Predictive Text Analytics and Text Classification Algorithms

by

Ahmet Yucel

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6, 2016

Keywords: singular value decomposition, sentiment analysis, predictive modeling, classification and regression trees, machine learning, neural network, text mining

Approved by

Mark Carpenter, Chair, Professor of Mathematics and Statistics
Erkan Nane, Associate Professor of Mathematics and Statistics
Bertram Zinner, Associate Professor of Mathematics and Statistics
Xiaoyu Li, Assistant Professor of Mathematics and Statistics

Abstract

In this dissertation, there are three research studies that are mainly based on text analysis. In the first study, a sentiment analysis is performed for extracting and identifying the general rating of the customer reviews for certain products. Classifying the sentiments of online reviews of products is important in that it provides the ability to extract critical information that can be used to improve the quality. Machine learning (ML) algorithms can be used effectively to analyze and therefore to automatically classify the reviews. The objective of this study is to develop a numerical composite variable from unstructured data for the estimation of the star rates of the customer reviews from different domains by employing popular tree-based ML algorithms by incorporating five-fold cross validation into the models. In the second study, a special text classification is used for extracting and identifying the subjective content of the customer reviews. Classifying people's feedback on a special subject is vital for analysts to understand the public behavior. Especially for the organizations dealing with big bodies of data consisting of people's reviews, understanding the reviews' contents and classify them by the subjective information is very important. Although Information Technology modernized process of data gathering, state of art methods are required to handle the available big data. On the other hand, traditional methods are not capable of delivering profound insights on the unstructured based feedbacks. Therefore, institutions are seeking novel methods for text analysis. Text mining (TM) is a machine-learning approach for dealing with people's reviews that can provide valuable insights about people's feedback. This study proposes a creation of composite variables for the learning process and utilizes Multilayer Perceptron-based Artificial Neural Network. In the third

study, a Turkish TM algorithm is developed for grading written exam papers automatically via TM techniques. Turkish grammar and natural language processing based algorithms are produced on the answer key prepared by the grader and then applied on the answer papers of the students. The main idea in this study is to build a TM tool in Turkish which is going to grade exam papers in Turkish.

Acknowledgments

First, I thank God, who created the hardship and the relief together. I would like to express my deep sense of thanks and gratitude for my adviser, Mark Carpenter, for the scientific advices and excellent guidance he patiently provided throughout my academic studies in Auburn. His excellent ideas both helped me throughout the dissertation and gave me a deep confidence for the future researches.

I also thank the committee members Erkan Nane, Bertram Zinner, Xiaoyu Li, and Latif Kalin for their valuable comments and contributions for the dissertation. Furthermore, I would like to thank my mathematics teacher Fatih Koyuncu who truly inspired me to follow an academic career.

I would like to thank Orhan Tokul the principal of Ankara Demetevler Anatolian Vocational Religious High School/Turkey for providing me with the data set that I used in the last chapter.

Finally, my wife Hanife, my daughter İkra Bilge, and my parents Zeynep- Hasan Hüseyin Yücel and Asiye-Orhan Tokul, I'd like to thank them for their love and supporting, they all kept me going with their prayers.

Contents

Abstract ii

Acknowledgements iv

List of Figures viii

List of Tables xi

1 Introduction 1

1.1 Introduction 1

2 Sentiment Analysis on Customer Reviews with Decision Tree-based Multinomial Classification 6

2.1 Sentiment Analysis on Customer Reviews with Decision Tree-based Multinomial Classification 6

2.2 Methodology 10

2.3 Data Acquisition and Preparation 12

2.4 K-fold Cross Validation 13

2.5 Data Cleansing, Tokenization, Stemming and Phrase Definition 14

2.6 Obtaining a Composite Variable 15

2.7 Singular Value Decomposition (SVD) 18

2.8 Feature Selection 18

2.9 Classification Models 20

2.9.1 Random Forest (RF) 20

2.9.2 Classification and Regression Trees (C&RT) 21
2.9.3 Boosted Decision Trees (BDT) 22
2.10 Performance Evaluation Metrics 22
2.11 Results and Discussion 23
2.12 Conclusion and Future Recommendation 34
3 Employing Text Mining for automatic document classification: Hotel Reviews as a Case Study 36
3.1 Employing Text Mining for automatic document classification: Hotel Reviews as a Case Study 36
3.2 Methodology 40
3.3 Data Acquisition and Preparation 42
3.4 K-fold Cross Validation 42
3.5 Data Cleansing, Tokenization, Stemming, and Feature Extraction 43
3.6 K-means Clustering Algorithm 44
3.7 Creating Composite Variables 45
3.8 Multi-Layer Perceptron-Based Artificial Neural Network (MLP-ANN) 46
3.9 Feature Selection 49
3.9.1 Chi-Square Test based Feature Selection (CSFS) 49
3.9.2 Correlation Based Feature Extraction/Creating Composite Variables 52
3.10 Combination of the features 54
3.11 Results 56
4 Text Mining Algorithm: Grading Written Exam Papers 57

4.1 Text Mining Algorithm: Grading Written Exam Papers 57
4.2 Turkish Grammar 59
4.2.1 Turkish Characters 60
4.2.2 Word orders in Turkish phrases 60
4.2.3 Turkish vowel and consonant harmony chain rules 61
4.3 Stemming 63
4.3.1 Synonyms 64
4.4 Case Study 71
4.5 Results 92
4.6 Future Work: Open-ended questions 95
Bibliography 97
Appendices 104

