Avrupa Bilim ve Teknoloji Dergisi
Sayı 52, S. 37-45, Kasım 2023

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 52, pp. 37-45, November 2023

Research Article

Şikayet İçeren Müşteri Yorumlarının Tespiti ve Sınıflandırılması

Elif Ayanoğlu[1*], Zeynep Çolak[2], Toygar Tanyel[3], Hasan Yunus Sarıoğlu[4], Banu Diri[5]

[1*] Yıldız Technical University, Faculty of Electrical and Electronics, Department of Computer Engineering, İstanbul, Türkiye, (ORCID: 0009-0000-1340-4870), elif.ayanoglu@std.yildiz.edu.tr

[2] Yıldız Technical University, Faculty of Electrical and Electronics, Department of Computer Engineering, İstanbul, Türkiye, (ORCID: 0009-0004-8515-3109), zeynep.colak@std.yildiz.edu.tr

[3] Yıldız Technical University, Faculty of Electrical and Electronics, Department of Computer Engineering, İstanbul, Türkiye, (ORCID: 0000-0002-2421-6880), toygar.tanyel@std.yildiz.edu.tr

[4] Yıldız Technical University, Faculty of Electrical and Electronics, Department of Computer Engineering, İstanbul, Türkiye, (ORCID: 0009-0000-5566-6190), yunus.sarioglu@std.yildiz.edu.tr

[5] Yıldız Technical University, Faculty of Electrical and Electronics, Department of Computer Engineering, İstanbul, Türkiye, (ORCID: 0000-0002-4052-0049), diri@yildiz.edu.tr

(İlk Geliş Tarihi 10 Mart 2023 ve Kabul Tarihi 24 Eylül 2023)

(**DOI:** 10.5281/zenodo.10254498)

ATIF/REFERENCE: Ayanoğlu, E., Çolak, Z., Tanyel, T., Sarıoğlu, H.Y. & Diri, B. (2023). Detection and Classification of Customer Comments Containing Complaints. *European Journal of Science and Technology*, (52), 37-45.

Öz

Şikayet içeren müşteri yorumlarının tespiti ve sınıflandırılması sistemi, e-ticaret sitelerindeki ürünlere yapılmış olan olumsuz yorumlarda, yorumu yapan kişinin aslında ürünün veya hizmetin tam olarak hangi özelliğinden şikayetçi olduğunu tespit etmek amacıyla geliştirilmiştir. Çalışmanın ilk aşamasında ürün hakkında yapılan bir yorumun olumlu veya olumsuz olup olmadığı %95 doğrulukla tespit edilmiştir. İkinci aşamasında da yapılan olumsuz yorumun yazarların belirlemiş olduğu 5 adet kategoriden hangisine dahil olduğu tespit edilmeye çalışılmıştır. Seçilen kategorileri ifade eden önceden belirlenmiş anahtar kelimelerin Word2Vec ile çıkarılmış kelime vektörleri ile yorum içerisinde geçen kelime ve BERT ile elde edilen cümle vektör değerleri arasındaki yakınlık kosinüs benzerliği ile ölçülerek yoruma ait olan şikayet kategorisi veya kategorileri belirlenmiştir. En başarılı yöntem Word2Vec ile çıkarılmış olan kelime vektörlerinin kullanıldığı yöntem olmuştur ve bu yöntemde yorumlara ait olan şikayet kategorisi tek etiketli yorumlar için %82,5, iki etiketli yorumlar için de %82 doğrulukla tespit edilmiştir.

Anahtar Kelimeler: Duygu Analizi, hedef tabanlı, BERT, Word2Vec, CNN, LSTM

Detection and Classification of Customer Comments Containing Complaints

Abstract

The detection and classification of customer reviews containing complaints system has been developed in order to determine exactly which feature of the product or service the person making the comment actually complains about in the negative comments made to the products on the e-commerce sites. In the first stage of the study, it was determined with 95% accuracy whether a comment about the product was positive or negative. In the second stage, it has been tried to determine which of the 5 categories that the authors have determined is included in the negative comment. Complaint category or categories belonging to the comment were determined by measuring the closeness between the word vectors extracted with Word2Vec of the predetermined keywords expressing the selected categories, the word in the comment and the sentence vector values obtained with the BERT, by measuring the cosine similarity. The most successful method was the method using word vectors extracted with Word2Vec, and in this method, the complaint category belonging to the comments was determined with an accuracy of 82.5% for single-label comments and 82% for two-label comments.

