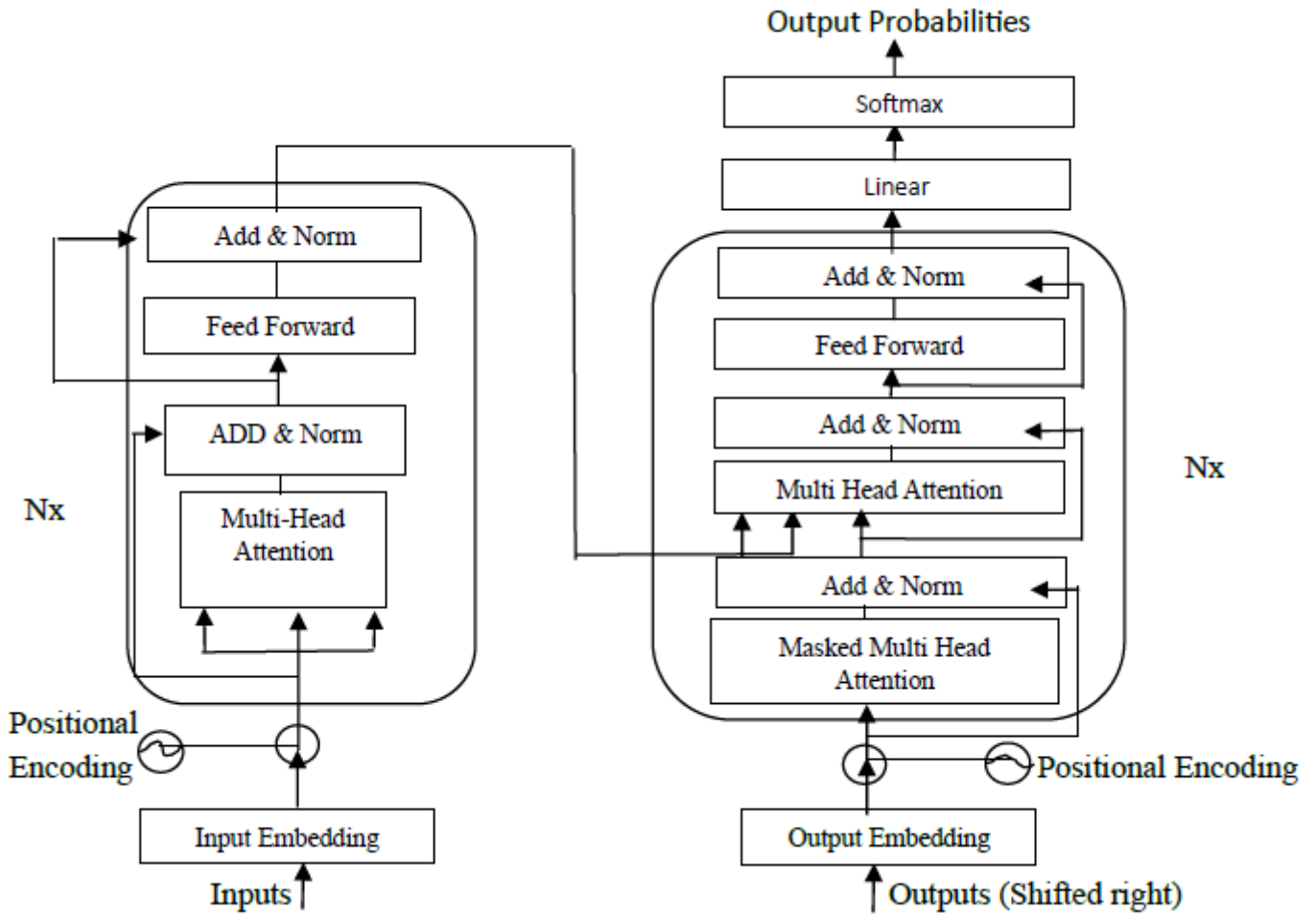


Fig 2. Transformer architecture

Models for spelling error detection and correction have been developed. Models A1, A2, A3, and A4 have been developed for four distinct datasets. The parameters such as hidden dimension, encoder/decoder layers, number of epochs and number of errors have been evaluated in different combinations. In addition, the influence of these parameter values on model performance has been systematically examined. In this context, eight different models have been developed. The parameter values of the developed models are presented in Table 4.

Table 4. The parameter values of the developed models

Model Name	Dataset	Hid Dim.	Enc/Dec. Layers	Epoch	Num of Errors
A1-1	A1	128	4/2	8	2
A1-2	A1	256	6/4	10	2
A2-1	A2	128	4/2	8	2
A2-2	A2	256	6/4	10	2
A3-1	A3	128	4/2	8	2
A3-2	A3	256	6/4	10	2
A4-1	A4	128	4/2	8	2
A4-2	A4	256	6/4	10	2

Additionally, the performance of the model in the Zemberek library has been tested as an alternative spelling corrector for the developed solution. The validation dataset has been provided as input to this module, and the results obtained have been compared.

