MULTIMODAL ANALYSIS OF CUSTOMER REVIEWS BY TRANSFER LEARNING

MERVE GÖNENÇAYOĞLU

Thesis for the Master’s Program in Computer Engineering

Graduate School
Izmir University of Economics
Izmir
2023

MULTIMODAL ANALYSIS OF CUSTOMER REVIEWS BY TRANSFER LEARNING

MERVE GÖNENÇAYOĞLU

THESIS ADVISOR: ASSOC. PROF. DR. SENEM KUMOVA METİN

A Master's Thesis
Submitted to
the Graduate School of Izmir University of Economics
the Department of Computer Engineering

Izmir
2023

ETHICAL DECLARATION

I hereby declare that I am the sole author of this thesis and that I have conducted my work in accordance with academic rules and ethical behaviour at every stage from the planning of the thesis to its defence. I confirm that I have cited all ideas, information and findings that are not specific to my study, as required by the code of ethical behaviour, and that all statements not cited are my own.

Name, Surname:

Date:

Signature:

ABSTRACT

MULTIMODAL ANALYSIS OF CUSTOMER REVIEWS
BY TRANSFER LEARNING

Gönençayoğlu, Merve

Master's Program in Computer Engineering

Advisor: Assoc. Prof. Dr. Senem Kumova Metin

June, 2023

It is undoubtedly true that people choose online shopping platforms as technology improves each day. E-commerce companies receive a huge number of valuable reviews in text format. Processing this data with respect to sentiment analysis is important for ensuring customer satisfaction and product quality. Sentiment analysis can give precious insights about customer's needs and opinions. Through the years, companies found new ways to enrich the customer experience and added image attachment feature to reviews. In this thesis, we examine the success of different transfer learning models on classifying sentiments of customer reviews and propose a multimodal approach to robust the success of text analysis. Our multimodal approach uses SBERT sentence embeddings for text and CLIP vision transformers for image. The final multimodal approach has 93.03% accuracy and 93.08% F1 considering the highest values.

Keywords: Sentiment Analysis, Customer Reviews, Sentence Embedding, Transfer Learning, Multimodal.

ÖZET

TRANSFER ÖĞRENME ILE MÜŞTERI YORUMLARININ
ÇOKLU MODEL ANALIZI

Gönençayoğlu, Merve

Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Senem Kumova Metin

Haziran, 2023

Teknolojinin gün geçtikçe gelişmesiyle birlikte insanlar online alışveriş platformlarını tercih etmektedir. Bu platformlarda e-ticaret şirketleri, metin formatında çok sayıda yorum almaktadır. Bu yorumların duygu analizine göre işlenmesi, müşteri memnuniyeti ve ürün kalitesinin sağlanması açısından önemlidir. Duygu analizi, müşterinin ihtiyaçları ve görüşleri hakkında değerli içgörüler sağlayabilir. Yıllar geçtikçe şirketler, müşteri deneyimini zenginleştirmenin yeni yollarını bulmuşlardır ve sistemlerine resim ekleme özelliği eklemişlerdir. Bu tezde, metin ve resim formatını birlikte kullanarak, farklı transfer öğrenme modellerinin müşteri yorumlarındaki duygu sınıflandırmasındaki başarısı incelenmiştir. Metin formatındaki başarıyı arttırmak için bir çoklu model yaklaşımı önerilmiştir. Çoklu model yaklaşımında metin için SBERT cümle vektörleri, görüntü için CLIP görüntü dönüştürücüleri kullanılmıştır. Bu yaklaşımda, en yüksek değerler dikkate alındığında %93.03 doğruluk ve %93.08 F1 performans değerlerine ulaşılmıştır.

Anahtar Kelimeler: Duygu Analizi, Müşteri Yorumları, Cümle Vektörü, Transfer Öğrenme, Çoklu Model.

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to my supervisor Assoc. Prof. Dr. Senem KUMOVA METİN for giving me the opportunity to work with her, for her excellent guidance and patience. Her immense knowledge and valuable comments encouraged me and directed me in all the time during my study.

I would like to thank my mother Sıdıka GÖNENÇAYOĞLU and my father İsmail GÖNENÇAYOĞLU for their continuous support, endless motivation, and trust. I always knew that they believed in me and wanted the best for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In today's world, analyzing customer review data has great importance and decision-making potential for companies as more and more people use e-commerce platforms each day. This importance can be summarized in three main categories: from the aspect of customer and seller experience that has a direct impact on a company's strategies, the aspect of brand loyalty and having a distinguished position in the market.

First, a customer review can have a subject on a seller, a product or a platform experience. By analyzing these reviews successfully, companies can assess customer satisfaction and build their strategies according to direct feedback (Marketou, 2017). These strategies can affect customers, sellers, and products positively. As e-commerce mainly focuses on the success of customer satisfaction, seller satisfaction and product quality, companies can make to-the-point decisions to increase the quality of services. Hence, each successful strategic decision made by using customer review data can lead to an increase in revenue which is one of the most important metrics for an e-commerce company.

The next aspect is customer loyalty, which is the direct outcome of customer satisfaction. It is important for e-commerce companies to create brand loyalty. According to Gartner (2022), 85% of buyers trust online reviews as much as personal recommendations. To achieve this, companies need to understand customers through their feedback and reviews and make them feel like they have a voice while using platforms (Marketou, 2017). In addition, encouraging customers to make more reviews has its own advantage as the more interaction an online platform receives, the more customers it can attract. Hence, the advertising cost for the company is reduced with the help of customer reviews.

Finally, as the usage of e-commerce platforms increases every day, new companies emerge in this sector. By using customer review data, companies can create an experience from previous sales and can make a difference in the market. Some of these differences can be implemented with technology investments such as gathering customer reviews not only with text but also images which this thesis will mainly focus on both data types.

There is no doubt that analyzing sentiments and classifying customer reviews has a huge impact on every company's success. Researchers study this problem to find the most effective solution in the means of both success and performance. As there is much research available in the literature, previous studies can be summarized with three main qualifications: working on text-based reviews only, language of review and using word embedding models while working with sentences.

Previous studies on sentiment analysis in customer review data mainly focused on text-based data only, such as product reviews and tweets on various topics. The main reason behind this is, adding an image to reviews is a considerably new development for online platforms when compared to the development of text-based reviews. In addition, most of the experiments had been done in the English language. There is less research that focuses on the Turkish language and because of the structural differences between English and Turkish languages, successful results are still being investigated.

Using the text vectorization method that provides the best representatives for texts is the key to reaching success in sentiment classification problems. Main text vectorization models that had been used in previous studies operate on word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). The main disadvantage of word embeddings is, when a sentence is given as an input, the model calculates a vector for each word without understanding its context in the sentence. Therefore, in current studies, the embeddings which represent sentences are proposed. These embeddings provide contextual integrity for vectorizing sentences. There are several sentence embedding techniques such as Doc2Vec (Le and Mikolov, 2014), SkipThought (Kiros et al., 2015), InferSent (Conneau et al., 2017), Universal Sentence Encoder (USE) (Cer et al., 2018) and SentenceBERT (SBERT) (Reimers and Gurevych, 2019).

In this thesis, SBERT variants are employed in experiments. SBERT uses BERT-based models which use transformers (Vaswani et al., 2017) architecture. This architecture has several advantages that can be summarized as processing relationships between sequential elements which are far from each other, accuracy and training and processing more data in less time (Srivastava, 2022).

All in all, there are three research questions that are studied in this thesis to analyze the product reviews utilizing transformer-based models in a multimodal approach. The research questions (RQ) can be summarized as follows.

RQ1. Which transformer-based sentence embedding provides a successful classification of product reviews?

In this thesis, Turkish product reviews are represented by several sentence embeddings that are built by transformer-based models. The performances of alternative embeddings are compared by five different machine learning algorithms. The aim is to present the combination of the most successful transformer-based embedding model and the machine learning algorithm that provides the highest performance in classification.

RQ2. Is it possible to improve the classification performance by multimodal embeddings that involve image data together with text data?

In this study, the experiments had been done on not only text but also image data that is attached to the product review. Concatenating image embeddings to text embeddings, multimodal representations of product reviews are formed. The multimodal embeddings are given as inputs to classification methods to measure the change in classification performance.

RQ3. Can the image data be used individually in a pre-processing step in the product review classification process?

We followed up two approaches to utilize image data in pre-processing steps of classification processes. These are object detection and image-product name similarity measurements. In object detection, simply, the given image is analyzed to detect if it contains a valid object assuming that if there exists at least a valid object in the image, it potentially belongs to the regarding product. A new dataset had been created after the object (detection) filtering and sent to sentence transformers.

In image-product name similarity measurements, the embeddings of review images and product name texts are generated into the same vector space. (Reimers and Gurevych, 2019). Then, cosine similarity is calculated between two vectors. Finally, this comparison is added as a new layer to the proposed classification pipeline.

The thesis is structured as follows; Chapter 2 is where the literature research has been summarized. Chapter 3 contains information about the methodology and experimental setup of our approaches. Chapter 4 is where experimental results have been shared. Chapter 5 is where the thesis has concluded.

CHAPTER 2: RELATED WORK

In today's world, sentiment analysis has been used to classify product reviews for their positivity or negativity. As technology is advancing every day, there are different methodologies researched to understand product reviews better. These methodologies mainly experimented with English product reviews and proved to be successful. However, the number of experiments on Turkish datasets is limited. Examining the previous works, it is also observed that most of the research regarding sentiment analysis had been done with word embeddings. And the works that employ sentence embeddings are relatively new. In most of the previous studies, sentiment analysis had been done mostly using only text dataset. Multimodal experiments are recently developed with the advances in technology.

This section will be divided into three main parts. First, studies on the English datasets will be explained. Second, multimodal studies will be represented. Finally, studies on Turkish Language will be explained.

Table 1 provides information on example studies on product review data that are comparable to ours and are accepted to be related to our research questions. In Table 1, the scores with an asterisk (*) represent the highest performance scores for the experiments done in the research. The details of research Id (RId) 1, 2, 3, 4 and 5 are given in section 2.1 Studies on English Customer Reviews. The details of RId 6, 7 and 8 are given in section 2.2 Multimodal Studies. The details of RId 9 and 10 are given in section 2.3 Studies on Turkish Datasets.

Table 1. Studies on Product Review

RId	Research Name	Method	Data Type	Accuracy	F1	Dataset	Dataset Size	Language
1	Kaynar et al. (2016)	TF-IDF	Text	*75% with ANN and SVM, 67% with Centroid, 68% with Naive	*76% with ANN and SVM, 66% with Centroid, 72% with Naive	IMDB Movie Review	2000 movie review (1000 negative, 1000 positive)	English
2	Singla, Randhawa and Jain (2017)	Sentiment Scores by NRC Sentiment Dictionary	Text	*81.77% with SVM, 66.95% with NB, 74.75% with DT		Amazon Product Reviews (mobile phones)	3,000 product reviews (1500 positive, 1500 negative)	English
3	Saha (2023)	BERT, Word2Vec	Text			Amazon product reviews (consumer electronics category)	1,177 reviews with 5 ratings.	English
4	Reimers and Gurevych (2019)	SBERT	Text	MR: *84.88% with SBERT-NLI-large, 83.64% with SBERT-NLI-base, 80.09% with Universal Sentence Encoder, 81.57% with InferSent – GloVe SST: *90.66% with SBERT-NLI-large, 88.96% with SBERT-NLI-base, 86.38% with Universal Sentence Encoder, 84.18% with InferSent - GloVe		MR - movie reviews snippets (Pang and Lee, 2005), SST - Stanford Sentiment Treebank (Socher et al.,2013) (Rottentomatoes movie reviews)	MR: 5006 reviews, (3 class and 4 class ratings), SST: 10662 sentences (5331 positive, 5331 negative)	English

Table 1 (Continued). Studies on Product Review

RId	Research Name	Method	Data Type	Accuracy	F1	Dataset	Dataset Size	Language
5	Mishev et al.(2020)	InferSent	Text	*85.8% (max reported performance)	*85.4% (max reported performance)	The Financial News Statements and Headlines, The Financial Phrase-Bank	2,000 sentences	English
6	Yu, and Jiang (2019)	TomBERT	Text, Image	*77.15% with TomBERT - FIRST, 76.57% with TomBERT - CLS, 74.15% with BERT	*71.75% with TomBERT-FIRST, 71.17% with TomBERT -CLS, 68.86% with BERT	Text only: SemEval-2014 Task 4 Amazon Customer Reviews (Pontiki et al., 2014) (Laptop and Restaurant category), TWITTER-14 (Dong et al., 2014) Multimodal: TWITTER-15 (Zhang et al., 2018), TWITTER-17 (Lu et al., 2018)	TWITTER-15 - 5338 tweets (1548 Positive, 630 Negative, 3160 Neutral), TWITTER-17 - 5972 tweets (2516 Positive, 728 Negative, 2728 Neutral)	English
7	Wöllmer et al. (2013)	Bag-of-Words, Bag-of-N-Gram	Video , Text	*73.2% (max reported performance)	*73.2 % (max reported performance)	Youtube and ExpoTV (Video), Metacritic Movie Review (Text)	370 review video (Youtube: 228 positive, 23 neutral, 57 negative, ExpoTV: 62 negative) ,102,622 text review	English
8	Bhat et al. (2022)	BERT for text, ResNet-50 for video	Text, Video	71.5% with proposed model, *75.33% with TFN	71.6% with proposed model, *76.2% with TFN	CMU-MOSI Multimodal Sentimental Dataset	2,199 reviews	English

Table 1 (Continued). Studies on Product Review

RId	Research Name	Method	Data Type	Accuracy	F1	Dataset	Dataset Size	Language
9	Hayran, and Sert (2017)	Word2Vec	Text	*80.05% with variance + average + sum (dvot), 78.8% with average + sum (dot), 79.72% with average and variance (dov), 79.13% with sum + variance (dtv), 64.02% with variance (dv), 78.34% with average (do), 78.31% with sum (dt), 65.17% with CBOW, 66.13% with SKIP-GRAM		Turkish Twitter	32,000 tweet (16,000 negative, 16,000 positive)	Turkish
10	Rumelli et al. (2020)	Lexicon based approach	Text	*73.8% with kNN, 73.2% with NB, 73.3% with RF, 73.4% with SVM	*74.7% with NB, 73.7% with kNN, 73.6% with RF, 72.8% with SVM	Hepsiburada and SentiTurkNet	26,000 product reviews (13,000 positive, 13,000 negative)	Turkish
11	Guven (2021)	mBER Turkish ELECTRA and Turkish ALBERT	Text	*89.95% with NB, 89.91% with LR, 86.26% with RO, 88.42% with BERT-M, *92.54% with ELECTRA-Tr, 91.29% with ALBERT-Tr		Hepsiburada Product Reviews	27,352 product reviews (13,676 positive, 13,676 negative)	Turkish

2.1. Studies on English Customer Reviews

Research until recent years on sentiment analysis problems has been done mostly in the English language. The studies continue by using different approaches to represent reviews better such as word embeddings and sentence embeddings. Some of these approaches are summarized in the following paragraphs.

Kaynar et al. (2016) studied on IMDB movie review dataset. They applied the TF-IDF method and achieved the highest accuracy of 89.73% with artificial neural networks in training and an accuracy of 75% with the testing dataset.