Keywords: Sentiment Analysis, Aspect Based, BERT, Word2Vec, CNN, LSTM

[*] Sorumlu Yazar: elif.ayanoglu@std.yildiz.edu.tr

1. Introduction

The amount of content that is created through different channels (e.g. video, audio, text) is dramatically increased in social media. A wide range of content on social media comes from people who produce natural language, complaints, hence comments. Identifying customer comments that contain complaints and classifying them under certain headings is quite an important concern for companies. By making inferences from the existing customer comments, companies can make more reasonable decisions to improve their relations with their customers, and to determine their marketing strategies. Although there are many existing studies for different languages in this problem, the number of studies for Turkish is quite limited.

Sentiment analysis studies are carried out in two different stages as text/sentence or target category level. Text/sentence level analyses aim to figure out the emotional polarity of the idea in the text. Emotion polarity is evaluated according to two (positive, negative) or three (positive, negative and neutral) categories. Most of the existing studies are based on classification of the dominant emotion in the text/sentence. In some of the studies, Dictionary-based approaches have been used to determine the dominant emotion. (Akgül et al., 2016; Karamollaoğlu et al., 2018; Yurtalan et al., 2019) utilized Twitter data, (Vural et al., 2013; Dehkharghani et al., 2017; Dehkharghani, 2018; Uslu et al., 2019) worked with movie reviews, (Öztürk and Ayvaz, 2018) investigated the issue of Syrian crisis and refugees on Twitter, (Karaöz and Burcu, 2018) used TV program comments from Twitter, and (Atan and Çınar, 2019) have worked on newspaper financial news. (Velioğlu et al., 2018; Shehu et al., 2019; Karcıoğlu and Aydın, 2019) also utilized Twitter data, (Shehu and Tokat, 2020; Uslu et al., 2019; Akba et al., 2014; Kaynar et al., 2016) investigated film reviews, (Parlar et al., 2018) worked on both film reviews and product reviews of Hepsiburada.com, (Nizam and Akın, 2014), (Nalçakan et al., 2015) used product comments on Twitter, (Öztürk et al., 2017) analyzed student comments made on Twitter, (Demirci et al., 2019), (Kaya et al., 2012) have worked on the prediction of emotion using machine learning methods on political events.

Deep learning architectures have also been used in some studies in the field. (Ciftci and Apaydin, 2018) RNN (Recurrent Neural Network) on product and movie reviews from Hepsiburada.com and Beyazperde, (Santur, 2019) GRU (Gated Recurrent Unit), (Ahmetoğlu and Resul, 2020) used the BERT (Bidirectional Encoder Representations from Transformers) model to predict mood on hotel reviews and product and movie reviews from RNN, (Acikalin et al., 2020) Hepsiburada.com and Beyazperde.

Sentiment Analysis studies are called “Aspect Based Sentiment Analysis” if they are carried out to determine the emotional state in the target category, that is, the sub-categories of an event. There are very few studies for Turkish literature in this area. (Kama et al., 2017; Karagoz et al., 2019) determined the target features in the sentence using the comments in the “DonanımHaber” which is one of the first websites built on computer technology in Turkey, and then they scored the emotional states. (Bayraktar et al., 2019) proposed a holistic method using statistical, linguistic and rule-based approaches for sentiment analysis based on Turkish lingustics on Turkish restaurant data set shared within the scope of SemEval-Aspect Based Sentiment Analysis (ABSA-2016). (Çetin and Eryiğit, 2018) worked on determining the target category and target term, the category and target at the same time, and the emotion class, based on the Turkish restaurant comments shared at SemEval ABSA-2016 competition. For the first three tasks, they tried to solve the problem in one step by using CRF (Conditional Random Fields) with word vectors and natural language processing outputs.

The aim of this study is to report the complaints about the product under the specified headings among the negative comments made about any clothing brand. Section 2 introduces the utilized data set, section 3 includes the methods used are briefly mentioned, section 4 provides information about the proposed system, section 5 and section 6, discuss the success of the system and its positive/negative points.

2. Material & Methods

2.1. Dataset

The data set used in this study was prepared by retrieving the customer comments published on Trendyol between 17th of March 2022 and 9th of April 2022. Since the developed system will be modeled for the clothing category, only the comments of the products belonging to this category are taken. In evaluation, users can give 1 to 5 stars, and make comments on Trendyol's website. While we automatically labeled the emotional states of the comments within the scope of the study, we gave negative labels for 1 and 2-star comments, and positive labels for 4 and 5-star comments. Ultimately, the data set contains 103,303 comments, 52,754 of comments were positive which labeled as “1”, and 50,544 of the comments were negative which labeled as “0”. Some samples from the dataset are shown in Table 1.