List of Figures

2.1 An Overview of The Proposed Sentiment Classification Approach 12

2.2 Testing Set - Prediction EER – E-reader Tablet 24

2.3 Testing Set - Prediction EER – Wireless Color Photo Printer 25

2.4 Testing Set - Prediction EER – Hotel reviews 26

2.5 Testing Set - Prediction OACER – E-reader Tablet 28

2.6 Testing Set - Prediction OACER – Wireless Color Photo Printer 29

2.7 Testing Set - Prediction OACER – Hotel reviews 30

2.8 Testing Set - Prediction mean EER – 1000 cases data sets 32

2.9 Testing Set - Prediction mean EER – Original data sets 33

2.10 Testing Set - Prediction mean EER – Comparison 34

3.1 Overall Structure of the Proposed Method 41

3.2 The algorithm for the creation of the composite variables 46

3.3 Schematic of Information Flow For a) A Single Neuron, b) A Multi-Layered ANN 47

4.1 Intersection of the stemming lists 68

4.2 Intersection of the synonyms and related words lists 69

4.3 Answer key 71

4.4 Categories of the answer key 72

4.5 Selected categories of the answer keys 73

4.6 Categorized answer key 74

4.7 Selected category 75
4.8 Original and translated versions of the selected category 75
4.9 Sub-categories of the selected category 76
4.10 Sub-categories of the selected category 77
4.11 Extracted features of the selected sub-categories 78
4.12 Synonym words list extracted from the selected sub-category 79
4.13 Reduced suffix list 80
4.14 Stem+suffix matchings 81
4.15 Selected stem+suffix matchings 82
4.16 Student answer 82
4.17 Original and translated versions of the student answer 83
4.18 Selected sub-categories of the student answer 84
4.19 Sub-categories of the student answer 85
4.20 Extracted features from the seleted sub-category 86
4.21 Intersection of the stemming lists 86
4.22 Answer key 87
4.23 Student answer 87
4.24 Synonym and related words 88
4.25 Stem+suffix matchings 89
4.26 Selected stem+suffix matchings 89
4.27 Intersection of the stemming lists 90
4.28 Selected category from the answer key 90
4.29 Original and translated versions of the selected category 91
4.30 Student answer 91

4.31 Original and translated versions of the student answer 92

4.32 Intersection of the synonyms and related words lists 92

4.33 Comparison of the results 94

4.34 Absolute error rates 95

List of Tables

2.1 Testing Set - Prediction EER – E-reader Tablet 24

2.2 Testing Set - Prediction EER – Wireless Color Photo Printer 24

2.3 Testing Set - Prediction EER – Hotel Reviews 25

2.4 Testing Set - Prediction OACER – E-reader Tablet 27

2.5 Testing Set - Prediction OACER – Wireless Color Photo Printer 28

2.6 Testing Set - Prediction OACER – Hotel Reviews 29

2.7 Testing Set - Prediction EER – 1000 cases data sets 30

2.8 Testing Set - Prediction EER – Original data sets 30

3.1 Neural Network Classification Models Based on The Selected Features 50

3.2 Neural Network Classification Models Based on The Composite Variables 53

3.3 Neural Network Classification Models Based on The Union of The Sets of Independent Variables 54

3.4 The Comparison of The Neural Network Classification Models Performances 56

4.1 Recalculated Grades and Comparison 92

Chapter 1

Introduction

1.1 Introduction

Analyzing user-generated content about a specific product or a company is of a crucial importance in today's competitive world. It provides the business owners with reaching specific information about the weak and strong sides of their product when compared to the competitors'. Therefore, companies have allocated enormous sums for conducting customer satisfaction surveys even though the limitations on the sample size and difficulties on creating effective questions are some of the obstacles that decrease the efficiency of these surveys.

Exponential growth of user-generated data in the Internet and information age has provided an important opportunity to businesses with effectively analyzing the product-centered contents. Such datasets do not only have importance for the business owners but also allows potential customers to learn about the experiences of existing users on a specific product or a company. Examples of this electronic data include the messages posted by users on social media sites, the opinions of users that are posted to product-associated websites, etc. However, such datasets to be analyzed are typically in unstructured format, which is challenging to decipher and analyze. Having said that, effective analysis of such a valuable content can be utilized as a critical tool to help businesses to dramatically improve their business quality. Such analysis would allow companies to answer critical questions such as: 1) What do people think about our product/company 2) What are the weak sides of our product 3) Which functions of the product

should we improve based on the adapted business strategy 4) How much should we financially target when we invest on certain features of the product etc.

Sentiment analysis, also referred to as emotional polarity computation, is a computational process that aims to identify the opinion of a writer with respect to a specific topic. It specifically targets to classify the author's opinion (that is presented in a piece of text) into either binary (i.e. good-bad, positive-negative etc.) or multinomial classes (multi-polarity). Therefore, such analysis has been effectively utilized in analyzing the large, complex unstructured text corpuses to uncover the linear/non-linear relations between the existing words and/or phrases and the overall sentiment of the writer [1-5].