Singla, Randhawa and Jain (2017) experimented on Amazon product reviews, specifically in the mobile phone category. For classification, they calculated sentiment scores for each review. Sentiment scores established by NRC (National Research Council) sentiment dictionary. The polarity for sentiment scores is calculated by extracting the negative score from the positive score. Then, they sent these scores to classifiers to predict the sentiments of product reviews. They used Naive Bayesian, Support Vector Machine and Decision Tree classifiers. They calculated accuracies of 66.95%, 81.77% and 74.75% respectively. Our multimodal approach outperformed these performance results using sentence embeddings and multimodal representation.

Saha (2023) experimented on the Amazon product reviews dataset to benchmark different word embedding methods and their impact on clustering algorithms. The research examined three different clustering algorithms: partitioning-based (KMeans), single linkage agglomerative hierarchical and density-based scan (HDBSCAN, DBSCAN). To vectorize the product reviews, BERT and Word2Vec models are used. Although this research aims to find the best-performing class of clustering, in three different clustering classes, BERT with [CLS] token embedding outperformed Word2Vec and Bert with average token embeddings. In our thesis, we used SBERT (Sentence-BERT) with mean pooling strategy as it is reported to outperform other token embedding approaches (Reimers and Gurevych, 2019).

The embedding method that is used in this thesis is introduced in "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks" (Reimers and Gurevych, 2019). In BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), there are some issues faced in solving problems like sentence similarity. These approaches require both sentences to be fed into the network. This resulted in a computational overload. For example, with BERT, in a collection of n = 10 000 sentences, $\frac{n(n-1)}{2}$= 49,995,000 computations must be done. It takes about 65 hours of computational workload. With SBERT, it can be done in five seconds. BERT method takes input of individual sentences and outputs sentence embeddings. Then, the average of BERT outputs (BERT embeddings) or output of first token ([CLS] token) has been calculated. On the other hand, SBERT adds a pooling operation to the output of BERT to produce sentence embeddings. Three different pooling strategies which are using the output of CLS token, computing mean of output vectors and computing max-over-time of the output vectors are experimented (Reimers and Gurevych, 2019). The best-performing pooling strategy is reported as mean pooling with 80.78% on NLI and 87.44% on STSb performances. The proposed SBERT method is a pre-trained model. It had been trained on SNLI (Bowman et al., 2015) and Multi-Genre NLI (Williams et al., 2018) datasets. SBERT had been experimented on different datasets. For the sentiment analysis problem, Reimers and Gurevych (2019) experimented SBERT on 1- sentiment prediction for movie reviews snippets (Pang and Lee, 2005), 2- sentiment prediction of customer product reviews (Hu and Liu, 2005) and 3- Stanford Sentiment Treebank with binary labels (Socher et al., 2013). The prediction accuracy measured by SentEval is 83.64%, 80.43% and 88.96% in order. On the other hand, BERT performances for the same datasets are 78.66%, 86.25% and 84.40% in order. The experimental results in Reimers and Gurevych's work (2019) showed that SBERT is the highest-performing sentence embedding model compared to the previous approaches. Also, SBERT is more efficient in computing power compared to BERT and RoBERTa (Liu et al., 2019).

Mishev et al. (2020) studied on English finance dataset and applied word and sentence embedding methods to conduct sentiment analysis. In sentence encoders, the highest-performing method is reported to be InferSent with 85.8% accuracy and 85.4% F1.

2.2. Multimodal Studies

To robust the success in sentiment analysis, multimodal approaches were introduced in previous studies. These studies employed text data with image, video, or audio datasets. In the following paragraphs, some examples of multimodal studies are summarized.

Wöllmer et al. (2013) studied multimodal architecture consisting of movie reviews in text and video format. For the video set, they gathered movie critic videos from Youtube and ExpoTV. They used Metacritic for the written text dataset. Multimodal feature extraction is performed in two parts. In the first part, they detected faces in every frame with the commercial software Okao Vision. Then, they applied Generalized Adaptive View-based Appearance Model and created video features. In the second part, they used Bag-of-Words and Bag-of-N-Gram on the text data. For the classification, they used linear Support Vector Machines. The highest accuracy and F1 scores are reported as 73% and 73% respectively in text-only experiments. In multimodal experiments, the highest accuracy is 73.2% and the highest F1 is 73.2%. It is examined that the performances are higher in multimodal representation, but it cannot be stated that multimodality provides a significant improvement in performance.

Yu and Jiang (2019), worked on Target-oriented Sentiment Classification and proposed a multimodal BERT architecture. Their motivation was identifying sentiment polarities in each sentence and adding image content to enhance the robustness. First, they divided each sentence into two sub-sentences: individual opinion target words and the remaining context words. Second, they applied BERT to text. Then, they designed a system to learn the alignment between opinions and images. They named this approach TomBERT (Target Oriented Multimodal BERT). The experiments were performed in two parts. First, fine-tuned BERT had been utilized on three benchmark datasets and results outperformed previous experiments. Then, they applied the proposed multimodal approach to the Twitter dataset and outperformed other multimodal methods. The highest performances were posted as 77.15% accuracy and 71.75% F1 in publicly available Twitter-15 (Zhang et al., 2018) dataset. Although this approach outperformed previous sentiment classification and multimodal approaches, it relies on BERT word embeddings.

Bhat et al. (2022) suggested a multimodal approach consisting of text and video datasets for sentiment classification. They used BERT for text embedding and ResNet-50 for video encoding. Then, these representations were joined together with the cartesian product. This approach is like the Tensor Fusion Network (Zadeh et al.,2017). CMU-MOSI Multimodal Sentimental Dataset which is the second largest dataset for multimodal sentiment analysis had been used in this study. However, the results were lower than the experiments done in Tensor Fusion Network (TFN). The highest TFN performance score was 75.33% in accuracy and 76.2% in F1. But the approach proposed by Bhat et al. (2022) resulted in 71.5% accuracy and 71.6% F1. There are three main reasons for this decrease in performance: quality of dataset, hypermeter tuning and loss of inference due to compressors and classifiers (Bhat et al., 2022). Our thesis not only had higher performance scores than previous TFN experiments but also with the image-product name similarity layer, the multimodal approach had higher performance than text embeddings.

2.3. Studies on Turkish Datasets

Experiments with Turkish datasets are limited, and this area needs to be studied more. In the following paragraphs, examples of studies that employ Turkish datasets are briefly explained.

Hayran and Sert (2017) experimented on Turkish Twitter dataset. Their approach had four stages. In the first stage, they cleaned the dataset from irrelevant data. In the second stage, they created word embeddings with Word2Vec (Mikolov et al., 2013). In the third stage, they created feature vectors. Finally, they sent the embeddings to SVM (support vector machine) and classified sentences according to their sentiments. Also, they experimented with different fusion techniques. The highest performer fusion technique had an accuracy of 80.05%. Although it is a high performance compared to previous English experiments, this experiment uses word embeddings, too. However, our experiments done with sentence embeddings, used multiple supervised machine learning algorithms, and resulted with higher accuracy.

Rumelli et al. (2020) proposed a lexicon-based approach using Hepsiburada and SentiTurkNet datasets. They preprocessed the dataset by clearing or correcting words. Then, they calculated the polarity for each word in a sentence. They used four different machine learning algorithms to classify the reviews. They achieved the highest accuracy with k-Nearest Neighbour with an accuracy of 73.8% and the highest F1 with Naive Bayes with 74.7%.

Guven (2021) experimented on Turkish product review dataset collected from Hepsiburada. The effect of multilingual BERT, Turkish ELECTRA and Turkish ALBERT had been investigated. The results had been compared with the results of Random Forest, Naive Bayes and Logistic Regression algorithms. The highest accuracy had been obtained with Naive Bayes algorithm with an accuracy score of 89.95% and Turkish Electra with an accuracy score of 92.54%.

As we mentioned before, there is a lot of research on the English language in sentiment classification in the literature. However, experiments on the Turkish language are limited. In addition, multimodal experiments are more limited. However, with the improvements in technology, a multimodal approach with sentence embeddings must be taken into consideration to robust the success of sentiment classification. This thesis focuses on this area.

CHAPTER 3: EXPERIMENTAL SETUP

In this thesis, a transformer-based approach had been followed to create sentence embeddings for each product review to capture the meaning of the whole sentences. First, the experiments had been done on text-only data to research if transformer-based sentence embeddings provide successful classification results. Second, multimodal, concatenated text and image, embeddings had been built to improve the classification performance. Third, two pre-processing approaches had been experimented on to determine reliable image data: object detection and image-product name similarity. As a result, a hybrid method with an image-product name similarity approach that outperforms the alternatives is suggested. The method uses text-only data for some product reviews and concatenated text and image data for the rest of the product reviews.

For all experiments, each product review had been labelled with one of the two sentiment categories: positive and negative. In addition, performance metrics had been analyzed with supervised machine learning methods which will be detailed in the following sections.

3.1. Dataset

There is a total of 155,849,352 product reviews in the database from which the datasets in this thesis are sampled. There are six main categories such as electronics, house and furniture, cosmetics etc. regarding the corresponding product. To produce a relevant sentiment output and because this category had more samples than other categories, the *Textile and Accessories* category had been chosen to sample reviews randomly. Contents of the reviews from the chosen *Textile and Accessories* category range from clothing, shoes, and bags to accessories like jewelry and watches. Among the 155,849,352 product reviews, there are 14,630,988 reviews which also have image data. The *Textile and Accessory* category owns the highest number of reviews with image data. Overall product review text data count in the database with categories can be seen in Table 2.

Table 2. Number of Product Reviews by Categories

Category	**Number of Product Reviews**	
	Text Data	**Image Data**
Textile and Accessory	77,996,742	5,929,204
House and Furniture	27,010,519	3,662,321
Consumer Goods	20,145,810	1,997,379
Cosmetics	13,312,460	1,513,704
Electronics	11,501,249	1,210,545
Youth and Sport	5,882,572	317,835
TOTAL	155,849,352	14,630,988

There are two main datasets which are used in this thesis. The first dataset (DS1) has 15,136 product reviews and it has text-only data. The second dataset (DS2) has 1,866 product reviews and it has text together with image data (given in Table 3.).

Table 3. Datasets

Number of Dataset	**Number of Product Reviews**	**Data Type**
DS1	15,136	Text
DS2	1,866	Text and Image

Labeling for sentiment analysis was conducted with a team of three bachelor's degree graduated annotator/judge that is native in Turkish language and belong to the 40-60 years old group. Two judges read each product review text and categorized each review with positive or negative labels. When two judges could not agree on a common decision, the third judge presented a decision. The final decision had been made with majority voting (Simply, the number of positive and negative votes are counted for each sample, and the label that owns the highest number of votes is assigned to the regarding sample). The resulting sentiment distribution in both datasets is balanced.

In the review datasets, not all product reviews have an image for the given text or vice versa as adding an image to a review is relevantly new technology. So, the first dataset (DS1) is a text-only dataset. However, to see the effect of image enrichment, it was important to choose samples that have both text and image pairs. So, the second dataset (DS2) is built as a multimodal set.

It should be noted that to obey the Personal Data Protection Authority and confidentiality agreement, all data that had been written as an example through the thesis has been modified. Because product reviews which are written as free text, contain personal opinions and may contain personal information about a customer, seller or company.

3.1.1. Text-only Dataset (DS1)

Text-only data set is a collection of 15,136 Turkish reviews labelled as positive or negative (given in Table 4.). The number of product reviews in this dataset is like the number of product reviews in Reimers and Gurevych (2019) study where 10,662 movie reviews are employed to reach 90.66% of accuracy.

Table 4. Number of Product Reviews by Sentiment Labels from DS1

Sentiment	**Number of Product Reviews**
Positive	7,870
Negative	7,266
TOTAL	15,136

The input file for the text-only dataset contains two columns which are comment and sentiment. The comment column stores Turkish product review text data. The sentiment column stores positive or negative labels assigned by the judges. The examples of samples in DS1 are given in Table 5. In DS1, the text length per review varies between two words to 170 words.

Table 5. Examples from DS1

Comment	Sentiment
Çoraplar hiç yumuşak değil, lastikleri ilk giymede gevşedi ve tüylendi. 100 TL çöpe gitti. (The socks are not soft at all, the elastics got loose and feathered in the first wearing. 100 TL wasted.)	Negative
Ürün güzel. 2. siparişim oldu bu markadan. ilk aldığım tshirt kutulu bir şekilde özenle gönderilmişti bu siparişimde kutusuz gönderildi ve üzerinde biraz toz vardı. ürün güzel problem yok. (The product is beautiful. This is my 2nd order from this brand. The first t-shirt I bought was carefully sent with a box, this order was sent without a box and there was a little dust on it. The product is good, no problem.)	Positive

3.1.2. Multimodal Dataset (DS2)

DS2 is a collection of 1,866 Turkish text and image pairs labeled as positive or negative as given in Table 6.

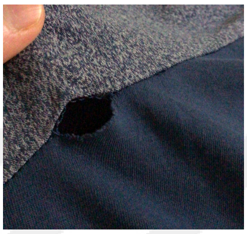
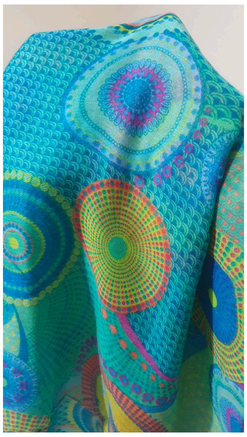
Table 6. Number of Product Reviews by Sentiment Labels from DS2

Sentiment	Number of Product Reviews
Positive	942
Negative	924
TOTAL	1,866

The input file for DS2 contains four columns which are *Comment, url, Sentiment* and *Product Name*. The *Comment* column stores Turkish product review text data. The *url* column stores an image of the product that the customer uploaded to the system while writing the comment. It generally shows the defect of the product if the review is negative. On the other hand, it usually shows the customer of the product if the review is positive. *Sentiment* column stores positive or negative labels assigned by the judges.

Product Name column stores the name of the product which we included in our experiments and will be discussed further. The resulting data after the labelling process can be seen with some examples in Table 7. The length of the text per review varies between two words to 79 words in this dataset DS2.

Table 7. Examples from DS2

Comment	Url (Image)	Sentiment	Product Name
Bu fiyata daha iyi ceketler alınabilir. Çok pişmanım. 1. si sıcak tutmuyor. 2.si cebi delikti ve dikişleri hep attı. Düğmeleri düşecek gibi duruyor. Terziye verilip sağlamlaştırılması lazım. (You can get better jackets for this price. I'm very regretful. Firstly, it does not keep warm. Secondly, it had a hole in the pocket and the seams took off. The buttons seem to fall off. It should be given to a tailor and fixed.)		Negative	Yuvarlak Düğmeli Kumaş Ceket (Round Buttons Tweed Jacket)
Pareo görseldekiyle aynı. Ürün çok kaliteli. Ben bir çok farklı desenini aldım. Yanında hediye olarak toka da göndermişler. Çok teşekkür ederim herkese tavsiye ediyorum. (Pareo is the same as in the picture. The product is very high quality. I bought many different patterns. They also sent a hairpin as a gift. Thank you very much, I recommend it to everyone.)		Positive	Mavi Renkli Otantik Desenli Pareo (Blue Colored Otantic Patterned Pareo)

3.2. Transfer Learning Methods

Transfer learning is using pre-trained models on a new problem. It is especially useful on computing complex problems such as computer vision and natural language processing because a pre-trained model can be reused on a loosely related different problem (Sharma, 2021).