2.2. Methods

In this section, brief information will be given about Word2Vec and BERT, which are used to obtain the vector values of words and sentences, and machine and deep learning approaches as classification methods.

2.2.1. Vector Acquisition Techniques

Word2Vec is an unsupervised, prediction-based model used to represent words as vectors. Word2Vec creates word vectors from the text that it takes as input by using CBOW (Continuous Bag of Words) or Skip-Gram algorithms.

In CBOW, the model predicts each word in a text by using a context window that is the words close to the word to be predicted. The order of words in context window is not important in this approach because it assumes that similar words should be used in similar context and there is no need to keep the order of words to predict a word in a context.

In Skip-Gram, the model do not use a context to predict a word. It predicts the context using only the word in the center.

BERT (Bidirectional Encoder Representations from Transformers) is a relatively recent article by Google AI Language researchers (2018). BERT's ability to deliver state-of-the-art (best and most advanced) results on various NLP tasks has made a significant impact on the deep learning community.

BERT does not process words sequentially like RNNs and LSTMs so when we call BERT as a bidirectional model, we do not mean left-to-right and right-to-left. It processes all the words at the same time. The order of words disappears when we process words together but BERT uses positional embeddings to solve this problem.

BERT uses Masked-language Modeling (MLM) and Next Sentence Prediction (NSP) tasks in the training process. In MLM, the model predicts the masked words in a text. In NSP, it tries to predict whether the second sentence of input is actually comes after the first sentence or not. Therefore, it captures contextual information at word and sentence level. We utilized BERTurk which is a pre-trained version of BERT for Turkish.

Both Word2Vec and BERT were trained for some tasks but while they were training, they learned how to represent words. Thanks to their skills to create meaningful embeddings, we use them for different tasks by training just additional task-specific layer(s).

2.2.2. Machine Learning Techniques

In this study, machine learning techniques MNB (Multinomial Naive Bayes), SVM (Support Vector Machine) and LR (Logistic Regression) are used. Naive Bayes is a classification method which is a collection of many Bayes' Theorem algorithms. The multinomial model is used to classify data that cannot be represented numerically. Support Vector Machines are used in classification problems. Draws a line to separate points placed on a plane. SVM aims to have this line at the maximum distance within the points of its two classes. Logistic Regression is also utilized to classify categorical and numerical data. The kernel type for SVM was selected as 'rbf', gamma parameter as 'auto' and max_iter as 5000. For Logistic Regression, solver type is implemented as 'lbfgs' and max_iter is 5000. Completely default parameters are applied for MNB.

2.2.3. Deep Learning Techniques

Convolutional Neural Networks are known for their ability to extract as many features as possible from its inputs. Also, CNN reduces the overfitting by overriding the contribution of some neurons to the next layer. LSTM (Long Short Term Memory) keeps the chronological order between words in the document like RNNs. However, unlike RNNs, it uses input, forget and output gates so it has the ability to decide how much information should be transferred to next steps.

In this study, it is preferred to use a model that combines the strengths of LSTM with the feature extraction feature of CNN (Tanyel et al., 2022). In this study, CNN and LSTM are combined and used as a single model. The features extracted by CNN are used as input for LSTM (CNN to LSTM).

2.3. Recommended System

The developed system consists of two stages. In the first stage, it is decided whether a comment given to the system contains a complaint or not. In the second stage, it is determined which complaint category the comment found to contain a complaint belongs to. Each stage of the system is presented in Figure 1. Task codes, can be found at our GitHub repository https://github.com/elifayanoglu/complaint-classification.

2.3.1. Sentiment Analysis

We created word vectors using different techniques for both words and sentences for 103,303 comments whose class was set as 0 “negative” and 1 “positive”. Before extracting the word vectors, initially preprocessing steps are applied on the comments. The steps of removing unnecessary words (stop words), converting to lowercase, removing punctuation marks and emojis have been carried out. Spelling errors were corrected using the zemberek library. For the Word2Vec part, Tensorflow's Tokenizer function is used for tokenization. (Bu cümleden sonra BERT'te tokenization işlemini nasıl yaptık o gelecek)After obtaining the word vectors with BOW, we created our second new word vectors with Word2Vec, which we trained with our own dataset. Vector values for sentences are also extracted with the pre-trained BERTurk model. After providing the vectors, which are the output of the BOW model, providing as input to the MNB model. Word2Vector vectors were given to the model consisting of the combination of CNN and LSTM. Output vectors of the Word2Vec and BERTurk are given to SVM and another machine learning model, LR. Our aim in the experiments is to experience more than one model to compare the differences between the results and to identify the model that works best for our data set.