The main goal of the first research given in chapter 2 is to analyze the large, unstructured customer reviews by developing a novel text-mining framework. The proposed study specifically aims at classifying the online customer reviews on a multinomial scale by employing sentiment analysis techniques. The adapted framework also aims to identify the critical factors that determine the reviewers' decision on assigning star rates. The remainder of the chapter 2 is organized as follows. A brief literature review on text classification (particularly used in business related domains) that employed sentiment analysis techniques is provided. The overall text analytics methodology used in this study is presented along with the data collection and preparation phase. A brief discussion on the statistical and machine learning (ML) models that are employed in this study follow this. The multinomial classification results obtained through using the novel framework is presented. Finally, the conclusions are summarized, future research thoughts and directions are discussed in a brief manner.

Institutions are seeking novel methods for improving their products or services to be perfectly matched with their customers' expectations. Therefore, obtaining their feedback is

crucial to provide customer oriented services. Before the advent of internet customer surveys as the major way of data gathering from customers was labor intensive and time consuming. The quality of predefined and structured questionnaires (i.e. sampling, unbiasedness, etc.) is always challenging [1] and might not reflect the all aspects of a subject. On the other hand, unstructured customer interviews may provide better insights [2], but they were even more expensive and challenging.

Recently, the internet facilitated market analysis by its automatic and economic data gathering capabilities. Nowadays, people use the internet as a media to share their ideas about daily life events. Therefore, analyzing the huge crowds of the shared data might provide valuable insights about public opinion on a special subject (e.g. presidential elections, products, places, etc) [3]. Institutions use such type of data to find out their weaknesses and strengths to improve their competitiveness. Knowing previous users' experiences about a service or product is beneficial for the potential customers as well. They can use this data as a decision making tool for service or product selection [3]. An example of such type of data is hotel reviews, which previous customers post their comments about the received services and may use star icons to rank the overall quality. The huge crowds and unstructured nature of the hotel reviews require state-of-the-art methods for data management and content interpreting. Institutions widely employ big data and sentiment analysis methods for investigating the valuable customers' feedbacks about their products or services.

Sentiment analysis is a package of text analysis and natural language processing (NLP) for investigating the author's idea reflected in a text. This methodology basically establishes probabilistic relations between words used in the text corpuses and possible subjects from a reservoir of text to predict the most probable subject that reflects author's opinion. On the other

hand, when the subject is defined, this method is able to measure the performance of the subject based on the writers' text. Customer reviews on a service or product contains valuable information regarding the performance quality. This quality measure is either binary classes (i.e. positive-negative, good-bad, etc.) or multinomial polarity (e.g. 1 star to 5 stars). Sentiment analysis is capable of establishing the probabilistic relation and provides insights about the quality [4-7]. Moreover, the application area is not limited to business. Security agencies (e.g. NSA, DHS, etc.) have employed this idea to identify possible threats (such as terrorist attacks) [8,9]. They established probabilistic relation of some words to a range of threats and can estimate possibility of a potential threat by investigating a text or transcript of voices (e.g. email, phone calls, etc.).

The statistical background of the studies represented in chapter 2 and 3 are mainly based on creating composite variables with high correlation with the dependent variables. The classification models are utilizing these composite variables to make better predictions. The correlation between the independent and dependent variables affects directly the performance of the model. In other words, a strong correlation between the independent and dependent variables let us to build more accurate predictive models based on the independent variable [86].

The study presented in chapter 3 introduces a novel text classification approach for investigating the hotel reviews as the customers' feedback to predict their market section (hotels' cities). The proposed sentiment analysis method creates three composite variables that are obtained through well-known unsupervised clustering algorithm. The results are presented on a multinomial scale for classifying the hotel reviews based on the city they are located. Such methods can be effectively utilized for similar type of classification problems in the text mining concept.

The general steps of the study given in chapter 3 is as given following: Literature review on the applications of text mining in business and market analysis is delivered, the data preparation phase along with the obtained descriptive statistics is presented, the details of the proposed machine-learning methodology that is utilized for investigating the hotel reviews is explained, the results of the proposed methodology and performance comparison of the utilized machine-learning algorithm are delivered, and finally the conclusion and future studies are discussed.

In the study given in chapter 4, a Turkish text mining algorithm is developed which is grading written exam papers automatically via text mining techniques. For the technique, Turkish grammar and natural language processing based algorithms are produced on the answer key prepared by the grader and then applied it on the answer papers of the students. The main idea in the study is to build a text mining tool in Turkish which is going to grade exam papers in Turkish. The general steps of the study are as following: Natural language processing, classification, preparation of dictionaries for the each class, feature extraction, developing algorithms, recalculating the grades based on the algorithms, and comparison.