As explained in Cohere (2022), pre-trained models have huge advantages compared to training a model from scratch. First, the same or better performance can be achieved faster with pre-trained models. Secondly, training a model from scratch requires the processing of a huge amount of data which makes the process more time and resource-consuming. However, there is an important downside of transfer learning. In transfer learning, initial training, and the problem which the pre-trained model will use must be similar (Joshi, 2020). For example, if a model is trained for detecting animals in an image, it may not perform well in the problem of detecting vehicles in an image. It is called negative transfer.

In this thesis, we used an infrastructure of sentence-embeddings with Siamese BERT-Networks, in short SBERT (Sentence-Bert) model, to vectorize input data which is one of the state-of-art models for creating meaningful, semantically similar sentence embeddings that are located close in vector space (Reimers and Gurevych, 2019).

SBERT models use Sentence Transformers Python framework to create sentence and image embeddings. This framework supports more than 100 languages and offers a very large collection of pre–trained models.

In this thesis, we used pre-trained SBERT models that have Turkish language support specifically, to understand product review sentences correctly and to create meaningful sentence embeddings. For the image vectorization in multimodal experiments, we used OpenAI CLIP (Contrastive Language-Image Pre-Training) Model (Radford et al., 2021) by using Sentence Transformers as a wrapper. The list of pre-trained models that are employed in our experiments is presented in Table 8.
In following subsections, SBERT and its variants (SBERT-BERT, SBERT-DBERT, SBERT-XLMR, SBERT-XLMR, SBERT-CLIP), and vision transformers employed in our experiments will be briefly presented.

Table 8. Pre-trained Models

Model Name	**Transformer Model**	**Model Url**	**Data Type**
paraphrase-multilingual-MiniLM-L12-v2	SBERT-BERT (Sentence Transformer - BERT Based Model)	https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2	Text
distiluse-base-multilingual-cased-v1	SBERT-DBERT (Sentence Transformer - distilBERT Based Model)	https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1	Text
paraphrase-multilingual-mpnet-base-v2	SBERT-XLMR (Sentence Transformer - XLMRoberta Based Model)	https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2	Text
clip-ViT-B-32-multilingual-v1	SBERT-CLIP (Sentence Transformer – CLIP Model)	https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1	Text
clip-ViT-B-32	CLIP-ViT (Vision Transformer (ViT))	https://huggingface.co/sentence-transformers/clip-ViT-B-32	Image
yolos-tiny	YOLOS-ViT (Vision Transformer (ViT))	https://huggingface.co/hustvl/yolos-tiny	Image

3.2.1. SBERT

SBERT is a modification of the BERT network using Siamese networks that can derive semantically meaningful sentence embeddings. This enables BERT to be used for certain new tasks such as large-scale semantic similarity comparison, clustering, and information retrieval via semantic search (Reimers and Gurevych, 2019).

SBERT model had been used in several different natural language processing tasks. Some examples can be listed as follows.

- Guo et al. (2023) used SBERT in question-answering problem. They proposed a new approach to downsize models to support devices with different memory configurations.
- Bhandare and Haribhakta (2022) create embeddings for the questions by SBERT

in a database to find questions that require a similar thinking process.

- Sasaki and Masada (2022) used SBERT to produce document embeddings and decide whether essays are good or bad with essay scoring.
- Ajallouda et al. (2022) employed SBERT for representing noun phrases.
- Madhusudhan, Mahurkar and Nagarajan (2020) applied SBERT to fake news detection problem and experimented on multimodal datasets using the ResNet-18 model.

As Reimers and Gurevych (2019) stated, one of the reported advantages of SBERT is its high performance in text similarity problems. Performance improvement can be explained with the sentence similarity problem as follows. In sentence similarity problems, to obtain an accurate similarity score, we need a sentence embedding that can represent the meaning hidden in the sentence. Before sentence transformers, BERT uses a cross-encoder structure which requires sending two sentences to the BERT network as input and adding a classification head on top for measuring a similarity score (given in Figure 1). However, this solution was not scalable because if it is fed by a 100K sentence set, it is required to perform 100K computations which means comparing each sentence with others in a 100K dataset. This is why, the ideal method would be computing each sentence vector before and using it when it is required. For scalability, SBERT produces sentence embeddings beforehand, so it is not needed to operate for each sentence-pair comparison. As a result, for sentence similarity problems, BERT could complete its operations in 65 hours for 10K sentences. On the other hand, SBERT could create the same number of embeddings in nearly five seconds and could calculate cosine similarity in 0.01 seconds (Reimers and Gurevych, 2019).

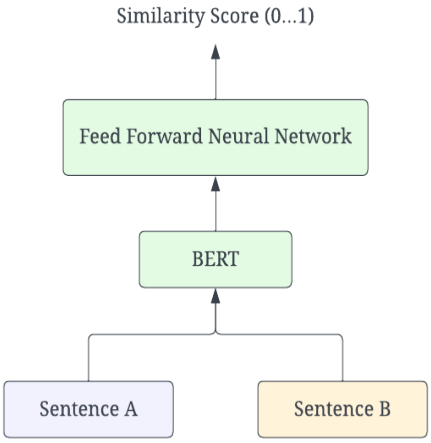


Figure 1. BERT Architecture for Sentence Similarity Problem

SBERT is different from BERT in many ways. For example, it does not have a classification head, and it processes one sentence at a time. In addition, different from BERT, SBERT uses mean pooling on the final output layer and finally, calculates sentence embeddings. As there are different pooling algorithms, SBERT employs mean pooling as it conceives the best results. Mean pooling is the layer for generalizing features by averaging groups of features of the BERT. After the pooling, there are two embeddings for the sentences. SBERT concatenates them and sends them to a SoftMax classifier. Then, training ends with the addition of SoftMax-loss function.

SBERT is trained and fine-tuned on sentence pairs using what is called Siamese architecture. However, it consists of a single BERT model. But because training is conducted by sentence A followed by sentence B as pairs, we can think that it has two identical BERT architectures with the same network weights and they run in parallel (Briggs, n.d.). This is the reason why we demonstrate it as two BERT models although the architecture has one BERT model (Figure 2).

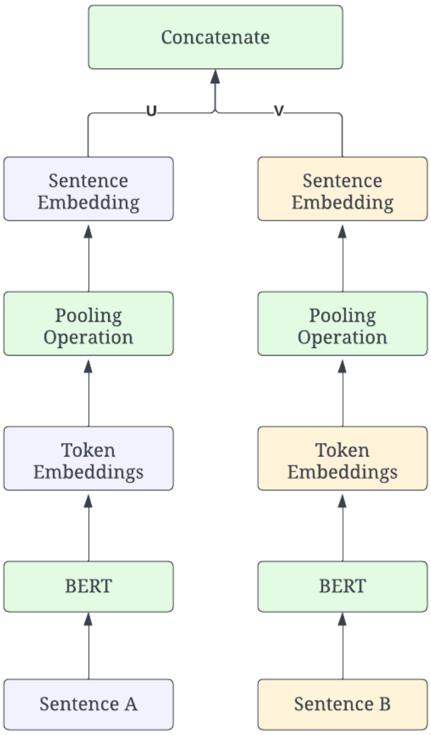


Figure 2. SBERT Architecture to demonstrate Siamese Networks

3.2.2. SBERT with BERT Based Model (SBERT-BERT)

We used paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2020) model as a SBERT model that uses BERT as a transformers model. SBERT-BERT is a pretrained multilingual model that supports more than 50 languages including the Turkish language. In training, it uses paraphrase-MiniLM-L12-v2 as teacher model and Microsoft/Multilingual-MiniLM-L12-H384 (Wang et al.,2020) as student model. It maps sentences to a 384-dimensional dense vector space.

The teacher model paraphrase-MiniLM-L12-v2 uses Microsoft/MiniLM-L12-H384-uncased model as a base model. The teacher model is trained on multiple datasets that can be summarized as AllNLI (concatenation of the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets), sentence-compression, SimpleWiki, altlex, msmarco-triplets, quora_duplicates, coco_captions, flickr30k_captions, yahoo_answers_title_question, S2ORC_citation_pairs (Semantic Scholar Open Research Corpus), stackexchange_duplicate_questions and wiki-atomic-edits.

The base model uses BERT transformers. There are two BERT models that have been proposed by Google AI (BERT Base and BERT Large) (Devlin et al., 2019). In paraphrase-multilingual-MiniLM-L12-v2 model, BERT Base model had been used. BERT Base has twelve layers with twelve attention heads and 110 million parameters. It consists of twelve encoder transformer blocks that had been stacked.

BERT model architecture consists of four main concepts. As Shreya (2022) explains in detail, these are token embeddings, position embeddings, self-attention layer and feed-forward neural network. The BERT model uses bidirectional training which considers both previous and next tokens simultaneously to capture the context of the sentence. It has a multi-layered architecture.

The first member of the architecture is calculating token embeddings. The BERT encoder architecture expects a sequence of tokens to the first of the encoder as an input. [CLS] tokens are special tokens that BERT uses at the beginning of the first sentence. [SEP] tokens are also special tokens for BERT that are placed at the end of each sentence. These representations create token embeddings (Shreya, 2022).

The second member of the BERT architecture is position embeddings which encode the position of each word. Token embeddings and position embeddings vectors are the same size so they can be summed to have one embedding. The final representation becomes an input for the self-attention mechanism which calculates the relation of the words in the sentence (Uçar, 2020). For example, in the sentence of "Yüzüklerin Efendisi okuduğum en güzel kitaptı, onun sayesinde fantastik kitaplara olan ilgim arttı." ("Lord of the Rings was the best book I've ever read, thanks to it my interest in fantasy books increased") relation between "Yüzüklerin Efendisi" ("Lord of the Rings"), "kitaptı" ("book") and "onun" ("it") words have a similar meaning and this has been calculated in self-attention layer. These results are sent to the encoder's final layer, feed-forward neural network and it is passed to the next encoder. As a result, each position has a corresponding vector which is the word embedding.

3.2.3. SBERT with XLMRoberta Based Model (SBERT-XLMR)

We used the paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2020) model as a SBERT model that uses XLMRoberta as a transformers model. It is a pre-trained multilingual model that supports more than 50 languages including the Turkish language. In training, it uses paraphrase-mpnet-base-v2 as teacher model and xlm-roberta-base as student model. It maps sentences to a 768-dimensional dense vector space.

The teacher model paraphrase-mpnet-base-v2 uses Microsoft/mpnet-base model as base model. The teacher model is trained on multiple datasets that can be summarized as AllNLI (concatenation of the SNLI and MultiNLI datasets), sentence-compression, SimpleWiki, altlex, msmarco-triplets, quora_duplicates, coco_captions,flickr30k_captions, yahoo_answers_title_question, S2ORC_citation_pairs (Semantic Scholar Open Research Corpus), stackexchange_duplicate_questions and wiki-atomic-edits.

The student model xlm-roberta-base is the multilingual version of RoBERTa. As stated by (Khan, 2019) it takes BERT architecture one step forward with removing the Next Sentence Prediction (NSP) and proposing dynamic masking which masked token changes while training epochs.

Lastly, the biggest change comes with the data that is used in pre-training. XLMRoberta is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages while BERT was trained on 16GB of Books Corpus and English Wikipedia (Conneau et al., 2019).

3.2.4. SBERT with DistilBERT Model (SBERT-DBERT)

We used distiluse-base-multilingual-cased-v1 (Reimers and Gurevych, 2020) model as an SBERT model that uses distilBERT as a transformers model. It is a pre-trained multilingual model that supports 15 languages including the Turkish language. In training, it uses Multilingual Universal Sentence Encoder (mUSE) as the teacher model and distilbert-base-multilingual as the student model. It maps sentences to a 512-dimensional dense vector space.

DistilBERT is like the BERT, but it is a smaller, distilled version. It retained 97% of the BERT's performance but used only half of the parameters (Sanh et al., 2019). It has half of the BERT's layers and does not have token-type embeddings. It uses a distillation technique on BERT which means approximating a larger network by a smaller one. This had been achieved with Kullback Leiber divergence.

3.2.5. SBERT with Clip Model (SBERT-CLIP)

We used the clip-ViT-B-32-multilingual-v1 (Reimers and Gurevych, 2019) model as thee SBERT model which is the multilingual text encoder for the OpenAI CLIP model. It is a pre-trained model that supports more than 50 languages including the Turkish language. It can encode text and can match the image vectors from the clip-ViT-B-32 model which we also used to vectorize images in multimodal experiments.

This model had been prepared with multilingual knowledge distillation. OpenAI's original CLIP model had been used as the teacher model and distilbert-base-multilingual-cased model had been used as the student model. The multilingual student model learns to align to the teacher model's vector space in more than 50 languages. Hence, a multilingual text model had been trained. For the text part, it has the same architecture model as distilBERT representation in *3.2.4. SBERT with distilBERT Model.*

3.3. Vision Transformers with CLIP (VIT-CLIP)

In this study, to measure the performance of image-product name similarity, we used sentence-transformers/clip-ViT-B-32-multilingual-v1 as the text embedding model that vectorizes product names and clip-ViT-B-32 as image embedding model that vectorizes product review images.

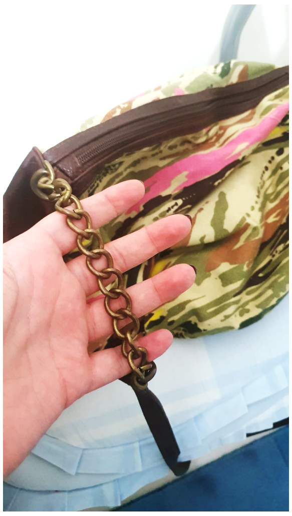
After the vectorization process, the cosine similarity of image and text embeddings is measured. If two vectors share a similar direction, the cosine similarity value would be high. This calculation had been done for each product name per image. Then, we calculated top one, top five and top ten highest cosine similarity scores in order to have a cleaner image dataset.

3.4. Vision Transformers with YOLOS (VIT-YOLOS)

Object detection is a computer vision technique to recognize and to localize an object in a given image (Keita, 2022). We used object detection in this thesis to separate blurred and irrelevant dirty images from the input dataset. Our aim was to recognize an object to accept it as a clean/reliable input as some customers upload blurred or unrecognizable images to collect the reward.

We used hustvl/yolos-tiny (Fang et al., 2021) pre-trained model that is fine-tuned on COCO 2017 Object Detection which consists of 118K labelled images. It is trained on bipartite matching loss which is the comparison of the union of predicted classes and bounding boxes with true labels. There are 100 classes of objects. For example, if an image has five objects out of 100 pre-identified classes, 95 labels will have "no class" and "no bounding box" as labels. Table 9 contains an example output with labels for the image. Then, the Hungarian matching algorithm is applied to create a mapping for each N queries (where N equals 100 at most). Finally, for classes, standard cross-entropy is calculated. For the bounding boxes, a linear combination of L1 and IoU loss is calculated.

Table 9. Sample Object Detection Output for an Image

Image	Object Detection Output	Label	Confidence Score
	Detected handbag with confidence 0.816 in image	handbag	0.816
	Detected person with confidence 0.73 in image	person	0.713
	Detected person with confidence 0.726	person	0.726
	Detected scissors with confidence 0.654	scissors	0.654

YOLOS model is pre-trained on ImageNet-1K dataset which consists of 118K labelled images for training and 5K labelled images for validation.

3.5. Supervised Machine Learning Algorithms

In this thesis, we used five supervised machine-learning algorithms to classify product reviews. These are logistic regression (LR), gaussian naive Bayes (GNB), decision tree classifier (DT), support vector machine (SVM) and multi-layer perceptron (MLP).