6. RESULTS AND DISCUSSION

The prediction performance of the developed models has been evaluated with the Accuracy metric. Accuracy has been chosen because the problem at hand is a classification problem. It is widely recognized in the literature as a frequently used and reliable metric for such problems. The Accuracy values of the models obtained using different datasets are presented in Table 5.

Table 5. The Accuracy values of the models

Method/Achievement (Accuracy)	Trendyol Reviews-1	Trendyol Reviews-2	Trendyol Reviews & Turkish Spelling Check Dataset-1	Trendyol Reviews & Turkish Spelling Check Dataset-2
A1-1	%78.40	%75.15	%69.50	%68.37
A1-2	%75.75	%71.45	%69.17	%68.98
A2-1	%77.40	%79.10	%70.16	%67.17
A2-2	%73.21	%75.15	%69.35	%70.15
A3-1	%79.45	%79.63	%85.72	%84.37
A3-2	%77.16	%77.72	%80.64	%84.02
A4-1	%79.08	%78.98	%83.27	%81.29
A4-2	%77.02	%76.17	%82.27	%82.79
Model in the Zemberek	%54.48	%55.31	%61.27	%61.74

- Models within the A3 series consistently achieved higher Accuracy rates relative to their counterparts.
- The A3-1 model stood out as the most successful model by providing 85.72% Accuracy with the Trendyol Reviews & Turkish Spelling Check Dataset-1 and 84.37% Accuracy with the Trendyol Reviews & Turkish Spelling Check Dataset-2. The closest performance to this model has been shown by A3-2 with 84.02% Accuracy in the Trendyol Reviews & Turkish Spelling Check Dataset-2.
- The A1-1 and A1-2 models lag behind the leading A3 models, recording Accuracy rates of 78.40% and 75.75%, respectively.
- The Accuracy rates obtained through the model in the Zemberek library, specifically 54.48%, 55.31%, 61.27%, and 61.74%, are markedly lower than those of the other models. This finding suggested that the model in the Zemberek library is less effective for the datasets under consideration.
- The prediction models developed using the Trendyol reviews datasets exhibited similar performance.
- The models developed with the Trendyol Reviews & Turkish Spelling Check Datasets delivered more accurate forecast results than those developed with the Trendyol reviews datasets.
- It has been observed that the forecast models developed with the Trendyol Reviews & Turkish Spelling Check Dataset-1 and Trendyol Reviews & Turkish Spelling Check Dataset-2 datasets showed similar performance.

Overall, the A3 series models, particularly A3-1 and A3-2, achieved the highest performance. Incorporating the Trendyol Reviews & Turkish Spelling Check Dataset appears to enhance overall accuracy rates, while the Zemberek library model demonstrates lower performance across all models, failing to produce sufficient results for these datasets. When the results obtained with the developed prediction models have been examined, it has been observed that the transformer-based architecture exhibits satisfactory performance in detecting and correcting spelling errors.

An example of a comment processed by the developed system is provided below. The first sentence displays the raw comment received as input from the user. The second sentence shows the corrected version of this comment, with spelling errors addressed by the system.

“2 tne aldım denemek için çok güzeldi 3 tne daha sipariş verdim stoklamak için teşekkürler.”

“2 tane aldım denemek için çok güzeldi 3 tane daha sipariş verdim stoklamak için teşekkürler.”

Sample outputs have been obtained using FastAPI Swagger. The screenshot of the outputs is shown in Figure 3.








Fig 3. FastAPI swagger example outputs

The developed system contributed to improving the customer experience by correcting typos in comments. These findings offer a significant foundation for evaluating the efficacy of the proposed methodologies and highlight opportunities for further refinement and optimization in future research endeavors.

7. CONCLUSION

E-commerce has experienced rapid growth in recent years and continues to expand dynamically. To maximize customer satisfaction in this sector, accurately identifying customer needs and developing strategies to meet demand are essential. Customer reviews are valuable for e-commerce platforms; however, they frequently contain typographical errors. In this study, a system has been developed to automatically correct typos in user comments on products sold on the e-commerce site Trendyol.com and models for spelling error detection and correction have been developed. Model performance has been evaluated using the Accuracy metric, revealing that the A3 series models demonstrated the highest performance. This study presents a more up-to-date system compared to previous research. The effect of using Turkish Spelling Check Dataset has been analyzed. In addition, the performance of the developed models has been also compared with the performance of the model in the Zemberek library. The study emphasizes the importance of spelling error detection and provides valuable insights for e-commerce businesses.