2.3.2. Identifying Subcategories

We decided to continue with an unsupervised learning method since the comments containing complaints do not have a category label, but we observed that the clustering methods we chose did not yield successful results as we expected. For this reason, we preferred a rule-based classification method using cosine similarity, which measures the similarity between two vectors (1).

$$cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{||A|| \, ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 B_i^2} \qquad (1)$$

The goal is to categorize negative comments therefore, half of the dataset is used at this stage and the label information of the categories was not available in the dataset. To implement the proposed system, first the dataset was examined in detail and 5 complaint categories, which were mostly mentioned in the comments, were determined. These categories are named as "fabric and stitch", "model and size", "colour", "image relevance", "cargo and delivery". Then, the 96 words that were mentioned the most in the negative comments in the data set was detected. These words were carefully selected by examining the 50K negative comments in the dataset. It was determined which words were most commonly used to express complaints for each category. Words that pass above a certain frequency are included in the list. These words are used as keywords to define the categories we have determined. A total of 31 keywords were defined for the "fabric and stitch" category, 8 for the "colour" category, 29 for the "model and size" category, 16 for the "image relevance" category, and 12 for the "cargo and delivery" category. These keywords are as presented in Table 3. We used these keywords directly to evaluate our Word2Vec-based system. But, since it makes more sense to get sentence embedding from BERT than word embedding, we created a small list of negative comments that include our keywords to evaluate our BERT-based system.

In this section, first of all, the Word2Vec vectors of the keywords representing our target categories and the Word2Vec vectors of the words belonging to the negative comments to which category they belong to were obtained. Afterwards, the following operations are performed: Cosine similarity is checked between the vector of the first word belonging to the first category from our 5 categories and all the words of the sentence to be categorized. The highest cosine similarity value is retained. This process is repeated for all keywords of the first category. The obtained 31 (the number of fabric/stitch category keywords) similarity values higher than the threshold are counted and divided by 31 to normalize. The same process is repeated for the other categories. The number we divide to normalize is the number of keywords in each category. Thus, the 5 values obtained were examined. The category with the highest value was assigned as the first label. If the second highest value was not much lower than the first, it was processed and labeled as the second label. While performing this process, if no similarity value remains above the threshold, it is labeled as "does not belong to any category".

In the BERT section, where we work on a sentence basis, the process changes as follows: The keywords belonging to the categories are now converted into sentences, and the similarity between the vector of these sentences and the negative comment vector that is desired to be categorized is checked.