3.5.1. Logistic Regression (LR)

Logistic regression is a supervised machine learning algorithm which has been used mainly in classification problems to predict the probability of a predefined target variable. It takes the continuous output of the logistic regression and sends it as an input to the sigmoid function to predict the probability for each class (Pedregosa et al., 2011). The sigmoid function transforms a continuous variable to a probability between 0 and 1. The parameters which we used in our experiments are, L2 penalty has been used. Suitable for L2 penalty, lbfgs optimization has been used.

3.5.2. Gaussian Naive Bayes (GNB)

Naive Bayes is a supervised machine learning algorithm which has been used mainly in classification problems and it is based on Bayes theorem. In Bayes rule, first, the conditional probability of two events P (X | Y) is calculated. From there, P (Y | X) has been calculated. In the Gaussian distribution, we need to calculate the mean and standard deviation for the training data. When X is a continuous variable and it follows a Gaussian distribution, the probability density of the normal distribution can be subtracted, and it is named Gaussian Naive Bayes (Vats, 2021).

3.5.3. Decision Tree (DT)

Decision tree is a supervised machine learning algorithm which has been used mostly in classification problems. It has a tree structure in which each note represents a label, branches represent decision rules and leaves represent the output. Leaves does not have another branch structure as it has the outcome of that decision (JavaTpoint, n.d.). The parameters which we used in our experiments are gini function (It has been used to ensure the impurity in splitting data) max_depth parameter (It is set to none), min_samples_split parameter (It is set to two. This means, the nodes will be expanded until all leaves contain less than two samples), and max_features parameter (It is set to two as our experiment has two labels).

3.5.4. Support Vector Machine (SVM)

Support vector machine is a supervised machine learning algorithm which has been used mostly in classification problems. SVM separates n-dimensional space into classes and creates the best decision boundary which is called a hyperplane (Pedregosa et al., 2011). The algorithm chooses extreme vectors to help create these hyperplanes. These extreme cases are called support vectors. The parameters which we used in our experiments are, for the kernel, linear kernel type has been used.

3.5.5. Multi-Layer Perceptron (MLP)

Multi-layer perceptron is a type of artificial neural network. The difference between MLP and LR is that there can be hidden layers between the input and output layers in MLP (Pedregosa et al., 2011). The parameters which we used in our experiments are, for the activation function in the hidden layer, the rectified linear unit function which returns f(x) = max(0, x). Also, for the solver in weight optimization, Adam optimizer had been used. Our network has one hidden layer with 100 units. The alpha parameter which is the strength of L2 regularization is set to one.

3.6. Evaluation Methods

In this thesis, five supervised machine learning algorithms (LR, MLP, GNB, SVM, DT) are used to classify product reviews. 5-fold cross-validation is applied in experiments to overcome the problem of overfitting. Overfitting can be explained as the outcome of learning and predicting on the same data and having the perfect score but failing in an unseen dataset (Pedregosa et al., 2011). To avoid this situation, we applied the k-fold cross-validation method in all our experiments.

The procedure for k-fold cross-validation is as follows. Firstly, the dataset splits into k equal-sized smaller sets. Then, for each of the k folds, the prediction model is trained for k-1 of the folds and learns from them. The remaining fold is used for the validation. The performance metrics are calculated by the average metric of each fold. The representation can be seen in Figure 3. Although we used 5-fold cross-validation in all experiments, in the last experiment where we found the best-performed machine learning algorithm, 10-fold cross-validation had been used.

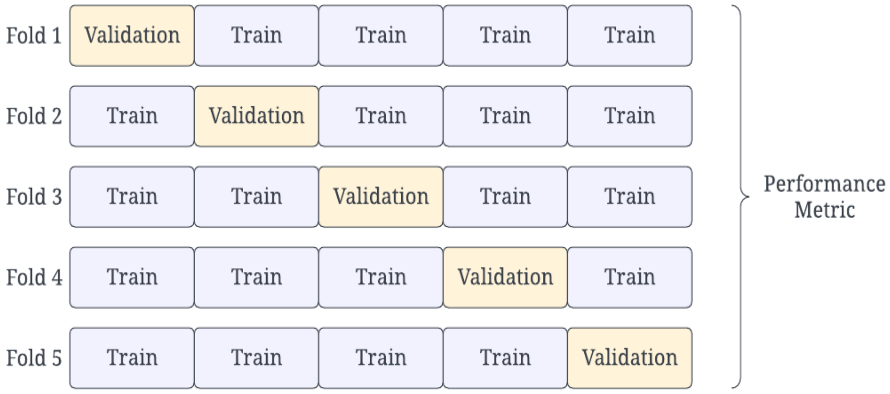


Figure 3. Representation of K-Fold Cross Validation

We reported the classification performance by average accuracy, F1, recall and precision scores for k-folds. Lastly, we applied the statistical ANOVA method and Tukey HSD test to decide the best-performing algorithm with k-fold cross-validation where k=10. In this section, the details on evaluation metrics and how they are employed in this study will be given.

3.6.1. Classification Performance Metrics

Our product review dataset had been labelled with positive and negative labels as covered in section *3.1. Dataset*. We experimented on different algorithms to classify reviews correctly as positive and negative. In Table 10, four product review examples are given as an example to show true label, predicted label and classification group. Here, *Product Review* is the input data which has been given as an input to the machine learning algorithm. *True Label* is the label that the judges assigned to the review after reading it. *Predicted Label* is the output of the machine learning algorithm for that product review after learning from the training dataset.

Table 10. Examples from Classification

Product Review	True Label	Predicted Label	Classification Group
Yanlış ürün gönderildi. (Wrong product sent.)	Negative	Negative	True Negative (TN)
Kullanılabilir ama bu fiyata uygun bir ürün değil! Markasına güvenerek almıştım ama kalitesi konusunda beni şaşırttı. (It can be used, but not a suitable product for this price!! I bought it with confidence in its brand, but it surprised me about its quality.)	Negative	Positive	False Positive (FP)
Şimdilik çok güzel. Ben kaba durmasını istemedim ve kaba gelmedi. Sadece dikişlerini beğenmedim ama onları da boyarım sorun değil. (It's beautiful for now. I didn't want it to seem rude and it didn't come out seemingly rude. I just didn't like the stitches, but I paint them so no problem.)	Positive	Negative	False Negative (FN)
Bir numara büyük almanızı öneririm. Çok beğendim. (I suggest you get a big number. I like it very much.)	Positive	Positive	True Positive (TP)

Accuracy (A) is a widely used metric in classification experiments. Simply, it is the number of correctly classified data over the total number of data as given in below.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The second performance metric is Precision (P). It shows the proportion of true positives to samples that are classified as positives (True positive + False positive) as given below.

$$P = \frac{TP}{TP + FP} \tag{2}$$

The third performance metric is Recall (R). It presents the ratio of true positives to the total number of correct positives in the data set as given below.

$$R = \frac{TP}{TP + FN} \tag{3}$$

F1 is the metric that considers precision and recall values together. It is actually the harmonic mean of precision and recall scores as given below.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

3.6.2. Significance Metrics

In statistics, a significance method tests if a hypothesis is true compared to observed data. There are many significant methods in the literature. In this thesis, we used the One-Way Analysis of Variance (ANOVA) method to decide if supervised machine learning algorithms have significant performance improvement in a given problem. The main reason we choose ANOVA is, it supports more than two groups as we have five classification algorithms. For two groups, generally t-test is being used (Zubair, 2022).

ANOVA tests the null hypothesis. In our experiments, the null hypothesis is "there is no difference in means between classification algorithms". If the null hypothesis is rejected, we will decide which algorithm should be chosen with the outputs of statistical tests. The ANOVA creates f-statistics value which is the ratio of the variance calculated among the means within the samples. The higher f-statistics means the higher chance that there is a difference between groups (Bobbitt, 2021). The ANOVA also calculates p value. If p value is less than constant 0.05, the null hypothesis is rejected. This means there is a significant difference between the means of groups. If p value is greater than constant 0.05, the null hypothesis is not rejected. This means we don't have enough evidence to support the hypothesis and it is not statistically significant.

Representation for ANOVA values can be seen in Table 11. MSB is the mean of the total of squares between groups. MSW is the mean of the total of squares within groups. SST is the total sum of squares. SSB is the sum of squares between groups. SSW is the sum of squares within groups. N is the total number of observations of folds in all groups (Penn State University, n.d.).

Table 11. Representation for ANOVA Summary Values

Source	**Degrees of Freedom DF**	**Sum of Squares SS**	**Mean Square MS**	**F-statistics**	**P-Value**
Between Groups	$k-1$	$SSB = \Sigma nj(\bar{X}j - \bar{X})2$	$MSB = SSB / (k-1)$	$F = MSB / MSW$	$F(k-1,N-k)$
Within Groups	$N-k$	$SSW = \Sigma nj(\bar{X} - \bar{X}j)2$	$MSW = SSE / (N-k)$		
Total	$N-1$	$SST = SSB + SSW$			

When p value is less than 0.05, then there is a significant difference between the means of groups. If there is a significant difference between the groups, post-hoc tests can be used to find which groups differ from each other. In this thesis, we used the Tukey HSD (Honestly Significant Difference) test which is used for making pairwise comparisons between groups. Pairwise comparisons consist of pairs of two different groups' means. For five groups there are a total of $k(k-1)/2$ which is ten pairs of comparisons in this research. We can find the most useful algorithm with the output of pairwise comparison.

CHAPTER 4: EXPERIMENTS AND RESULTS

There are three main research questions that this thesis focused on. These are

- Which transformer-based sentence embedding provides a successful classification of product reviews? (RQ1).
- Is it possible to improve the classification performance by multimodal embeddings that involve image data together with text data? (RQ2).
- Can the image data be used individually in a pre-processing step in the product review classification process? (RQ3)

In this section, we will explain our experiments in detail and present findings to compare the performance of different sentence embedding models and see the effect of multimodal structure in the review classification task. Then, we will summarize the results of object detection by YOLOS and the image-product name similarity method which increases the success of evaluation metrics.

4.1. Transfer Learning Experiments

Transfer learning experiments are conducted with four different sentence embedding models (SBERT-XLMR, SBERT-BERT, SBERT-DBERT and SBERT-CLIP) for text data and two image embedding models (VIT-CLIP and VIT-YOLOS) for image data. In below subsections details on

- transfer learning with text-only data (DS1)
- multimodal transfer learning (DS2)
- object detection by YOLOS
- image-product name similarity

experiments will be presented respectively.

4.1.1. Transfer Learning Experiments with Text-only Data (DS1)

In this set of experiments, we used 15,136 Turkish product reviews (DS1) that are in text format labelled as positive or negative. We converted the input file into .csv file and vectorized each review by

- SBERT-XLMR,
- SBERT-BERT,

- SBERT-DBERT,
- SBERT-CLIP

models. We encoded each category and applied 5-fold cross-validation. We fed each fold to supervised machine learning machines (LR, GNB, DT, SVM and MLP). For each fold, the performance score is measured, and finally, the average scores of folds are reported. These experiments had been repeated for each machine-learning algorithm.

Table 12 presents the results of transfer learning experiments with text. In Table 12, shaded cells refer to the highest scores for each metric. For example, considering all classification methods and embedding models, the highest F1 (92.24%) is obtained when SBERT-XLMR is classified with the LR model. AVG column and row represent the average of the performance metrics.

The outputs observed from Table 12 can be summarized as below.

- SBERT-XLMR is the highest performing embedding technique with the average performances of 89.63% accuracy, 89.99% F1, 89.29% recall and 90.67% precision.
- Considering average values, SVM is the highest performing classifier with 89.00% accuracy and 89.40% F1.
- Considering maximum performance values, LR has the highest performance measures (91.95% accuracy, 92.24% F1, 91.96% recall) except precision.

These experimental results prove that transformer-based sentence embeddings provide successful classification of product reviews (related to RQ1). The highest accuracy in this experiment (91.95%) is higher than the previous researches (Kaynar et al. (2016), Singla, Randhawa and Jain (2017), Reimers and Gurevych (2019), Mishev et al.(2020), Yu and Jiang (2019), Wöllmer et al. (2013), Bhat et al. (2022), Hayran and Sert (2017), Rumelli et al. (2020)) shown in Table 2.1.

Table 12. Classification Results of Transfer Learning Experiments with Text-only Data (DS1)

SBERT-XLMR	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	91.95	89.59	83.98	91.77	90.86	89.63
F1 (%)	92.24	89.62	84.82	92.08	91.19	89.99
R (%)	91.96	86.42	84.36	91.96	91.78	89.29
P (%)	92.52	93.07	84.47	92.21	91.10	90.67
SBERT-BERT	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	89.09	86.59	79.88	89.30	88.68	86.71
F1 (%)	89.47	86.67	80.96	89.70	88.93	87.15
R (%)	89.11	83.85	80.58	89.66	89.97	86.64
P (%)	89.83	89.69	81.18	89.75	88.98	87.89
SBERT-DBERT	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	88.31	80.96	78.16	88.58	87.64	84.73
F1 (%)	88.72	81.20	79.33	88.97	88.07	85.26
R (%)	88.41	79.10	79.38	88.58	87.33	84.56
P (%)	89.04	83.42	79.06	89.37	88.88	85.95
SBERT-CLIP	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	86.21	73.62	70.59	86.35	85.24	80.40
F1(%)	86.68	75.21	72.28	86.85	85.25	81.26
R (%)	86.32	76.94	72.78	86.71	84.87	81.52
P (%)	87.07	73.57	71.89	87.00	86.84	81.28
AVG	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	88.89	82.69	78.15	89.00	88.10	85.37
F1(%)	89.28	83.17	79.35	89.40	88.36	85.91

4.1.2. Multimodal Transfer Learning Experiments (DS2)

In multimodal transfer learning experiments, two processes are followed up. Firstly, we used 1,866 product reviews in text format and calculated the classification performance scores as it was done in the previous experiments. In this group of experiments, we added images of the product reviews and created a multimodal structure (DS2 is employed).

In the second group of experiments, firstly we vectorized product review texts with SBERT-XLMR, SBERT-BERT, SBERT-DBERT and SBERT-CLIP models. Then we vectorized the matching image with the VIT-CLIP model. We concatenated both text and image vectors and created a new vector for each line of the product review.

We encoded positive and negative labels beforehand and sent concatenated vectors and labels to the same supervised machine-learning algorithms. All the experiments were done with 5-fold cross-validation and performance scores had been calculated. The pipeline of multimodal transfer learning experiments can be seen in Figure 4.

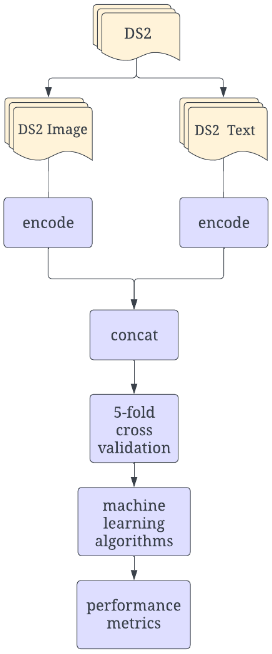


Figure 4. The Pipeline of Multimodal Learning Experiment for DS2

Table 13 represents the results of the text-only and multimodal transfer learning experiments. The highest performances had been shaded with green and the increase in performance in multimodal structure compared to text-only experiments had been shaded with yellow in Table 13. AVG column and row represent the average of performance scores.