Chapter 2

Sentiment Analysis on Customer Reviews with Decision Tree-based Multinomial Classification

2.1 Sentiment Analysis on Customer Reviews with Decision Tree-based Multinomial Classification

In the literature, a large body of research exists that employ sentiment analysis techniques to classify text corpuses into categories. Such research can be grouped into two, widely employed, main approaches. The first approach, called '*Semantic Orientation'* [6-8] aims to classify the words into binary classes such as 'good' or 'bad', by mostly using a domain-specific lexical resources. Thus, if a specific text document contains more pre-determined 'good' terms than 'bad' ones, that particular piece of text would be deemed as 'good' and vice versa. The second approach called '*Machine Learning*' (ML), a branch of artificial intelligence, has been employed not only for text analytics but also for variety of areas from finance to health informatics [9-14]. ML models classify text into either binary or multinomial categories by learning through labeled historical data. The conceptual difference between these two main streams is that the machine learning models can be employed effectively in analyzing the sentences that express the sentiments in a subtler manner. To exemplify, to correctly classify a sentence such as "how *could anyone buy this car* "can be challenging since it does not include any explicit negative term, although it obviously reflects a negative opinion about a product. In such situations, the latter approach (machine learning models) can play a crucial role in deciphering the subtlety. In addition, these models do not require any prior knowledge about the data. Therefore, as an efficient and effective set of business-analytics tools, machine-learning

models have been employed successfully to classify texts in large corpuses [11, 15-18]. What follows next is a brief discussion about the research conducted by employing these main approaches.

Semantic orientation of a single word is typically considered as a starting point to analyze the sentiment of the entire document. Therefore, lexical resources such as *WordNet* [19] have been widely utilized in such analysis to automatically identify primarily the emotion-related adjectives. To exemplify the earlier works, Hatzivassiloglou and McKeown [6] focused on semantic orientation of conjoined adjectives by employing a log-linear regression model that uses preselected set of seed words. Soricut and Marcu [20] introduced two probabilistic models that can be implemented by a discourse-parsing algorithm to identify elementary discourse units and build sentence-level discourse parse trees by employing syntactic and lexical features. Kamps et al. [21] developed a distance measure on *WordNet* to determine the semantic orientation of adjectives, by employing a graph theory-based model. Ding and Liu [22] proposed to use some linguistic rules that involve a new opinion aggregation function in determining the semantic orientations of opinions, rather than using the conventional approaches that use a set of opinion words for the same task. Kaji and Masaru Kitsuregawa [23] developed the structural clues to extract polar sentences from documents by aiming at achieving extremely high precision at the cost of recall, which in turn leads to build lexicon from the extracted polar sentences.

Studies that can be considered under the first approach have not only used the lexical resources but some of them have employed unsupervised learning methods in determining the semantic orientation of text. For example, Turney and Littman [7] have introduced a simple algorithm that involves query issuing to a Web search engine, for unsupervised learning of semantic orientation from extremely large corpora. The results obtained were analyzed by

employing point wise mutual information. In a similar study, Turney [24] presented an unsupervised learning algorithm, which uses the average semantic orientation of the phrases to classify the reviews into binary class (i.e. “up” and “down”). In their study, a review is classified as recommended if the semantic orientation of its phrases has associations with the “up” class (e.g. “excellent”) and vice versa. Li and Liu [25] have come up with a novel unsupervised learning-based clustering approach that involves TF – IDF weighting method, voting mechanism and importing term scores. One of the main advantages that were provided by employing such method (when compared with the most of the existing similar work) is that the human participation was eliminated.

On the other hand, the second and final main stream involves machine-learning methods for sentiment classification. As one of the pioneering studies for this stream, Pang and Lee [11] employed Support Vector Machines (SVM), Naïve Bayes (NB) and maximum entropy classification method by considering different features such as unigram, bigrams, combination of both and parts of speech to classify the movie reviews into binary classes. The primary goal of their work was to show that machine learning methods could perform well on classifying the unstructured data rather than comparing their performance with the existing techniques employed in this domain. In a similar study, Chaovalit and Zhou [26] compared the performances of machine learning models with the semantic orientation approach. They found out that the machine learning models slightly outperformed the semantic orientation approach. A similar study from movie domain, Mullen and Collier [17] used the semantic orientation values derived from phrases that were extracted from variety of different sources. Therefore these values were employed to create a feature space, which in turn can be separated into classes using SVM. Boiy and Moens [27] employed Support Vector Machines (SVM), Multinomial Naïve

Bayes (MNB) and maximum entropy (ME) classification method to classify the associated text pieces into three categories such as *negative*, *positive* and *neutral*. The primary goal was to determine the sentiments towards a certain entity (consumption products) in Web sentences written in three languages (English, French and Dutch) by employing the aforementioned supervised learning techniques and reducing the amount of training examples by means of active learning. Prabowo and Thelwall [28] classified the sentiments that were collected from product reviews, movie reviews and MySpace comments by combining General inquirer based classifier (GIBC), Rule-based classifier (RBC), Statistics based classifier (SBC) and Support Vector Machines (SVM). Their results showed that hybridized methods could improve the performance of the classifier on both precision and recall measures. In a more recent study, Sarkar et al. [18] contributed a new approach by employing both a supervised learning algorithm (artificial neural networks) and unsupervised method (k-means clustering algorithm) to classify the sentiments into a multinomial scale (i.e. strong like, weak like, doubtful, weak dislike and strong dislike).