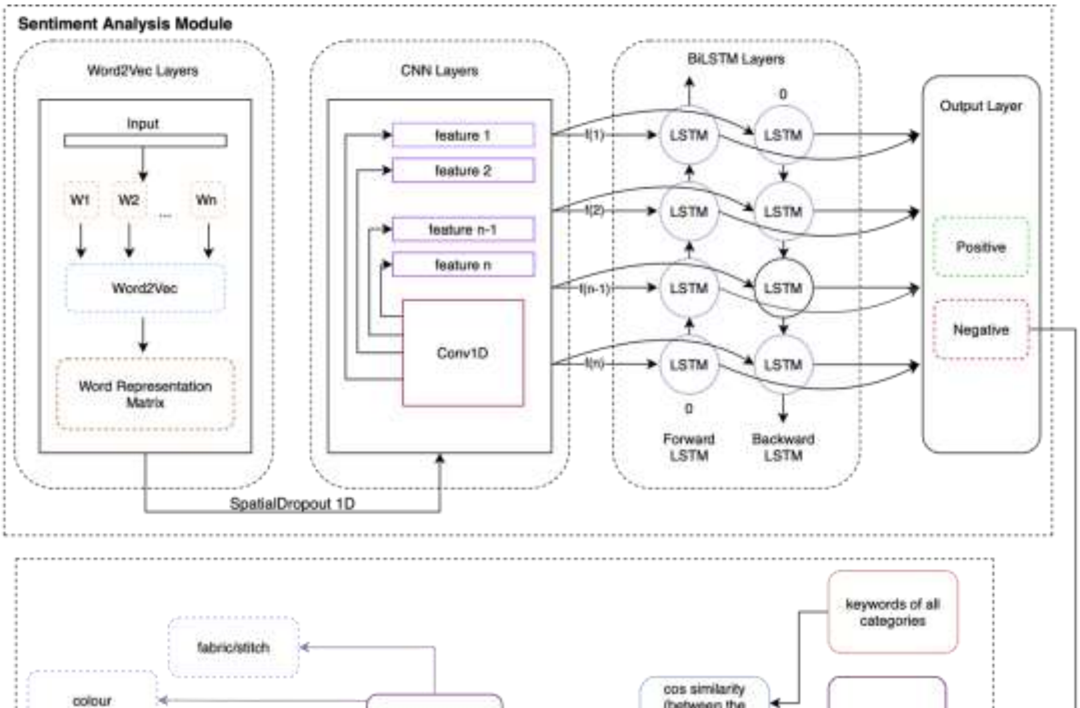

Figure 1. System work flow.

When we investigated the points where the system failed, the following conclusions were drawn for Word2Vec and BERT.

Evaluation of Word2Vec for Subcategory Detection.

Example - 1:

TR - "beğenmedim kumasni filan"

EN - "I didn't like the fabrci or somethin"

Prediction: "Does not belong to any category"

Word2Vec is a weak model against typos. While the comment in the example-1 should be in the "fabric and sewing" category, an important word in the sentence was misspelled, so the system could not find a word that has high cosine similarity with the keywords, and the sentence was evaluated incorrectly. While some typos can be corrected with the Zemberek library, some of them could not be corrected as in this example.

Example - 2:

TR - "gerçekten pes beden olarak xl seçmeme rağmen birim gelmiş biri l bravoooooo iadede edemiyorum şimdi ne yapacağım ben bunları ? tebrikler trendyol"

EN - "I really can't believe it. Although I chose XL for the size, one of them came in L, bravoooooo I can't send them back, what should I do now? Congratulations trendyol"

Prediction: "fit and size"

If the complaint comment is written in long and rarely used patterns, it can be miscategorized. In the example-2, although the size is mentioned in the comment, the user is actually trying to explain that the wrong size was sent to him. This is actually a "cargo and delivery" issue. In addition, the system can assign two tags for a comment. However, the second label could not be assigned because the cosine similarity calculated in this interpretation was below the specified threshold value.

Table 3. Keywords for categories

Category	Keywords
Kumaş ve dikiş	ince, incecik, inceydi, yırtık, delik, çekti, yırtıldı, sökük, kalitesi, kaliteli, kalitesiz, gösteriyor, naylon, kalın, yamuk, dandik, dikiş, dikişleri, dikişi, dikmişler, defolu, terleten, terletiyor, terletecek, terletir, kumaş, kumaşı, kumaşını, kumaşın, küçüldü, kayboldum
Renk	rengi, rengini, renginin, soluk, solmuş, soluyor, soldu, canlı
Kalıp ve beden	bol, boldu, büyük, küçük, dardı, dar, geniş, pot, potluk, kesiminde, kesiminden, oversize, kesim, kesimi, kesimleri, kolları, uzun, kısa, boyu, kalıbı, kalıp, kalıbını, beden, bedeni, bedenim, bedene, bedenler, bedenleri
Görselle alaka	alakası, fotoğraf, fotoğrafta, fotoğraftaki, fotoğraftakinden, fotoğrafla, fotoğraftakiyle, göründüğü, görseldeki, görseldekiyle, görselle, görsel, resimdeki, resimdekinden, resimdekiyle, resimde
Kargo ve teslimat	teslimat, yavaş, geç, paketleme, kargo, leke, lekeli, etiketsiz, kusurlu, yanlış, eksik, yerine

Example - 3:

TR - "kalıbını beğenmeyip iade ettim ."

EN - "I didn't like the fit and returned it."

Prediction: "fit and size & fabric and sewing "

In example sentence-3: "I returned it because I didn't like its fit."

System-provided labels: Label one: "Fit and Size", Label two: "Fabric and Sewing".

The categories of 'Fit and Size' and 'Fabric and Sewing' are the ones that are most frequently confused with each other. It is possible for a sentence that should belong to only one category to be mistakenly labeled as both 'Fabric and Sewing' and 'Fit and Size' at the same time.

Example - 4:

TR - "ürün genel anlamda duruşu kumaşı güzel fakat lekeli geldi keşke kontrol etseydiniz ."

EN - "overall, the product's fit and its fabric is good but it came stained I wish you had checked it."