The outputs observed from Table 13 can be summarized as below.

- SBERT-XLMR is the highest performing embedding technique in both text and multimodal. It has an average of 90.84% accuracy, 90.74% F1, 90.45% recall and 91.52% precision in text.
- Considering average values, multimodal approach does not have a positive impact on performance. Multimodal average performances are lower than text average performances.

Table 13. Results of Multimodal Transfer Learning Experiments

SBERT-XLMR	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	92.60	92.01	84.19	92.71	92.66	90.84	90.52	92.39	82.53	87.94	90.46	88.77
F1(%)	92.66	91.89	83.58	92.80	92.78	90.74	90.57	92.37	82.26	88.05	90.17	88.68
R (%)	92.36	89.60	84.71	92.78	92.78	90.45	90.23	91.29	82.37	88.00	89.81	88.34
P (%)	93.03	94.34	83.71	92.88	93.64	91.52	90.92	93.51	82.81	88.10	90.46	89.16
SBERT-BERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	91.32	90.30	81.57	90.68	92.18	89.21	88.91	90.35	79.42	86.39	89.28	86.87
F1 (%)	91.35	90.13	82.20	90.71	91.79	89.24	88.97	90.32	80.31	86.47	89.18	87.05
R (%)	90.87	87.79	80.79	90.24	90.66	88.07	88.75	89.17	79.40	86.20	88.96	86.50
P (%)	91.84	92.63	82.32	91.20	91.98	89.99	89.21	91.52	81.54	86.76	88.80	87.57
SBERT-DBERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	88.59	84.14	78.46	89.28	88.80	85.85	84.14	84.67	79.21	83.07	82.85	82.79
F1 (%)	88.68	83.88	77.92	89.32	88.74	85.71	84.39	84.65	77.94	83.09	84.20	82.85
R (%)	88.54	81.74	76.86	88.75	88.33	84.84	85.35	83.86	77.60	82.70	82.91	82.48
P (%)	88.90	86.17	77.76	89.97	89.06	86.37	83.69	85.50	80.40	83.68	83.33	83.32
SBERT-CLIP	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	87.25	76.95	69.45	87.83	87.08	81.71	85.10	78.13	68.86	83.55	85.42	80.21
F1 (%)	87.27	76.69	69.06	87.86	87.64	81.70	85.12	78.01	68.62	83.61	85.61	80.19
R (%)	86.84	74.84	67.20	87.48	87.27	80.73	84.39	76.75	69.42	83.23	84.08	79.58
P (%)	87.84	78.70	70.44	88.35	86.11	82.29	85.92	79.34	70.00	84.03	83.51	80.56
AVG	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	89.94	85.85	78.42	90.13	90.18	86.90	87.17	86.39	77.51	85.24	87.01	84.66
F1 (%)	89.99	85.65	78.19	90.17	90.24	86.85	87.26	86.34	77.28	85.31	87.29	84.70

Considering the average score (last two rows in Table 13), multimodal structure does not seem to improve the classification performance, but it can also be seen that there are increases per machine learning algorithm compared with text experiments. For example, GNB has higher performance with multimodal experiments than with text experiments. Also, there are increases in DT multimodal experiments compared with text experiments, too.

As a result, it can be stated the multimodal results did not provide enough evidence to state that multimodal embeddings increase the success of classification (related to RQ2). However, it provides higher performance results than multimodal experiments of the related work in Yu and Jiang (2019), Wöllmer et al. (2013) and Bhat et al. (2022). In the following sections, two new stages regarding the multimodal approach will be added to the pipeline and experimented on.

4.1.3. Object Detection Experiments with YOLOS

Object detection with YOLOS is considered as a pre-processing step related to RQ3. Since it filters the data beforehand and creates a new dataset to send to the pipeline.

There is usually a reward system when a customer adds an image to the review. This is called the loyalty rewards. For example, SHEIN is a global e-commerce company focused on fashion. According to their loyalty program, when customers upload a photograph to review, they gain 10 points. Every 100 points is 1$ and customers can spend the rewarded points during shopping (SHEIN, n.d.). Another example is SEEN which is an e-commerce company focused on hair and skin care products. Their loyalty program is similar, customers gain 10 points when uploading a review with a photograph and 100 points is 10$. Customers can spend their points any time during shopping (SEEN, n.d.).

To gain loyalty rewards with minimum effort, customers often use irrelevant or very blurred images to quickly add the image and gather the reward. In previous experiments, we saw that some customers added straight-colored or very blurred images where there are no objects in them. Some examples are in Figure 5. These images would not provide a reliable result, so we added an object detection stage to our pipeline.

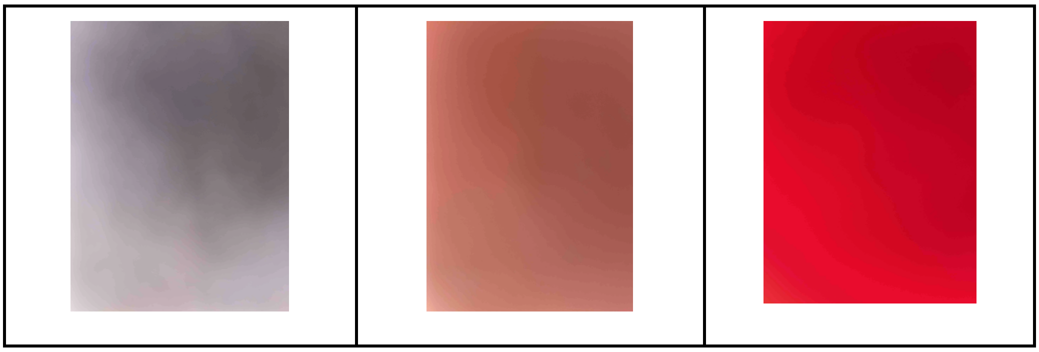


Figure 5. Examples of Very Blurred Images

A set of blurred images are encoded to VIT-YOLOS object detection model and their confidence scores are gathered. The confidence scores for the blurred images ranged between 0.001 to 0.099. Hence, we accepted 0.1 as the lowest threshold for excluding blurred images in our experiment. As the confidence score is the probability of detecting an object in the image correctly, a higher score means higher potential success for the detection of a reliable object. So, our experiments include tests with 0.1, 0.5 and 0.8 confidence scores.

Secondly, we repeated experiments with the dataset that is below the 0.1 threshold and below the 0.8 threshold to see the effect of the image on performance metrics. We called these datasets "remaining" datasets. We created a new dataset for each threshold value and did transfer learning experiments with text and multimodal datasets. The data size for each dataset can be seen in Table 14. The pipeline of object detection experiments with different datasets can be seen in Figure 6, Figure 7 and Figure 8.

Table 14. Object Detection Experiments Datasets

# of Dataset	Dataset	Data Size	Number of Positive Samples	Number of Negative Samples
OD1	Object Detection with 0.1 Threshold	1,681	864	817
OD2	Object Detection with 0.1 Threshold - Remaining	185	61	124
OD3	Object Detection with 0.5 Threshold	1,398	711	687

Table 14 (Continued). Object Detection Experiments Datasets

# of Dataset	Dataset	Data Size	Number of Positive Samples	Number of Negative Samples
OD4	Object Detection with 0.8 Threshold	992	495	497
OD5	Object Detection with 0.8 Threshold - Remaining	874	427	447

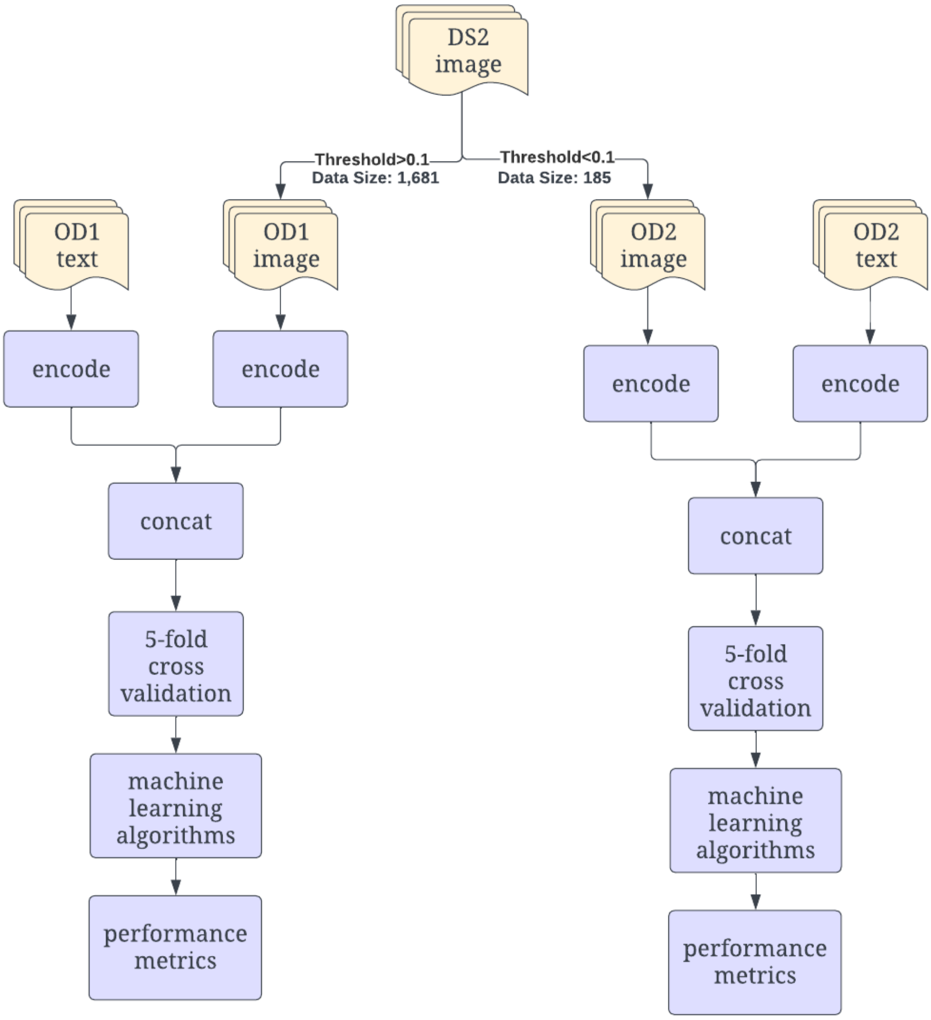


Figure 6. The Pipeline of Object Detection Experiments for OD1 and OD2

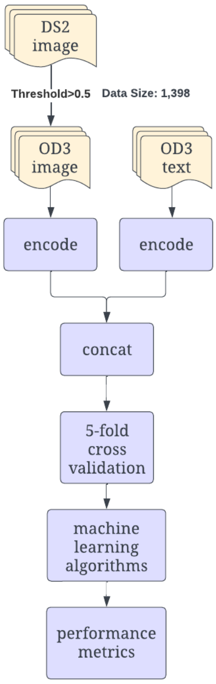


Figure 7. The Pipeline of Object Detection Experiments for OD3

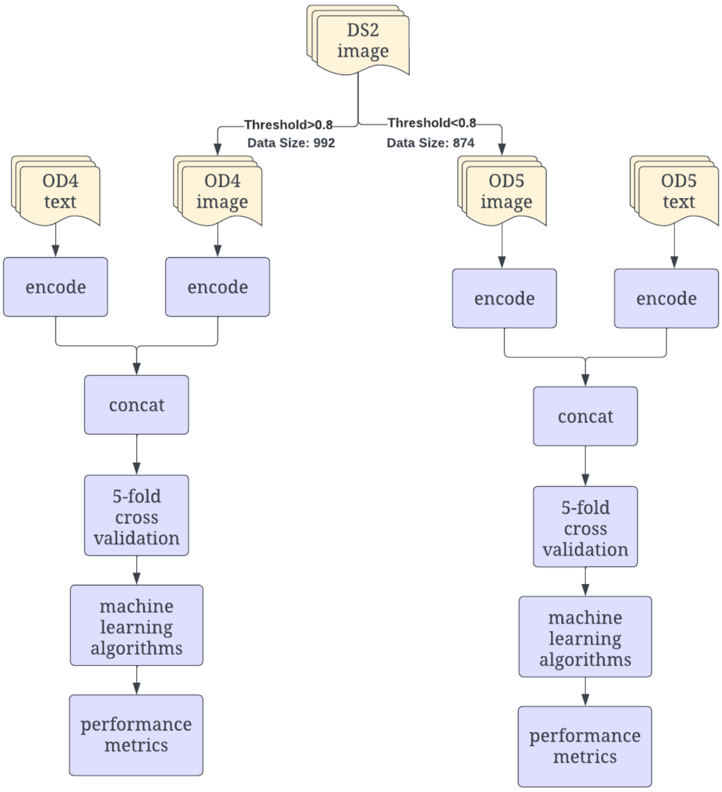


Figure 8. The Pipeline of Object Detection Experiments for OD4 and OD5

In Table 15, results with the OD1 dataset are reported. In Table 16, results with the OD3 dataset are reported. In Table 16, results with the OD4 dataset are reported. Each table represents the results of the text-only and multimodal transfer learning experiments. Highest performances in Tables 15, 16 and 17 had been shaded with green and the increased scores in multimodal structure compared to text-only experiments had been shaded with yellow. AVG column and row represent the average of the performance metrics.