To exemplify the studies that specifically focused on sentiment classification on travel blogs, Ye et al. [29] compared three supervised machine-learning algorithms of Support Vector Machines, Naïve Bayes (NB) and the character based n-gram model to classify (in a binary manner) the sentiments of the reviews for seven popular travel destinations in the U.S. and Europe. Based on the results obtained, NB model was outperformed by the other two machine learning approaches. Having said that, the minimum accuracy level that all three approaches have achieved was 80 % (where the random chance was obviously 50 % in a binary classification case). Wang et al. [30] have introduced a novel approach called Latent Aspect Rating Analysis (LARA) to analyze the opinions expressed about the certain aspect of a product (hotel reviews in their case study). The novel probabilistic approach that was proposed in this

study also considered the relative emphasis on different aspects when evaluating the overall judgment of a specific entity. In a more recent study, Bjørkelund et al. [31] compared machine learning algorithms with lexical-resource semantic orientation methods in terms of sentiment classification performance on a multinomial scale (i.e. strong positive, weak positive, neutral, weak negative and strong negative) for hotel reviews. The study has also incorporated the temporal and spatial aspects of the analysis into the opinion mining process to better visualize the data by using Google Maps features. Based on the existing studies that have done sentiment classification on online (product) reviews, it is evident that they employed the lexical-based resources, unsupervised learning or supervised learning algorithms. In our study, we introduce a novel approach that employs both *semantic orientation approaches* and decision tree – based machine learning algorithms to classify the sentiments into multinomial categories (star rates i.e. 1,2,3,4 and 5). By doing so, the current study presents uniqueness in that it utilizes the semantic orientation approaches during the data preparation phase of the study, then the machine learning approaches that are capable of learning complex relation(s) between the input(s) and the output are incorporated during the second phase of the study (learning and prediction). In the following section, the detailed methodology along with the background information on the data source, preparation and classification models employed in this study are provided in detail.

2.2 Methodology

In this study, a novel sentiment (multinomial) classification methodology (as depicted in Figure 2.1) that consists of four sequential phases is proposed. The first phase consists of the data collection procedure by which the unstructured data is obtained from the associated sites on the World Wide Web and *k-fold* cross-validation where the entire dataset is systematically split into test and train sets. Phase 2, which is considered to be as the data preparation phase, is composed

of three steps: step 1) cleaning stage, where the stops words, nonsense words as well as typos were removed from the unstructured train data; step 2) Feature Extraction stage, where the variables are extracted through tokenization, phrase definition and stemming; and finally step 3) consists of the processes in which SVD concepts are extracted, a composite variable is created via using binary weighting method as well as correlation coefficients of the features and a chi-square feature selection is applied to select the most significant features. Phase 3 represents the processes in which tree-based machine learning multinomial classification algorithms are separately applied to the structured training data for the composite variable and for the set of selected features. In the final phase (phase 4), trained models are employed to predict the star rates (multinomial) that were assigned by the reviewers/customers. During this particular phase, the machine learning models are calibrated until they reach to their optimum performance (via trial and error based experiments) and, the models with best performance (trained models) are then kept for further processes. Finally, the results obtained via applying these algorithms are evaluated. Additional detailed information on each of these phases and steps is provided in the following subsections.