Prediction: "cargo and delivery & fabric and sewing"

Since both the complained side and the good side of the product are described in the sentence, the positive description of the product is also labeled as a complaint.

Apart from the above examples, the accuracy of the system is high. The reason for this is that enough keywords are defined for the categories. In addition, Word2Vec's analysis of word-based similarities allowed it to yield very good results. An example of a successful conclusion is given in example-5.

Example - 5:

TR - "bu lekeler nedir ya . hiç mi kontrol eidlmiyor . xs aldığım halde omuzları oturması . beden çok bol"

EN - "What are these stains. is it not contrloled at all. even though I bought xs, shoulders don't fit. size is too large"

Evaluation of BERT for Subcategory Detection.

Example - 6:

TR - "S aldım. 170 boy 56 kg. Ama aşırı büyük oldu.. Üzülerek iade ediyorum"
EN - "I bought an S. 170 height 56 kg. But it was too large.. I'm sorry to return it"
Prediction: "Does not belong to any category."
The prediction of example-6 was incorrect because the similarity between the vector of the sentence and the BERT vectors of the sentences representing the "fit and size" category was not high enough.

Example -7:

TR - "Ürünün kumaşı kalın ve güzel ama kalıbı o kadar kötü ki . Giyilecek gibi değil ."
EN - "The fabric of the product is thick and beautiful, but its fit is so bad. It is not suitable for wearing."
Prediction: "cargo and delivery & color"
In example-7, the customer is satisfied with the fabric but complains about the fit of the product. However, the system also perceives the positive part as a complaint.

Example - 8:

TR - "her yerinde kalem izleri var m neden değil sanki kalıbı da aşırı büyük değişim istiyorum"
EN - "It have pen marks all over it why it is not m as it is too large for m I want to change it"
Prediction: "Does not belong to any category."
There are 4 sentences in the example-8, but a conjunction or punctuation is not used between the sentences. In this case, the expression "it is too large for m" could not affect the vector of the sentence as much as it should, and the comment could not be labeled with the "fit and size" category.

***Example - 9*:**

TR - "tek kullanımlık yıkandıktan sonra giyilmez"
EN - "it is for single use cannot be worn after washing"
Prediction: "fabric and sewing"
In the example-9, the poor quality of the fabric is indirectly explained. However, the system using BERT vectors and cosine similarity correctly predicted the label of this interpretation.

Example - 10:

TR - "ürün aşırı bol salaş görseldekine aldanmayın ama iade işiyle uğraşmayacağım"
EN - "the product is too baggy shabby don't be fooled by the image but I will not bother to return it"
Prediction: "image relevance & fit and size"
Successful estimation of both complaint tags with the system using BERT vectors can be seen in example-10.

3. Results & Discussion

3.3.1. Results of the Sentiment Analysis Stage

The success of the sentiment analysis of the model, which is a combination of the BOW model and the MNB model, was found to be 92%. When the Word2Vector vectorizer was added to the model consisting of a combination of CNN and LSTM, the classification success was 95%. Output vectors of Word2Vec and BERTurk were given to SVM and the model achieved 88% and 92% accuracy, respectively. Then, using the vector outputs of Word2Vec and BERTurk and another machine learning model, LR, 86% and 91% rates were achieved. As a result, 95% success was achieved with W2V + CNN + LSTM. The accuracy, f1 score, precision, recall, and AUC (Area Under the ROC Curve) of all methods are presented in Table 2.

Table 2. Results of the models

	Accuracy	F1 Score	Precision	Recall	ROC_AUC
W2V + CNN + LSTM	0.95	0.95	0.95	0.95	**0.95**
BERT + SVM	0.92	0.92	0.92	0.93	0.92
W2V + SVM	0.88	0.88	0.88	0.89	0.88
BERT + LR	0.91	0.91	0.91	0.91	0.91
W2V + LR	0.86	0.87	0.86	0.86	0.86
BOW + Multinomial NB	0.91	0.92	0.91	0.92	0.92

3.3.2. Results of Subcategories Determination Stage

To test the system we proposed, first of all, we chose 200 random negative comments with a single category in order to measure the success of Word2Vec and BERT separately. Then, we labeled these comments manually and then got the results with Word2Vec and BERT. Since we obtained the most successful results with Word2Vec in the sentiment analysis, which is the first stage of the system, only the error matrices of the operation performed with Word2Vec are shown while the results of the first stage are shared. The F1 scores of these error matrices, which can also be seen in Table 4, are as follows: 93% for "fabric and sewing" and "fit and size" categories, 96% for "color", 100% for "image relevance" and 88% for "cargo-delivery".