Table 15. Results of OD1

SBERT-XLMR	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	92.50	91.37	83.52	92.03	92.33	90.35	91.14	92.45	84.24	87.87	90.90	89.32
F1(%)	92.30	90.99	83.68	91.82	92.40	90.24	90.86	92.16	83.31	87.61	90.42	88.87
R (%)	91.84	89.16	84.04	91.48	91.11	89.53	90.26	90.99	83.80	87.82	90.75	88.72
P (%)	92.85	92.93	83.93	92.23	92.46	90.88	91.49	93.41	83.69	87.41	90.80	89.36
SBERT-BERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	91.37	90.54	81.74	90.60	91.20	89.09	89.35	90.78	81.20	87.51	89.47	87.66
F1 (%)	91.09	90.08	80.92	90.34	90.65	88.62	89.10	90.48	81.13	87.20	89.63	87.51
R (%)	90.26	87.94	82.10	90.02	90.01	88.07	89.16	89.53	82.46	87.09	89.89	87.63
P (%)	91.98	92.34	80.24	90.68	92.66	89.58	89.07	91.49	79.89	87.37	89.97	87.56
SBERT-DBERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	88.70	84.29	78.58	89.05	89.29	85.98	83.40	85.54	77.99	81.92	83.76	82.52
F1 (%)	88.36	83.64	77.90	88.79	88.55	85.45	83.03	85.11	77.86	81.51	83.03	82.11
R (%)	87.70	82.21	79.53	88.67	88.06	85.24	83.43	84.53	78.44	81.73	82.70	82.16
P (%)	89.09	85.19	78.17	88.99	89.37	86.16	82.88	85.79	78.05	81.37	83.70	82.36
SBERT-CLIP	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	87.86	78.94	71.74	90.60	87.09	83.25	85.48	80.43	69.84	82.99	85.96	80.94
F1 (%)	87.44	78.33	70.88	90.34	87.02	82.80	85.04	80.21	69.71	82.60	85.44	80.60
R (%)	86.72	77.59	71.74	90.02	86.23	82.46	84.53	80.99	70.64	82.71	85.62	80.90
P (%)	88.31	79.14	70.64	90.68	88.29	83.41	85.58	79.56	68.12	82.53	85.61	80.28
AVG	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	90.11	86.29	78.90	90.57	89.98	87.17	87.34	87.30	78.32	85.07	87.52	85.11
F1 (%)	89.80	85.76	78.35	90.32	89.65	86.78	87.01	86.99	78.00	84.73	87.13	84.77

Table 16. Results of OD3

SBERT-XLMR	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	92.13	91.42	83.33	92.20	92.42	90.30	91.27	92.27	83.76	88.27	90.27	89.17
F1 (%)	92.13	91.24	83.42	92.20	92.35	90.27	91.30	92.20	83.70	88.36	90.89	89.29
R (%)	91.45	89.03	83.76	91.74	91.88	89.57	91.17	90.88	82.19	88.61	89.17	88.40
P (%)	92.92	93.61	83.00	92.75	92.79	91.01	91.46	93.62	83.25	88.13	91.39	89.57
SBERT-BERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	91.06	90.63	80.26	90.20	91.20	88.67	88.77	90.42	79.40	86.84	88.91	86.87
F1 (%)	91.06	90.36	80.66	90.26	90.99	88.67	88.81	90.32	79.42	86.99	89.32	86.97
R (%)	90.60	87.47	81.35	90.46	90.47	88.07	88.89	88.75	80.34	87.75	90.46	87.24
P (%)	91.57	93.49	80.56	90.11	91.70	89.49	88.78	92.03	78.74	86.32	88.51	86.88
SBERT-DBERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	87.98	84.26	76.32	87.99	88.06	84.92	83.69	85.84	77.40	81.62	83.33	82.38
F1 (%)	87.97	84.11	76.65	88.05	88.05	84.97	83.72	85.76	77.74	81.78	83.92	82.58
R (%)	87.47	82.91	77.93	88.19	88.18	84.94	83.77	85.04	76.36	82.34	82.62	82.03
P (%)	88.54	85.38	78.08	87.98	88.52	85.70	83.77	86.55	77.83	81.37	81.41	82.18
SBERT-CLIP	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	86.77	79.61	67.95	86.27	86.41	81.40	86.27	81.26	67.10	82.12	85.62	80.47
F1 (%)	86.67	79.35	67.49	86.20	86.72	81.29	86.14	81.14	68.44	82.24	85.90	80.77
R (%)	86.05	77.93	68.24	85.77	85.63	80.72	85.05	80.20	68.25	82.48	85.18	80.23
P (%)	87.43	81.00	68.04	86.72	87.63	82.16	87.29	82.20	68.48	82.01	86.29	81.26
AVG	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	89.49	86.48	76.97	89.16	89.52	86.32	87.50	87.45	76.91	84.71	87.04	84.72
F1 (%)	89.46	86.27	77.06	89.18	89.53	86.30	87.49	87.35	77.32	84.84	87.51	84.90

Table 17. Results of OD4

SBERT-XLMR	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	92.54	91.03	82.16	91.83	92.14	89.94	89.92	92.03	80.04	87.40	89.41	87.76
F1 (%)	92.43	90.74	83.44	91.74	91.76	90.02	89.86	91.92	80.75	87.39	89.81	87.95
R (%)	91.11	88.28	82.02	90.71	91.31	88.69	89.70	90.91	81.62	87.47	89.49	87.84
P (%)	93.85	93.39	81.81	92.93	92.66	90.93	90.05	93.03	80.07	87.31	88.42	87.77
SBERT-BERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	90.53	89.52	78.54	89.72	90.42	87.74	89.01	90.12	77.93	85.99	88.91	86.39
F1 (%)	90.39	89.20	76.55	89.65	90.06	87.17	88.97	89.98	76.98	86.06	88.29	86.05
R (%)	89.29	86.67	76.36	89.09	89.49	86.18	88.89	88.69	75.96	86.67	89.49	85.94
P (%)	91.54	91.94	78.48	90.31	91.29	88.71	89.06	91.37	78.08	85.48	90.07	86.81
SBERT-DBERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	87.50	83.67	75.00	87.80	87.20	84.24	81.25	85.68	75.51	79.94	80.14	80.50
F1 (%)	87.39	83.44	75.85	87.52	87.20	84.28	81.30	85.45	76.19	79.57	80.65	80.63
R (%)	86.87	82.42	73.54	86.06	86.06	82.99	82.02	84.44	75.15	78.79	81.01	80.28
P (%)	88.06	84.56	76.60	89.20	87.46	85.18	80.68	86.58	75.15	80.58	80.70	80.74
SBERT-CLIP	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	84.48	79.74	69.05	84.18	84.28	80.34	84.47	81.35	67.33	83.67	84.78	80.32
F1 (%)	84.32	79.56	70.41	84.15	83.64	80.42	84.36	81.01	69.46	83.57	85.51	80.78
R (%)	83.64	78.79	70.91	84.44	84.44	80.44	84.04	79.80	67.88	83.23	84.04	79.80
P (%)	85.21	80.72	70.31	84.05	85.35	81.13	84.73	82.43	66.39	83.98	85.01	80.51
AVG	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	88.76	85.99	76.19	88.38	88.51	85.57	86.16	87.30	75.20	84.25	85.81	83.74
F1 (%)	88.63	85.73	76.56	88.27	88.17	85.47	86.12	87.09	75.85	84.15	86.07	83.85

The outputs observed from Table 15, 16 and 17 can be summarized as below.

- SBERT-XLMR is the highest performing embedding technique in both text and multimodal in all thresholds.
- Considering average values, multimodal approach does not have a positive impact on performance in all thresholds. Multimodal average performances are lower than text average performances.

- Considering maximum performance values, OD4 text experiment has the highest performance measures (92.54% accuracy, 92.43% F1, precision 93.85%) except recall.
- Considering average values, adding the object detection pre-process step to the pipeline did not increase the multimodal performance (RQ3). Hence, multimodal experiment performances did not increase.
- Between the three multimodal threshold experiments, maximum F1, recall and precision scores belong to OD3 with 92.20%, 91.17% and 93.62% respectively.
- Between the three threshold experiments, the highest average performance scores belong to OD1 with 90.57% accuracy and 90.32% F1 in the text-only experiments.

When we compare these average scores with previous average scores in *4.1.2 Multimodal Transfer Learning Experiments*, there is an increase in accuracy and F1 scores. In *4.1.2 Multimodal Transfer Learning Experiments, the* average text-only experiment has 90.18% accuracy and 90.24% F1. In this section, the average OD1 text experiment has 90.57% accuracy and 90.32% F1. This means that using cleaner/reliable images yields cleaner text reviews. Hence, the performance of text experiments has increased. Also, both average and highest performance values are higher than the previous multimodal studies (Yu and Jiang (2019), Wöllmer et al. (2013) and Bhat et al. (2022)) in the literature.

When we compare text and multimodal experiments per object detection threshold experiments, there is an increase in performance metrics in multimodal experiments. In OD1 experiments, the highest precision of the whole set of experiments belongs to multimodal data with 93.41%. Also, we can see increases in all metrics in GNB and some of the metrics in DT with multimodal experiments compared to GNB and DT with text experiments. These increases had been coloured with yellow in Table 15. In OD3 experiments, the highest precision of the whole table belongs to multimodal data with 93.62%. Also, we can see increases in performance metrics both in GNB and DT with multimodal experiments compared to text experiments. These increases had been coloured with yellow in Table 16. In OD4, there are increases in performance metrics in LR, GNB, DT and MLP with multimodal experiments compared to text experiments. These increases had been colored with yellow in Table 17.

The last experiment in object detection tests is the comparison of threshold data results and remaining data results. Here, we aim to find if a multimodal approach after cleaning images has benefits for our research. For OD1 and OD4 datasets, we stored the remaining data which has a threshold that is below 0.1 and 0.8 in order. These two new datasets are named as Object Detection with 0.1 Threshold - Remaining (OD2) and Object Detection with 0.8 Threshold - Remaining (OD5). These remaining datasets (OD2 and OD5) simply contain dirty image data with text comments. Table 18 and Table 19 represent the results of OD2 and OD5 experiments respectively.

The outputs observed from Table 18 and 19 can be summarized as below.

- The experimental results are conflicting. Considering highest values, multimodal OD5 had higher performance (92.56% accuracy, 92.58% F1) than multimodal OD4 (92.03% accuracy, 91.92% F1) dataset. This proved that separating dirty images object detection did not have a positive effect on multimodal experiments.

As a result, filtering images with a pre-processing step such as object detection did not increase the success (RQ3).

Table 18. Results of OD2

SBERT-XLMR	**Text**						**Multimodal**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	91.89	88.11	81.08	90.81	91.35	88.65	85.41	89.19	72.43	84.32	84.86	83.24
F1 (%)	94.09	90.33	85.54	93.35	93.65	91.39	89.04	91.40	80.86	88.06	89.15	87.70
R (%)	96.70	85.23	83.60	95.90	95.87	91.46	90.13	88.50	80.30	87.70	89.37	87.20
P (%)	91.73	96.44	81.39	91.06	91.67	90.46	88.34	95.17	82.65	88.72	88.40	88.66
SBERT-BERT	**Text**						**Multimodal**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	87.57	86.49	78.92	86.49	88.11	85.51	84.32	84.32	79.46	82.16	83.24	82.70
F1 (%)	90.88	89.59	85.24	90.10	91.20	89.40	88.47	88.10	83.78	86.53	87.52	86.88
R (%)	92.57	86.83	83.53	91.80	92.57	89.46	90.07	86.83	85.30	85.97	88.43	87.32
P (%)	89.61	93.08	86.17	88.78	89.64	89.46	87.69	90.38	84.32	88.07	87.38	87.57
SBERT-DBERT	**Text**						**Multimodal**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	82.70	80.00	76.22	85.41	87.57	82.38	72.43	76.22	75.68	68.11	70.27	72.54
F1 (%)	88.20	83.95	81.37	89.54	91.27	86.87	80.44	81.77	81.62	75.40	77.15	79.28
R (%)	96.73	78.70	83.63	93.40	93.43	89.18	86.83	80.30	82.67	75.33	78.67	80.76
P (%)	81.20	90.07	82.26	86.29	88.54	85.67	75.40	84.36	80.68	76.15	78.20	78.96
SBERT-CLIP	**Text**						**Multimodal**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	84.86	70.27	66.49	79.46	80.54	76.32	75.14	67.57	65.41	74.05	75.68	71.57
F1 (%)	89.16	77.46	76.08	84.91	85.73	82.67	82.14	75.63	71.78	79.70	80.57	77.96
R (%)	94.23	77.03	74.57	87.73	88.50	84.41	86.90	76.23	65.53	78.73	82.03	77.89
P (%)	84.71	78.00	77.88	82.41	83.33	81.27	78.05	75.38	77.27	81.73	80.21	78.53
AVG	**Text**						**Multimodal**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**	**LR**	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	86.76	81.22	75.68	85.54	86.89	83.22	79.32	79.32	73.24	77.16	78.51	77.51
F1 (%)	90.58	85.33	82.06	89.47	90.46	87.58	85.02	84.23	79.51	82.42	83.60	82.96

Table 19. Results of OD5

SBERT-XLMR	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	92.22	92.45	81.46	91.41	92.33	89.98	90.84	92.56	83.98	87.64	89.47	88.90
F1 (%)	92.35	92.36	82.32	91.60	92.19	90.17	91.05	92.58	84.75	87.84	91.39	89.52
R (%)	91.73	89.49	84.11	91.50	91.05	89.58	90.83	91.05	83.22	87.02	90.83	88.59
P (%)	93.03	95.49	82.95	91.74	92.26	91.09	91.32	94.26	83.07	88.78	90.98	89.68
SBERT-BERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	90.84	90.04	79.86	90.05	90.39	88.24	88.78	88.90	78.83	86.27	89.13	86.38
F1 (%)	91.05	89.98	82.26	90.32	90.11	88.75	89.05	88.94	80.39	86.71	88.80	86.78
R (%)	91.27	87.47	82.33	90.82	91.95	88.77	89.04	87.46	79.22	87.47	89.50	86.54
P (%)	90.88	92.73	79.72	89.87	89.72	88.58	89.15	90.59	79.93	85.97	88.95	86.92
SBERT-DBERT	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	88.44	82.49	78.60	88.33	89.25	85.42	78.25	81.00	76.66	79.51	78.71	78.83
F1 (%)	88.71	82.11	77.17	88.69	89.02	85.14	79.13	80.63	77.41	80.25	79.12	79.31
R (%)	88.81	78.95	78.97	89.26	89.48	85.09	80.08	77.84	76.28	81.20	82.10	79.50
P (%)	88.63	85.70	79.21	88.21	88.85	86.12	78.45	83.91	78.62	79.46	79.25	79.94
SBERT-CLIP	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	87.53	74.59	69.56	86.84	86.84	81.07	82.03	73.22	66.93	79.63	82.49	76.86
F1 (%)	87.68	74.44	68.84	87.19	87.51	81.13	82.65	72.72	68.25	80.26	83.02	77.38
R (%)	86.81	72.46	68.44	87.47	87.49	80.54	83.45	70.24	68.24	80.76	83.66	77.27
P (%)	88.65	76.67	71.59	86.97	88.61	82.50	81.90	75.64	69.00	79.81	83.23	77.91
AVG	Text						Multimodal					
	LR	GNB	DT	SVM	MLP	AVG	LR	GNB	DT	SVM	MLP	AVG
A (%)	89.76	84.89	77.37	89.16	89.70	86.18	84.98	83.92	76.60	83.26	84.95	82.74
F1 (%)	89.95	84.72	77.64	89.45	89.71	86.30	85.47	83.72	77.70	83.77	85.58	83.25

4.1.4. Image-Product Name Similarity Experiments

In RQ3, pre-processing steps had been described. Image-product name similarity is considered as a pre-process step. Because the validation begins with the image and the algorithm is fed if the image is related to the product or it takes only the text.

Previous experiments showed that there are some increases in performance metrics in multimodal approach between object detection models but these values are lower than the first experiment with text only experiment. In the previous experiment, we experimented with images that have different qualities using different threshold values. We discovered that some images contain objects that are detected by YOLOS but these objects may not belong to corresponding products. For example, customers may upload their pets' photographs as a t-shirt product review. Most of the time, they upload a random landscape photograph or screenshot of their phone screen. These images are irrelevant to the product and may cause trivial results so we added an image-product name similarity check stage to our pipeline. Some examples for irrelevant images can be seen in Figure 9. The pipeline for image-product name similarity can be seen in Figure 10.