ORCID

Okan ÇİFTÇİ  <https://orcid.org/0000-0002-9435-8980>
 Sumru NAYİR  <https://orcid.org/0009-0003-4782-3063>
 Emre Tolga AYAN  <https://orcid.org/0000-0002-4894-2190>
 Ceren ULUS  <https://orcid.org/0000-0003-2086-6381>
 M. Fatih AKAY  <https://orcid.org/0000-0003-0780-0679>

REFERENCES

- [1]. Sanbella, L., Van Versie, I., & Audiah, S. (2024). Online Marketing Strategy Optimization to Increase Sales and E-Commerce Development: An Integrated Approach in the Digital Age. *Startuppreneur Business Digital (SABDA Journal)*, 3(1), 54-66.
- [2]. Wang, Z., Zhu, Y., He, S., Yan, H., & Zhu, Z. (2024). LLM for sentiment analysis in e-commerce: A deep dive into customer feedback. *Applied Science and Engineering Journal for Advanced Research*, 3(4), 8-13.
- [3]. Aytan, B., & Şakar, C. O. (2023). Deep learning-based Turkish spelling error detection with a multi-class false positive reduction model. *Turkish Journal of Electrical Engineering and Computer Sciences*, 31(3), 581-595.
- [4]. Dutta, A., Polushin, G., Zhang, X., & Stein, D. (2024). Enhancing E-commerce Spelling Correction with Fine-Tuned Transformer Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 4928-4938). <https://doi.org/10.1145/3637528.3671625>
- [5]. Isbarov, J., Huseynova, K., & Rustamov, S. (2024, April). Robust automated spelling correction with deep ensembles. In *Proceedings of the 2024 8th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence* (pp. 26-30). <https://doi.org/10.1145/3665065.3665070>

- [6]. Kakkar, V., Sharma, C., Pande, M., & Kumar, S. (2023, July). Search Query Spell Correction with Weak Supervision in E-commerce. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track) (pp. 687-694).
- [7]. Kuznetsov, A., & Urdiales, H. Spelling correction with denoising transformer. arXiv 2021. arXiv preprint arXiv:2105.05977.
- [8]. Naziri, A., & Zeinali, H. (2024). A Comprehensive Approach to Misspelling Correction with BERT and Levenshtein Distance. arXiv preprint arXiv:2407.17383.
- [9]. Oral, E., Mancuhan, K., Erdem, H. V., & Hatipoglu, P. E. (2024, May). Turkish Typo Correction for E-Commerce Search Engines. In Proceedings of the Seventh Workshop on e-Commerce and NLP@ LREC-COLING 2024 (pp. 65-73).
- [10]. Pankam, I., Limkonchotiwat, P., & Chuangsuwanich, E. (2023, June). Two-stage Thai Misspelling Correction based on Pre-trained Language Models. In 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 7-12). IEEE.
- [11]. Ratnam, D. J., Karthika, A. N., Praveena, K., Taniya, R., Thara, S., & Prema, N. (2024). Phonogram-based Automatic Typo Correction in Malayalam Social Media Comments. *Procedia Computer Science*, 233, 391-400.
- [12]. Santoso, J. T., & Yan, S. (2024). A Hybrid Approach to Typo Correction in Indonesian Documents Using Levenshtein Distance. *Journal of Technology Informatics and Engineering*, 3(2), 151-168.
- [13]. Sharma, S., Valls-Vargas, J., King, T. H., Guerin, F., & Arora, C. (2023, July). Contextual multilingual spellchecker for user queries. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 3395-3399).
- [14]. Soyusiawaty, D., & Wolley, D. H. R. (2021). Hybrid spelling correction and query expansion for relevance document searching. *International Journal of Advanced Computer Science and Applications*, 12(8).
- [15]. Tohidian, F., Kashiri, A., & Lotfi, F. (2022, November). BEDSpell: Spelling Error Correction Using BERT-Based Masked Language Model and Edit Distance. In International Conference on Service-Oriented Computing (pp. 3-14). Cham: Springer Nature Switzerland.
- [16]. Toleu, A., Tolegen, G., Mussabayev, R., Krassovitskiy, A., & Ualiyeva, I. (2022). Data-driven approach for spellchecking and autocorrection. *Symmetry*, 14(11), 2261.
- [17]. Yanfi, Y., Soeparno, H., Setiawan, R., & Budiharto, W. (2024). Multi-head Attention Based Bidirectional LSTM for Spelling Error Detection in the Indonesian Language. *IEEE Access*.
- [18]. Phukan, R., Neog, M., & Baruah, N. (2023, July). A Deep Learning Based Approach For Spelling Error Detection In The Assamese Language. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.
- [19]. Aziz, R., Anwar, M. W., Jamal, M. H., & Bajwa, U. I. (2021). A hybrid model for spelling error detection and correction for Urdu language. *Neural Computing and Applications*, 33, 14707-14721.
- [20]. Tien, D. N., Minh, T. T. T., Vu, L. L., & Minh, T. D. (2022). Vietnamese Spelling Error Detection and Correction Using BERT and N-gram Language Model. In *Intelligent Systems and Networks: Selected Articles from ICISN 2022, Vietnam* (pp. 427-436). Singapore: Springer Nature Singapore.
- [21]. Al-Hussaini, L. (2017). Experience: insights into the benchmarking data of hunspell and aspell spell checkers. *Journal of Data and Information Quality (JDIQ)*, 8(3-4), 1-10.
- [22]. Do, D. T., Nguyen, H. T., Bui, T. N., & Vo, H. D. (2021). Vsec: Transformer-based model for vietnamese spelling correction. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18* (pp. 259-272). Springer International Publishing.
- [23]. Akin, A. A., & Akin, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure*, 10, 1-5.