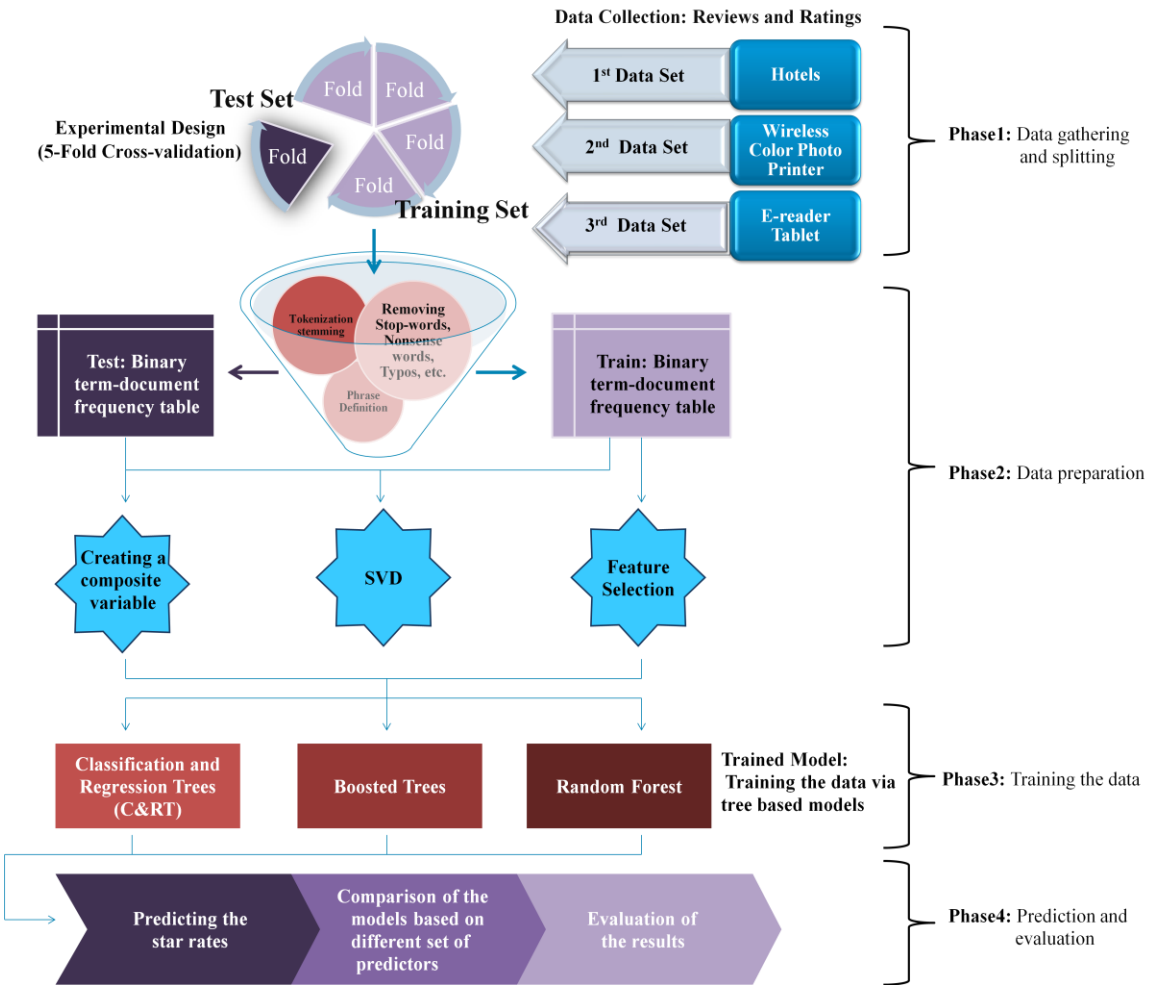


Figure 2.1: An Overview of the Proposed Sentiment Classification Approach

2.3 Data Acquisition and Preparation

Three unstructured data sets used in this study were obtained from Hotels.com which is the world's largest travel site [32] that operates in 45 countries worldwide and has 340 million unique monthly visitors [33] and from Amazon.com which is one of the most popular online retailers. Users leave their verbal comments (textual reviews) about the places that they have visited, hotels they have stayed, or products they have purchased etc. along with the multi-scaled star rating that change from one to five. In this sense, users aim to share their experiences and thoughts with prospective customers to better help them while making travel or shopping

decisions. Having the star rate for all of the reviews (labeled data) is advantageous in that it allows applying supervised learning (machine learning) algorithms in analyzing this large and complex unstructured data. In our analysis, we use 3 data sets at different sizes and from different domains. The data sets consist of 3200 hotel reviews, 2150 e-reader tablet reviews and 1000 wireless color photo printer reviews, along with their associated star rates. Also, the data sets used in our study are balanced in that it contains equal number of each star rate. By doing so, the main goal is to have a balanced datasets that represent each class equally, which in turn leads machine learning models to learn each class in a balanced manner. After completing all the data preparation process we get 3 term-document frequency tables in terms of the 3 data sets. The number of the variables take place in the term-document frequency tables extracted from hotel reviews, e-reader tablet reviews, and wireless color photo printer reviews are 4700, 616, and 385 respectively.

2.4 K-fold Cross Validation

The *k*-fold cross-validation is a widely employed approach that is used for comparing the performances of prediction methods. The main goal in using such approach is to minimize the potential bias that might be associated with the random sampling of training and test data samples [34]. In such technique, the entire data is randomly split into *k* mutually exclusive subsets of approximately equal size rather than splitting it into a single train and a single test set, which can be biased due to the high uncertainty on random sampling. The prediction model is tested *k* times by using the test sets. The overall performance of the models that employed *k*-fold cross validation is calculated by taking the average of the *k* individual performances as follows [35]:

$$CV = \frac{1}{k} \sum_{i=1}^{k} PM_i \quad (2.1)$$

where *CV* represents cross validation, *k* is the number of the mutually exclusive subsets, and *PM* stands for the performance measure that was used in our study. As depicted in Figure 2.1, the stratified 5-fold cross validation approach was employed to the unstructured dataset before beginning to prepare the data for machine learning models. The rationale behind that is the test set (newly added reviews to be tested) would always be obtained in a raw, unstructured manner. Additional details on *k*-fold cross validation can be found in any basic data-mining textbook (see. e.g. [36]).

2.5 Data Cleansing, Tokenization, Stemming and Phrase Definition

Stop words are common in text analytics and should be removed since they do not add any meaning to the sentiment and tend to make the text heavier from the analytics perspective. Therefore, stop words such as *the, a, is, at* etc. have been removed in our analysis. Similarly, words that do not have any meaning (*nonsense words*) as well as typos have been removed from the dataset analyzed in the proposed study.

Tokenization in text mining is simply breaking a stream of text up into tokens. In tokenization, documents (reviews in our case) are broken into meaningful components such as words, phrases, sentences etc. Stemming is simply identifying the morphological roots or bases of the words. Therefore, the words that have same morphological structure should be identified automatically and treated accordingly so that it could lead to a more accurate prediction performance. In our analysis, both tokenization and stemming processes have been employed through using commercial software *STATISTICA* 11 (www.statsoft.com).

In sentiment analysis, phrase definition plays an important role in that it helps discriminating the subtle structures that were caused by the change in the meaning of a phrase when the (immediate) neighbor words are taken into account. Also, longer phrases tend to be more informative in terms of identifying the polarity of the sentiments. For example, while term "convenient" is likely a positive sentiment, "not convenient" or "not very convenient" are less likely to appear in positive comments. Therefore, models like bag-of-unigrams or bag-of-bigrams would have the tendency to fail to handle "not convenient" and "not very convenient", respectively. Also, the existing related literature [37, 38] showed that "*sentiment classifiers combined with high order n-grams as features can achieve comparable, or better SA performance than state of the art on large-scale data sets*". In our proposed analysis, high order n-grams have been utilized in classifying the hotel users' reviews.