Figure 9. Examples of Irrelevant Images

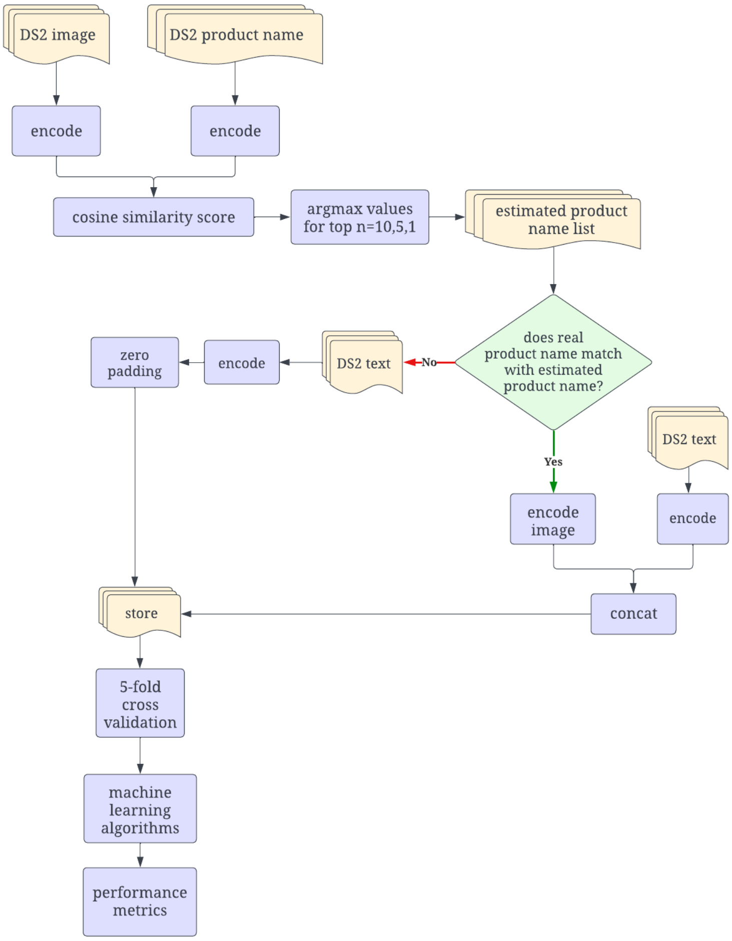


Figure 10. The Pipeline of Image-Product Name Similarity

First of all, we encoded product names for each product with the CLIP text model. Then, we encoded each image in the text-image pair with the CLIP image model. We calculated cosine similarity scores of images with every product name. We stored the top ten, top five and top one highest cosine similarity scores and their corresponding product names. Hence, we had three different lists for possible product names. We searched for real product names in these three different lists, if real product name matches with top n product names, we included that image data in our experiment. In order to achieve this, we encoded text and image data with XLMRoberta text model and CLIP image model because XLMRoberta had the highest performance in previous experiments. We concatenated these vectors and stored them. If real product name does not appear in top n predicted product names, we decided that it is an irrelevant image to the product so we discarded that image and used it’s text review only.

However, text only vectors do not have the same size as concatenated text-image vectors and machine learning algorithms take the same size vectors as an input. In order to solve this problem, we applied zero padding to text vectors and we could store text with zero padding data and multimodal data in the same data structure. Lastly, we applied k fold cross validation which separated the dataset into five folds and sent it to machine learning algorithms to calculate performance metrics.

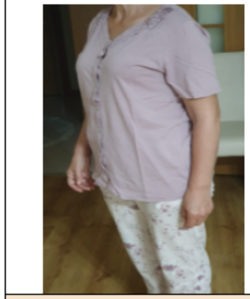
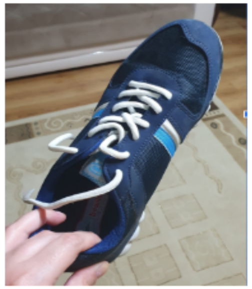
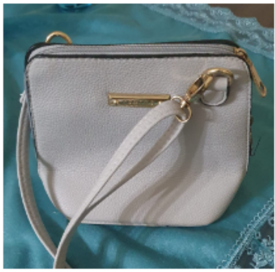
It should be noted that there are three distinct experiments in image-product name similarity experiments and they hold different sizes of multimodal data. These experiments had been done with top ten maximum cosine similarity scored product names, top five maximum cosine similarity scored product names and top one maximum cosine similarity scored product names which means possible product name is equal to real product name. Data sizes for these datasets can be seen in Table 20.

Table 20. Image-Product Name Similarity Experiments Datasets

# of Dataset	**Dataset**	**Multimodal Data Size**	**Text Data Size**
1	Top 10 Image-Product Name Similarity	235	1,631
2	Top 5 Image-Product Name Similarity	156	1,710
3	Top 1 Image-Product Name Similarity	72	1,794

In Table 21, there are examples for similarity outputs for each dataset after the image-product name similarity check.

Table 21. Examples for the Top 10, Top 5 and Top 1 Image-Product Name Similarity

Review Image	**Real Product Name**	**Top 10 Image-Product Name Similarity**	**Cosine Similarity Score**	**Match Position**
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Kadın Gecelik Yumuşak Peluş Pijama Takımı	0.369	1
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Kadın Yaka Düğmeli Çiçek Desenli Uzun Kollu Pijama Takımı	0.362	2
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Desenli Peluş Yazlık Kadın Pijama Takımı	0.353	3
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Kız Erkek Çocuk Sevimli Hayvanlı Pijama Takımı	0.344	4
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Kırmızı Kumaş Düğmeli Pijama Takımı Kadın Pijama Takım	0.336	5
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Kadın Askılı Şortlu Pijama Takımı	0.333	6
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Sabahlıklı Hamile Gecelik Pijama Takımı	0.328	7
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	0.328	8
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Kadın Kırmızı Pijama Takımı Kısa Kollu	0.326	9
	Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım	Kadın Pembe Dantelli Hamile Pijama Takımı	0.326	10
Review Image	**Real Product Name**	**Top 5 Image-Product Name Similarity**	**Cosine Similarity Score**	**Match Position**
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Işıklı Unisex Bej Spor Ayakkabı	0.288	1
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Unisex Günlük Yürüyüş Koşu Beyaz Sneaker Spor Ayakkabı	0.284	2
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	0.282	3
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Mavi Kız Çocuk Spor Ayakkabı	0.280	4
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Bej - Spor Ayakkabı	0.271	5
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Unisex Beyaz Sneaker Yürüyüş Ayakkabısı	0.270	6
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Unısex Beyaz Mavi Soğuğa Dayanıklı Çocuk Spor Ayakkabı	0.268	7
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Deri Bordo Kadın Oxford Ayakkabı	0.265	8
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Kadın Beyaz Bilekten Bağlamalı Bağcıklı Topuklu Sandalet Ayakkabı	0.265	9
	Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı	Lacivert Kız Çocuk Koşu Spor Ayakkabısı	0.265	10
Review Image	**Real Product Name**	**Top 1 Image-Product Name Similarity**	**Cosine Similarity Score**	**Match Position**
	Kadın Mavi Toka Detaylı Baget Çanta	Kadın Mavi Toka Detaylı Baget Çanta	0.276	1
	Kadın Mavi Toka Detaylı Baget Çanta	Kadın Mavi Kapitone Tokalı Baget Çanta	0.253	2
	Kadın Mavi Toka Detaylı Baget Çanta	Kadın Plastik Çanta	0.239	3
	Kadın Mavi Toka Detaylı Baget Çanta	Bej Kadın Çapraz Çanta	0.236	4
	Kadın Mavi Toka Detaylı Baget Çanta	İnce Askılı Kalem Elbise Mavi	0.236	5
	Kadın Mavi Toka Detaylı Baget Çanta	Cıtcıtlı Eşarp Şal Mavi Renk	0.236	6
	Kadın Mavi Toka Detaylı Baget Çanta	Mavi Deri Pantalon Boru Paça	0.235	7
	Kadın Mavi Toka Detaylı Baget Çanta	Mavi Unisex 3 Lü Set Valiz	0.235	8
	Kadın Mavi Toka Detaylı Baget Çanta	14 Inç Laptop Çantası Kılıf	0.235	9
	Kadın Mavi Toka Detaylı Baget Çanta	Fit Cepli Pamuk Şort	0.234	10

In the first example of Table 21, for the Top 10 Image-Product Name Similarity, the similarity check algorithm could find the correct product name in eighth position. Although the real name of the product is “Pamuklu Büyük Beden Gri Düğmeli Yakalı Pijama Takım”, the highest similarity score according to given product review image belonged to “Kadın Gecelik Yumuşak Peluş Pijama Takımı” with 0.369 similarity score. The correct match had a similarity score of 0.328.

In the second example, for the Top 5 Image-Product Name Similarity, the similarity check algorithm could find the correct product name in third position. Although the real name of the product is “Kadın Yetişkin Bağcıklı Yürüyüş Spor Ayakkabısı”, the highest similarity score according to given product review image belonged to “Işıklı Unisex Bej Spor Ayakkabı” with 0.288 similarity score. The correct match had a similarity score of 0.282.

In the last example, for the Top 1 Image-Product Name Similarity, the similarity check algorithm could find the correct product name at first position which means the predicted name was equal to the real name. In the example, the real name of the product is “Kadın Mavi Toka Detaylı Baget Çanta”, and the highest similarity score is 0.276. The correct matches are shaded with green.

In this experiment, we aim to increase performance metrics of the Text experiment in the *4.1.2 Multimodal Transfer Learning Experiments* section as it has the highest performance in order to prove a clean multimodal pipeline can be useful for detecting sentiment analysis of product reviews. Table 22 represents the results of Image-Product Name Similarity experiments. AVG column and row represent average of the performance metrics.

The outputs observed from Table 22 can be summarized as follows.

- Considering highest values, best performance belongs to Top 5 Image-Product Name Similarity experiments with the maximum accuracy of 93.03%, F1 93.08% and precision 94.72%.
- Considering average, Top 5 Image-Product Name Similarity experiments also have the highest performance with 91.23% F1 and 91.88% precision.

Table 22. Results of Image-Product Name Similarity Experiments

SBERT-XLMR	**TOP 10**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	92.71	92.12	82.96	92.77	92.17	90.55
F1 (%)	92.74	92.20	83.29	92.82	92.70	90.75
R (%)	91.94	92.05	84.62	92.47	91.42	90.50
P (%)	93.69	92.52	83.42	93.30	93.19	91.22
SBERT-XLMR	**TOP 5**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	92.98	91.85	83.92	93.03	92.39	90.84
F1 (%)	93.00	91.68	85.34	93.07	93.08	91.23
R (%)	92.26	88.87	85.47	92.58	92.37	90.31
P (%)	93.82	94.72	84.46	93.64	92.76	91.88
SBERT-XLMR	**TOP 1**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	92.50	92.39	82.37	92.55	92.60	90.48
F1 (%)	92.54	92.50	82.22	92.62	92.65	90.51
R (%)	92.05	93.00	85.16	92.47	92.26	90.99
P (%)	93.11	92.04	82.11	92.83	93.57	90.73
SBERT-XLMR	**Text**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	92.60	92.01	84.19	92.71	92.66	90.84
F1 (%)	92.66	91.89	83.58	92.80	92.78	90.74
R (%)	92.36	89.60	84.71	92.78	92.78	90.45
P (%)	93.03	94.34	83.71	92.88	93.64	91.52

- Considering average F1, performance results are higher than text-only experiments in the *4.1.2 Multimodal Transfer Learning Experiments* section.
- It proves that using relevant and clean images in a multimodal dataset with text has higher performance in classifying sentiment analysis than using only text dataset. Hence, the image can be used in the preprocessing step and after the validation, it can be used on a multimodal structure (RQ3).

Following the previous experiment, a more detailed test had been conducted with top five filtered image – product name similarity measurement. These detailed experiment results can be seen in Table 23. We created a hybrid representation from top five concatenated multimodal data and text with zero padding data. We simply called it hybrid as it is a hybrid of text and multimodal approach. TOP 5 Hybrid (Data Size: 1,866) experiment contains 156 multimodal representation and 1,710 text-only data with zero padding as an input. In total, this experiment has 1,866 data size. TOP 5 Multimodal (Data Size: 156) experiment contains 156 multimodal representation which text-image pairs filtered after the top five product name similarity process. TOP 5 Text Only (Data Size: 1,710) experiment contains 1,710 text reviews which their corresponding image reviews does not belong in top 5 image – product name similarity list. TOP 5 Text Only (Data Size: 156) experiment contains 156 text reviews which their corresponding image reviews belong in top 5 image – product name similarity list. Between these experiments, TOP 5 Hybrid (Data Size: 1,866) also had the highest performance.

The outputs observed from Table 23 can be summarized as below.

- Considering F1, recall and precision, the highest performance belongs to the hybrid approach which consists of multimodal data and text with zero padding data (RQ3).

Table 23. Comparison of Top 5 Datasets

SBERT-XLMR	**TOP 5 Hybrid (Data Size: 1,866)**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	92.98	91.85	83.92	93.03	92.39	90.84
F1 (%)	93.00	91.68	85.34	93.07	93.08	91.23
R (%)	92.26	88.87	85.47	92.58	92.37	90.31
P (%)	93.82	94.72	84.46	93.64	92.76	91.88
SBERT-XLMR	**TOP 5 Multimodal (Data Size: 156)**					
	LR	**GNB**	**DT**	**SVM**	**MLP**	**AVG**
A (%)	85.91	90.38	82.06	83.33	85.24	85.38
F1 (%)	85.77	90.08	84.56	82.69	85.59	85.74
R (%)	85.75	88.33	79.42	81.83	84.50	83.97
P (%)	86.16	92.63	88.48	84.39	86.76	87.68

Table 23 (Continued). Comparison of Top 5 Datasets

SBERT-XLMR	TOP 5 Text Only (Data Size: 1,710)					
	LR	GNB	DT	SVM	MLP	AVG
A (%)	92.46	91.93	84.44	92.98	92.63	90.89
F1 (%)	92.51	91.83	84.24	93.06	92.63	90.86
R (%)	92.02	89.71	83.47	92.95	91.68	89.97
P (%)	93.10	94.11	83.61	93.25	93.34	91.48
SBERT-XLMR	TOP 5 Text Only (Data Size: 156)					
	LR	GNB	DT	SVM	MLP	AVG
A (%)	91.65	92.30	82.04	90.38	91.03	89.48
F1 (%)	91.18	91.66	81.80	90.14	91.46	89.25
R (%)	88.25	87.08	80.58	88.33	90.83	87.02
P (%)	95.15	97.50	85.26	92.53	92.75	92.64

It should be noted that these experiments had been done with 5-fold cross validation. In the following experiment, we aimed to find the most successful supervised machine learning algorithm. One-Way ANOVA (Analysis of Variance) analysis and Tukey HSD (Honestly Significant Difference) tests applied to determine the best algorithm statistically. Therefore, we applied k fold cross validation where k equals ten for this experiment. We calculated F1 score for each algorithm for the top five product name similarity multimodal and zero padded text dataset as it produced the best results. The 10-fold cross validation results can be seen in Table 24.

Table 24. 10 Fold Cross Validation for F1 Score

SBERT-XLMR 10 Fold F1	TOP 5 Hybrid				
	LR	GNB	DT	SVM	MLP
Average F1 (%)	93.17	91.68	84.05	92.81	92.73
Fold 1	92.97	91.71	82.61	92.39	92.39
Fold 2	91.80	90.50	84.26	91.40	91.80
Fold 3	95.14	95.14	84.97	94.57	95.08
Fold 4	91.30	89.13	84.32	90.81	90.81
Fold 5	94.68	91.80	88.30	93.68	92.47
Fold 6	93.62	91.71	86.73	94.68	94.62
Fold 7	96.30	93.41	84.95	95.79	95.29
Fold 8	89.69	88.04	83.87	89.69	89.23
Fold 9	92.31	91.98	82.35	92.15	92.15
Fold 10	93.92	93.41	78.13	92.90	93.48

One-way ANOVA aims to find whether the experimental hypothesis is significant or not. The null hypothesis is there is no difference in means. The method takes the input of ten fold values for the five groups of Logistic Regression, Gaussian Naive Bayes, Decision Tree Classifier, Support Vector Classification and Multi-Layer Perceptron. We used the results of the TOP 5: 156 Multimodal + Text (Zero Padding) experiment. The F-statistics value is 32.39144 and p-value is lower than 0.00001. Hence, the result is significant at $p < 0.05$. The null hypothesis, which is that there is no difference in means, is rejected.