Some comments may contain sentences related to more than one category. In the dataset we use, each comment has a maximum of two categories. The number of correct detections of all categories using both Word2Vec and BERT is as seen in Table 5. While Word2Vec was more successful when there were two different categories in the comment, BERT was more successful in predicting a single category.

Table 4. Error matrix for all categories of single-label comments with Word2Vec

	Category						
Prediction	**Fabric**			Prediction	**Size**		
		True				True	
		Fabric	Other			Size	Other
	Fabric	62	6		Size	52	1
	Other	4	128		Other	7	140
Prediction	**Colour**			Prediction	**Relevance to the Image**		
		True				True	
		Colour	Other			Image	Other
	Colour	12	1		Image	16	0
	Other	0	187		Other	0	183
Prediction	**Cargo and Delivery**						
		True					
		Cargo	Other				
	Cargo	23	2				
	Other	4	171				

Table 5. Estimation by number of Word2Vec and BERT tags (for the selected 200 comments)

	Number of Comments		
Method	Two tags are correct	Only one tag is correct	Both tags are wrong
Word2Vec	164	36	0
BERT	63	115	12

4. Conclusion

In this study, a complaint detection and classification method for the Turkish language has been proposed, considering that it is critical for companies to ensure customer satisfaction in the developing competitive environment.

In the first stage of the study, the comments written by the customers about the product were divided into two categories, with and without complaints, with a success rate of 95%. In the second stage, the categories of comments containing complaints were tried to be detected. During the category detection phase, Word2Vec and BERT were used for word vectors and sentence vectors respectively. Categorization with Word2Vec vectors yielded more successful results than with BERT vectors. The success of the operation using Word2Vec was measured by the F1 score. The F1 score for single-label comments is 0.94. For double-labeled comments, the F1 score was 0.82.

While the successes with Word2Vec were higher in sentences containing common words, BERT gave more successful results if the complaint was expressed indirectly or if common words were not used.

In this study, when categorizing on a sentence basis, that is, when using BERT vectors comments consisting of more than one sentence were evaluated as a single sentence. However, if there is more than one sentence in the comments, using the sentences separately, can make the comment more similar to the key sentences. However, the fact that users generally do not use punctuation and conjunctions in comments complicates this situation.

Another point where the system fails is that the positive side of the comment is also perceived as a complaint. For the solution of this problem, it may be useful to divide the comment into sentences using the conjunctions in the comment. If the conjunction in the sentence is a conjunction that combines opposite meanings, such as "but" or "however", processing the sentences separately, can be a solution.

In addition, it was observed in this study that spelling mistakes affect success. While correcting the English keyboard characters by using the normalization function in the Zemberek library increased the scores, it was not successful at the desired level for other spelling mistakes.

References

Acikalin, U. U., Bardak, B., & Kutlu, M. (2020, October). Turkish sentiment analysis using bert. In *2020 28th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Tanyel, T., Alkurdi, B., & Ayvaz, S. (2022, September). Linguistic-based Data Augmentation Approach for Offensive Language Detection. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 1-6). IEEE.

Ahmetoğlu, H., & Resul, D. A. Ş. (2020). Türkçe Otel Yorumlarıyla Eğitilen Kelime Vektörü Modellerinin Duygu Analizi ile İncelenmesi. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *24*(2), 455-463.

Akba, F., Uçan, A., Sezer, E. A., & Sever, H. (2014, July). Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews. In *8th European Conference on Data Mining* (Vol. 191, No. 2002, pp. 180-184).

Akgül, E. S., Ertano, C., & Diri, B. (2016). Twitter verileri ile duygu analizi. *Pamukkale University Journal of Engineering Sciences*, *22*(2).

Suat, A. T. A. N., & ÇINAR, Y. (2019). Borsa İstanbul'da finansal haberler ile piyasa değeri ilişkisinin metin madenciliği ve duygu (sentiment) analizi ile incelenmesi. *Ankara Üniversitesi SBF Dergisi*, *74*(1), 1-34.

Bayraktar, K., Yavanoglu, U., & Ozbilen, A. (2019, December). A rule-based holistic approach for Turkish aspect-based sentiment analysis. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2154-2158). IEEE.

Çetin, F. S., & Eryiğit, G. (2018). Türkçe hedef tabanlı duygu analizi için alt görevlerin incelenmesi–hedef terim, hedef kategori ve duygu sınıfı belirleme. *Bilişim Teknolojileri Dergisi*, *11*(1), 43-56.