2.6 Obtaining a Composite Variable

The main idea of composite variables is very similar to the composite materials. A composite material is a combination of so many small and weak materials in a single and stronger body. For a composite variable, by using some special calculation techniques, so many weak learners and classifiers are brought together in a stronger variable which is called as a composite variable and this variable is used as a stronger indicator and classifier. The composite variables can be used in very different domains. For example, a composite variable can be used for criminal issues. It can be constructed based on several criminal categories and then it can be used for predicting the similar criminal events before they happen [87]. Another domain that the composite variables can be utilized is economy. OECD supports some researches based on composite variables. The main aim of the researches is to provide some special composite variable construction techniques based on the economic information so that the composite

variables can be used for comparing the countries' economic performances [90]. Also, in another research, a composite variable is constructed on the reasons affecting express package shipping flow, and then it is used for improving the shipment service design [88]. In addition, composite variables can be used as a dimension reduction technique for vary large and complex datasets. By applying a simple clustering algorithm based on the composite variable, the dimension of the dataset can be reduced [89]. So that it is easier and simpler for extracting information out of it.

In sentiment classification, initial observation of the term (phrase) effect on the outcome is critical since this in turn would allow the learning algorithm to gain raw information about the unstructured data. For our case, since the satisfactions or dissatisfactions of the customers consist of similar expressions, particular terms are expected to have higher occurrence among the reviews from the same star-rate group. Therefore, correlation scores can be used to extract information about the strength of the expressions that are used in the reviews. The correlation scores between the terms (phrases) and the outcome (star-rates) can provide important information about the terms that have either negative or positive or neutral effect on the star-rates.

Given a set of observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the formula for computing the correlation coefficient is given by:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \qquad (2.2)$$

Where r represents the correlation score between a particular variable (feature) and the response (outcome), and n represents the number of observations.

For the current sentiment classification problem, it should be noted that, test data should not be included in the analysis in calculating the correlation score of a feature. The rationale behind this is that the machine-learning algorithms should not consider any observation that belongs to the test set during the learning procedure. Therefore, in our analysis, the correlation scores have been computed five times independently, since 5-fold cross-validation procedure has been employed. By calculating the correlation score for each feature in the training set, the following correlation vector can be obtained:

$$\boldsymbol{R} = (r_1, r_2, \dots, r_f) \tag{2.3}$$

Where *f* denotes the number of features extracted from the associated training fold.

Another important element that is considered in our sentiment analysis in creating a composite variable is to see the existence of a particular term in the documents. Binary weighting score is a widely-employed scoring technique that is used to obtain such information. By doing so, binary weight of a particular feature in a specific document can be represented through the following matrix:

$$\mathbf{BW} = \begin{bmatrix} bw_{11} & bw_{12} & . & . & . & bw_{1f} \\ bw_{21} & bw_{22} & . & . & . & bw_{2f} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ bw_{n1} & bw_{n2} & . & . & . & bw_{nf} \end{bmatrix} \tag{2.4}$$

After obtaining the aforementioned correlation vector (***R***) and binary weight matrix (**BW**), which are the main measurements to evaluate the contribution of each term to the corresponding document, these two main elements are multiplied to create a new matrix that is expected to contain more information in determining the effect of the features on the outcome.

Finally, sum of the elements in a row is calculated and divided by the number of terms that have a binary weighting score of 1. The rationale behind this division process is that since the intensities of the documents are different, a term might cover very important idea in a short review while the same term might cover very little part of a long review. Therefore, the number of terms that have a binary weighting score of 1 are also taken into account in creating a composite variable as can be shown by the following formula:

$$(\mathbf{BW\ X\ R})/T = \begin{bmatrix} (bw_{11} * r_1 + bw_{12} * r_2 + \ldots\ldots\ldots + bw_{1f} * r_f)/T_1 \\ (bw_{21} * r_1 + bw_{22} * r_2 + \ldots\ldots\ldots + bw_{2f} * r_f)/T_2 \\ . \\ . \\ (bw_{n1} * r_1 + bw_{n2} * r_2 + \ldots\ldots\ldots + bw_{nf} * r_f)/T_n \end{bmatrix} \tag{2.5}$$

where T represents the number of terms in a document. Therefore, the composite variable, C, can be presented in a vector format that has $N\ X\ 1$ dimensions as following:

$$\mathbf{C} = \begin{bmatrix} c_1 \\ c_2 \\ . \\ . \\ c_n \end{bmatrix} \tag{2.6}$$

The procedure discussed above can also be summarized by applying the following formula:

$$C_n = \sum_{j=1}^{f} \frac{(BW_{nj} \times R_j)}{T_n} \tag{2.7}$$

where C_n represents the value of the composite variable for the n[th] document. After obtaining such composite variable, it can now can be used as the main feature in our sentiment classification problem, in that it contains very rich information that was gathered by using both correlation and binary weighting methods. Therefore, all of the other features that are utilized in calculating the composite variable have been eliminated since the new composite variable already encapsulates information that was gathered by combining these features.