However, it can be seen that Decision Tree has very low performance compared with other algorithms. This resulted with a high difference between algorithms and the null hypothesis was rejected. The Data Summary and ANOVA Tables can be seen in Table 25 and Table 26 respectively with this statistical information.

Table 25. Data Summary

Groups	**N**	**Σx**	**Mean**	**Σx^2**	**Std.Dev.**
1	10	931.73	93.173	86847.0659	1.9716
2	10	916.83	91.683	84096.9329	2.0872
3	10	840.49	84.049	70709.4047	2.7297
4	10	928.06	92.806	86162.1106	1.9025
5	10	927.32	92.732	86025.7698	1.9302

Table 26. ANOVA Summary

Source	**Degrees of Freedom DF**	**Sum of Squares SS**	**Mean Square MS**	**F-statistics**	**P-Value**
Between Groups	4	597.0423	149.2606	32.39144	0
Within Groups	45	207.3611	4.608		
Total	49	804.4034			

After determining our experiment is significant, we applied the Tukey HSD test to find the best algorithm. It provides the information of which pairwise comparisons have significant differences. Significant difference is shown with Q in Table 27. Here, the highest significant difference belongs to pairs of T1. Therefore, we can say that to classify sentiments of product reviews, the LR can be used.

However, we can also say that performance scores of T3 are very low compared with other algorithms. When Decision Tree has been used, it fails significantly. Hence, the problem is not completely independent from classifier type. From the Tukey Test, it can be said that Decision Tree is not one of the algorithms to be used in this problem.

Table 27. Tukey HSD Pairwise Comparisons

Pairwise Comparisons			

T1:T2	M1 = 93.17
	M2 = 91.68
		1.49	Q = 2.19 (p = .53508)

T1:T3	M1 = 93.17
	M3 = 84.05
		9.12	Q = 13.44 (p = .00000)

T1:T4	M1 = 93.17
	M4 = 92.81
		0.37	Q = 0.54 (p = .99530)

T1:T5	M1 = 93.17
	M5 = 92.73
		0.44	Q = 0.65 (p = .99052)

T2:T3	M2 = 91.68
	M3 = 84.05
		7.63	Q = 11.25 (p = .00000)

T2:T4	M2 = 91.68
	M4 = 92.81
		1.12	Q = 1.65 (p = .76827)

T2:T5	M2 = 91.68
	M5 = 92.73
		1.05	Q = 1.55 (p = .80934)

T3:T4	M3 = 84.05
	M4 = 92.81
		8.76	Q = 12.90 (p = .00000)

T3:T5	M3 = 84.05
	M5 = 92.73
		8.68	Q = 12.79 (p = .00000)

T4:T5	M4 = 92.81
	M5 = 92.73
		0.07	Q = 0.11 (p = .99999)

CHAPTER 5: CONCLUSION

In this thesis, three main research questions had been answered.

- Success of recent transformers based sentence embedding models had been investigated. Most successful sentence embedding had been decided.
- Effect of the multimodal structure on the success of sentiment classification of product reviews had been investigated.
- Success of two pre-processing steps regarding image data and integrating them to multimodal pipeline had been investigated.

Turkish product review dataset which contains text and image pairs has been used for experiments. Relating pre-processing approaches had been applied to these dataset before each experiment. The contributions and experiments can be summarized below.

- BERT, XLMRoberta, Clip and distilBERT based SBERT models are experimented with text only and multimodal datasets and benchmarked. Text dataset with the XLMRoberta based SBERT model had the highest performance metrics (RQ1).
- Image data had been cleared with object detection using YOLOS. Multimodal experiments had been done with different thresholds but this experiment did not yield a reasonable result (RQ3).
- A hybrid approach with an image-product name similarity stage had been created. This approach proved that if an image is relevant to the product review, it has the highest success in classifying sentiments (RQ2).
- The problem is not independent from the classifier. Using ANOVA and Tukey HSD methods, the Decision Tree algorithm fails significantly.

As future work, we plan to apply SBERT to the seller questions/answers dataset in order to find the similar questions and solve the problem of asking the same questions by customers to increase effectiveness and save time. Also, we are planning to research emotion analysis which is a multiclass problem and experiment it with different categories of product reviews such as food or groceries using the hybrid approach suggested in this paper.

REFERENCES

Ajallouda, L., Najmani, K., Zellou, A. and Benlahmar, E. H. *Doc2Vec, SBERT, InferSent, and USE: Which embedding technique for noun phrases?* In *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. Moulay Ismail University, Morocco. 03 April 2022.

Bhandare, S. and Haribhakta, Y. *Finding Patterns of Questions based on Cognitive Domain of Bloom's Taxonomy Using S-BERT.* In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. Greater Noida, India. 16-17 December 2022.

Bhat, A., Mahar, R., Punia, R. and Srivastava, R. "*Exploring Multimodal Sentiment Analysis through Cartesian Product approach using BERT Embeddings and ResNet-50 encodings and comparing performance with pre-existing models*." In *2022 3rd International Conference for Emerging Technology (INCET)*. Belgaum, Karnataka, India. 27-29 May 2022.

Bobbitt, Z. (2021, August 16). *How to Interpret the F-Value and P-Value in ANOVA* [Blog]. Available at: https://www.statology.org/anova-f-value-p-value/

Bowman, S. R., Angeli, G., Potts, C. and Manning, C. D. *A large annotated corpus for learning natural language inference.* In *Conference on Empirical Methods in Natural Language Processing.* Lisbon, Portugal. 17-21 September 2015.

Briggs, J. (n.d.). *Sentence Transformers: Meanings in Disguise*. Pinecone. Available at: https://www.pinecone.io/learn/sentence-embeddings/ (Accessed: 19 April 2023)

Cer, D. M., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y. H., Strope, B. and Kurzweil, R. (2018). *Universal Sentence Encoder.* ArXiv. Retrieved from https://arxiv.org/abs/1803.11175

Cohere. (2022, August 8). *Pre-Trained vs In-House NLP Models*. Cohere. Available at: https://txt.cohere.com/pre-trained-vs-in-house-nlp-models/ (Accessed: 16 May 2023).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2019). *Unsupervised Cross-lingual Representation Learning at Scale.* In Annual Meeting of the Association for Computational Linguistics.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L. and Bordes, A. (2017). *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data.* ArXiv. Retrieved from https://arxiv.org/abs/1705.02364

Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* ArXiv, abs/1810.04805.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M. and Xu, K. (2014). *Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification.* In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 49-54). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P14-2009. doi:10.3115/v1/P14-2009.

Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J. and Liu, W. (2021) *'You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection'*, CoRR, abs/2106.00666. [Online]. Available at: https://arxiv.org/abs/2106.00666.

Gartner (2022) *Gartner Digital Markets' 2022 Global Software Buyer Trends Survey.* Available at: https://www.gartner.com/en/digital-markets/insights/using-customer-reviews-in-marketing (Accessed: 3 June 2023).

Guo, Z., Kavuri, S., Lee, J. and Lee, M. *IDS-Extract: Downsizing Deep Learning Model For Question and Answering.* In *2023 International Conference on Electronics,* *Information, and Communication (ICEIC)*. Shangri La, Singapore. 5-8 February 2023.

Guven, Z. A. *The Effect of BERT, ELECTRA and ALBERT Language Models on Sentiment Analysis for Turkish Product Reviews.* In *2021 6th International Conference on Computer Science and Engineering (UBMK)*. Gazi University, Ankara, Turkey. 15-17 September 2021.

Hayran, A. and Sert, M. *Sentiment analysis on microblog data based on word embedding and fusion techniques.* In *2017 25th Signal Processing and Communications Applications Conference (SIU).* Antalya, Turkey. 15-18 May 2017.

JavaTpoint. (n.d.). *Machine Learning - Decision Tree Classification Algorithm.* Retrieved from https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (Accessed: 4 June 2023)

Joshi, N. (2020, 7 February). *Exploring the Limits of Transfer Learning.* [Blog]. Allerin. Retrieved from https://www.allerin.com/blog/exploring-the-limits-of-transfer-learning. (Accessed: 12 June 2023).

Minqing Hu and Bing Liu. *Mining and Summarizing Customer Reviews.* In *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA. 22-25 August 2004.

Kaynar, O., Görmez, Y., Yıldız, M. and Albayrak, A. (2016) *Sentiment Analysis with Machine Learning Techniques*. International Artificial Intelligence and Data Processing Symposium (IDAP'16), Sivas, Turkey.

Keita, Z. (2022, September). *YOLO Object Detection Explained* [Blog]. Available at: https://www.datacamp.com/blog/yolo-object-detection-explained

Khan, S. (2019, September 4). *BERT, RoBERTa, DistilBERT, XLNet: Which One to Use?* Towards Data Science. [Blog] Available at: https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8 (Accessed: 21 May 2023).

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Urtasun, R., Torralba, A. and Fidler, S. (2015). *Skip-Thought Vectors.* In NIPS.

Le, Q. V. and Mikolov, T. (2014). *Distributed Representations of Sentences and Documents.* ArXiv. Retrieved from https://arxiv.org/abs/1405.4053

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* ArXiv, abs/1907.11692.

Lu, D., Neves, L., Carvalho, V., Zhang, N. and Ji, H. *Visual Attention Model for Name Tagging in Multimodal Social Media.* In *Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium. October-November 2018.

Madhusudhan, S., Mahurkar, S. and Nagarajan, S. K. *Attributional analysis of Multi-Modal Fake News Detection Models (Grand Challenge).* In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM).* New Delhi, India. 24-26 September 2020.

Marketou, D. (2017, December 29). *7 Reasons Why Customer Reviews are Important* [Blog]. Available at: https://medium.com/@dmarketou/7-reasons-why-customer-reviews-are-important-630c135c2240

Mikolov, T., Chen, K., Corrado, G. S. and Dean, J. *Efficient Estimation of Word Representations in Vector Space.* In *International Conference on Learning Representations*. Scottsdale, Arizona. 2-4 May 2013.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T. and Trajanov, D. (2020). *Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers.* IEEE Access, 8, 131662-131682. doi: 10.1109/ACCESS.2020.3009626.

Pang, B. and Lee, L. (2005). *Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales.* In: Proceedings of the 43rd

Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 115-124. Ann Arbor, Michigan: Association for Computational Linguistics. Available at: https://aclanthology.org/P05-1015 (Accessed: 28 May 2023).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) *"Scikit-learn: Machine Learning in Python."* Journal of Machine Learning Research, 12, pp. 2825-2830.

Penn State University. (n.d.). *Lesson 13.2: The ANOVA Table*. Penn State Online Statistics Course. Available at: https://online.stat.psu.edu/stat415/lesson/13/13.2 (Accessed: 3 June 2023).

Pennington, J., Socher, R. and Manning, C. *GloVe: Global Vectors for Word Representation.* In *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* Association for Computational Linguistics, Doha, Qatar. 26-28 October 2014.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S. (2014). *SemEval-2014 Task 4: Aspect Based Sentiment Analysis.* In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 27-35). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S14-2004. doi:10.3115/v1/S14-2004.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. *Learning Transferable Visual Models From Natural Language Supervision.* In Meila, M. and Zhang, T. (eds.) *38th International Conference on Machine Learning.* 18-24 July 2023.

Reimers, N. and Gurevych, I. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *2019 Conference on Empirical Methods in Natural Language Processing.* Hong Kong. 3-7 November 2019.

Reimers, N. and Gurevych, I. (2020) *'Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation'*, arXiv preprint arXiv:2004.09813. Available at: [http://arxiv.org/abs/2004.09813]

Rumelli, M., Akkuş, D., Kart, Ö. and Isik, Z. *Sentiment Analysis in Turkish Text with Machine Learning Algorithms.* In 2019 Innovations in Intelligent Systems and Applications Conference (ASYU). Yaşar University, Izmir, Turkey. 31 October-2 November 2019.

Saha, R. (2023). *Influence of various text embeddings on clustering performance in NLP.* ArXiv, abs/2305.03144.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* CoRR, abs/1910.01108.

Sasaki, T. and Masada, T. *Sentence-BERT Distinguishes Good and Bad Essays in Cross-prompt Automated Essay Scoring.* In *2022 IEEE International Conference on Data Mining Workshops (ICDMW).* Orlando, USA. 28 November-1 December 2022.

SEEN, (n.d.). *Rewards.* https://helloseen.com/pages/rewards. (Accessed: 7 June 2023)

Sharma, P. (2021, October 30). *Understanding Transfer Learning for Deep Learning.* Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/understanding-transfer-learning-for-deep-learning/ (Accessed: 7 June 2023)

SHEIN, (n.d.). *Bonus Point Program.* https://us.shein.com/bonus-point-program-a-371.html (Accessed: 7 June 2023)

Shreya, G. (2022, November). *Comprehensive Guide to BERT: An Introduction to the Transformer Model for NLP.* Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2022/11/comprehensive-guide-to-bert/

Singla, Z., Randhawa, S. and Jain, S. (2017) *'Sentiment analysis of customer product* *reviews using machine learning'*, pp. 1-5. Available at: doi: 10.1109/I2C2.2017.8321910.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. and Potts, C. *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.* In *2013 Conference on Empirical Methods in Natural Language Processing.* Seattle, Washington, USA. 28 March 2013.

Srivastava, A. (2022, August 8). *What are Transformers in NLP and its Advantages.* Knoldus Blogs. [Blog]. Available at: https://blog.knoldus.com/what-are-transformers-in-nlp-and-its-advantages/ (Accessed: 6 June 2023)

Uçar, T. (2020, April 28). *BERT Modeli ile Türkçe Metinlerde Sınıflandırma Yapmak.* Medium. Available at: https://medium.com/@toprakucar/bert-modeli-ile-t%C3%BCrk%C3%A7e-metinlerde-s%C4%B1n%C4%B1fland%C4%B1rma-yapmak-260f15a65611 (Accessed: 30 April 2023).

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). *Attention is All you Need.* In NIPS.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. and Zhou, M. (2020). *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.* arXiv:2002.10957 [cs.CL]. Available at: https://arxiv.org/abs/2002.10957

Williams, A., Nangia, N. and Bowman, S. *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.* In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana. 2-4 June 2018.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K. and Morency, L.-P. (2013). "*YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context.*" IEEE Intelligent Systems, 28(3), pp .46-53. doi:10.1109/MIS.2013.34.

Yu, J. and Jiang, J. *Adapting BERT for Target-Oriented Multimodal Sentiment Classification,* In *Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization. Macao, China. 10-16 August 2019.

Vats, R. (2021, February 22). *Gaussian Naive Bayes: What You Need to Know?* [Blog]. Upgrad. Available at https://www.upgrad.com/blog/gaussian-naive-bayes/

Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L.-P. *Tensor Fusion Network for Multimodal Sentiment Analysis.* In *2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. 7-11 September 2017.

Zhang, Q., Fu, J., Liu, X. and Huang, X. *Adaptive Co-Attention Network for Named Entity Recognition in Tweets.* In *AAAI Conference on Artificial Intelligence (AAAI).* New Orleans, Louisiana, USA. 2-7 February 2018.

Zubair, M. (2022, September 26). *Statistical Comparison Among Multiple Groups With ANOVA (Stat-11)* [Blog]. Available at: https://towardsdatascience.com/statistical-comparison-among-multiple-groups-with-anova-d4ac27f6e59e