Ciftci, B., & Apaydin, M. S. (2018, September). A deep learning approach to sentiment analysis in Turkish. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*(pp. 1-5). IEEE.

Dehkharghani, R. (2018). A hybrid approach to generating adjective polarity lexicon and its application to turkish sentiment analysis. *International Journal of Modern Education and Computer Science*, *11*(11), 11.

Dehkharghani, R., Yanikoglu, B., Saygin, Y., & Oflazer, K. (2017). Sentiment analysis in Turkish at different granularity levels. *Natural Language Engineering*, *23*(4), 535-559.

Demirci, G. M., Keskin, Ş. R., & Doğan, G. (2019, December). Sentiment analysis in Turkish with deep learning. In *2019 IEEE international conference on big data (big data)* (pp. 2215-2221). IEEE.

Kama, B., Ozturk, M., Karagoz, P., Toroslu, I. H., & Kalender, M. (2017, November). Analyzing implicit aspects and aspect dependent sentiment polarity for aspect-based sentiment analysis on informal Turkish texts. In *Proceedings of the 9th international conference on management of digital EcoSystems* (pp. 134-141).

Öztürk, Z. K., Cicek, Z. I., & Ergül, Z. (2017). Sentiment analysis: an application to anadolu university.

Karagoz, P., Kama, B., Ozturk, M., Toroslu, I. H., & Canturk, D. (2019). A framework for aspect based sentiment analysis on turkish informal texts. *Journal of Intelligent Information Systems*, *53*, 431-451.

Karamollaoğlu, H., Doğru, İ. A., Dörterler, M., Utku, A., & Yıldız, O. (2018, September). Sentiment analysis on Turkish social media shares through lexicon based approach. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)* (pp. 45-49). IEEE.

Karaöz, B., & Gürsoy, U. T. (2018). Adaptif öğrenme sözlüğü temelli duygu analiz algoritması önerisi. *Bilişim Teknolojileri Dergisi*, *11*(3), 245-253.

Karcioğlu, A. A., & Aydin, T. (2019, April). Sentiment analysis of Turkish and english twitter feeds using Word2Vec model. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Kaya, M., Fidan, G., & Toroslu, I. H. (2012, December). Sentiment analysis of Turkish political news. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 174-180). IEEE.

Kaynar, O., Görmez, Y., Yıldız, M., & Albayrak, A. (2016, September). Makine öğrenmesi yöntemleri ile Duygu Analizi. In *International Artificial Intelligence and Data Processing Symposium (IDAP'16)* (Vol. 17, No. 18, pp. 17-18).

Nalçakan, Y., Bayramoğlu, Ş. S., & Tuna, S. (2015). *Sosyal Medya Verileri Üzerinde Yapay Öğrenme ile Duygu Analizi Çalışması*. Technical Report.

Nizam, H., & Akın, S. S. (2014). Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı*, *1*(6).

Parlar, T., Özel, S. A., & Song, F. (2018). QER: a new feature selection method for sentiment analysis. *Human-centric Computing and Information Sciences*, *8*, 1-19.

Santur, Y. (2019, September). Sentiment analysis based on gated recurrent unit. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1-5). IEEE.

Shehu, H. A., & Tokat, S. (2020). A hybrid approach for the sentiment analysis of Turkish Twitter data. In *Artificial Intelligence and Applied Mathematics in Engineering Problems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2019)* (pp. 182-190). Springer International Publishing.

Shehu, H. A., Tokat, S., Sharif, M. H., & Uyaver, S. (2019, December). Sentiment analysis of Turkish Twitter data. In *AIP Conference Proceedings* (Vol. 2183, No. 1, p. 080004). AIP Publishing LLC.

Uslu, A., Tekin, S., & Aytekin, T. (2019, April). Sentiment analysis in Turkish film comments. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Velioğlu, R., Yıldız, T., & Yıldırım, S. (2018, September). Sentiment analysis using learning approaches over emojis for Turkish tweets. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)* (pp. 303-307). IEEE.

Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z. O. (2013). A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish. In *Computer and information sciences III: 27th international symposium on computer and information sciences* (pp. 437-445). Springer London.

Yurtalan, G., Koyuncu, M., & Turhan, Ç. (2019). A polarity calculation approach for lexicon-based Turkish sentiment analysis. *Turkish Journal of Electrical Engineering and Computer Sciences*, *27*(2), 1325-1339.

Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, *35*(1), 136-147.