2.7 Singular Value Decomposition (SVD)

Singular value decomposition is a very well-known matrix dimension reduction technique and is used for handling big dimensional data sets in statistics. In this study it was utilized as a feature extraction technique. Therefore the SVD concepts would be the features. Since SVD can handle even very large matrices of word counts and documents, it is very common technique in text mining. Let A be an *m x n* term-document frequency matrix where m is the number of documents (reviews) and n is the number of extracted (selected) features. SVD computation produce the *m x r* orthogonal matrix *U*, *n x r* orthogonal matrix *V*, and *r x r* matrix *D* so that *A = UDV'*, where *V'* is the conjugate transpose of *V*, and *r* is the number of eigenvalues of *A'A*, similarly *A'* is the conjugate transpose of *A*. The concepts matrix that we deploy in to the models is corresponding to the matrix *U* which is called as document score matrix [49].

2.8 Feature Selection

For an appropriate text classification you need an appropriate data structure to represent the text data and appropriate objective functions to avoid overfitting to data and appropriate

algorithms to deal with the high dimensional matrices without losing so much information. Therefore for text miners one of the most challenging tasks is automated text classification. Feature selection is the process of selecting features that are going to be used in the text classification. It is the first and the most important step of text classification. It is used to simplify and speed up the run process of the learning algorithms. Feature selection is only applied on the features extracted from the train set [50].

From the vocabulary extracted from the train, the more effective ones are selected. In other words, from a set of complex classifiers we extract a simpler one. There is no any theoretical guarantee in general high performance. While a selected set of features perform very well, the other set of selected features does not. Indeed selecting the most optimal set of features is very intuitive, needs so many experience and so it is very difficult in general.

χ^2 is an important feature selection technique. In statistics it is used to check the independence of two events. From the feature selection perspective, instead of events, we are going to check occurrence of the terms, and rank them all with respect to the following measurement:

$$X^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - N_{e_t e_c})^2}{E_{e_t e_c}} \tag{2.8}$$

where D is document, e_t and e_c (term t and class c) are defined as e_t = 1 (if term t exists) and e_c = 0 (if t does not exist), N is the observed frequency in D and E is the expected frequency.

χ^2 measurement shows the deviation between the expected E and observed N. A high χ^2 means that the hypothesis of independence is incorrect. In other words, the counts of expected and observed are close. Dependency of the two events indicates that the occurrence of the term is related with the occurrence of the class and thus it is an important feature for the class [51].

2.9 Classification Models

As it is well known in the literature, the performance of a machine learning (ML) algorithm depends heavily on the dataset and dimension. Thus, a reasonable way to select an efficient ML algorithm should be based on trial and error experiments. Therefore, based on our preliminary results, tree-based ML algorithms (i.e. Classification & Regression Trees, Boosted Decision Trees and Random Forests) outperformed their counterparts in classifying the star rates that were assigned along with the unstructured hotel reviews. What follows is a brief description of the employed models, since these prediction algorithms are well known ML models, which in turn detailed information can be easily reached in most of the ML sources if needed.

2.9.1 Random Forest (RF)

Random forests (RF) are known to be as efficient and robust to noise [39] prediction algorithms by which the final classification decision is made based on a "voting" concept. They are usually efficient on large databases and capable of handling large number of features and missing data. When a new observation needs to be classified, each of the trees in the forest considers this specific observation as a vector and assigns or "votes" for a class for this particular case (star rates in our case). Finally, the decision is made based on the number of votes that were assigned by multiple decision trees. In other words, the class that has the highest number of votes assigned is assigned as a final decision [40].

The following steps can summarize the basic tree growth procedure in RF algorithm:

- A random sample is drawn from the entire data such that the size of this sample is equal to the number of cases in the training set.
- A number "v", which needs to relatively much smaller than the number of features (V) in the training set, is specified and kept constant through the tree-growing process. Then at

each node, v variables are randomly selected out of the V features. Finally, split node is selected as the best split on these v features that were selected.

- The trees are grown (without any pruning) to the largest extent possible via the nodes that were split using the aforementioned selected features.

2.9.2 Classification and Regression Trees (C&RT)

Decision trees are easy to interpret algorithms [41] and therefore, they have been widely employed in several prediction problems [42]. The classification procedure can be briefly described as follows:

The original training data is split into several subsets such that each of these subsets consists of more or fewer homogeneous states of the target variable [43]. By doing so, the effect of each independent feature on the target/outcome is measured. This, simple procedure occurs until a stable state is reached.

ID3, C4.5, C5 [44, 45] and C&RT (Classification and Regression Trees) [43] are well-known, popular decision tree algorithms. In our current study, the C&RT algorithm has been selected to employ, due to its high performance in the trial & error based preliminary analysis, when compared to other aforementioned counterparts.

2.9.3 Boosted Decision Trees (BDT)

Boosted decision trees, as implied by their name, aims to fix the misclassification problems occurred in Decision tree models. The boosting methods in decision trees are considered to be as remedies to the problems caused because of huge changes happen in the final tree structure when there is a small change in the training data [46, 47]. In such misclassification situations, the deviations of the predicted values from the respective means (residuals for each

partition) are computed to re-weight each training example by how incorrectly they were classified. This procedure then repeats for the new tree obtained. By doing so, many trees are built up and the final decision is made based on 'voting' of the weighted scores of the individual leaves. Finally, such 'additive weighted expansions' concept aims to eventually produce an excellent fit of the predicted values to the actual ones.

2.10 Performance Evaluation Metrics

For classification problems, confusion matrices are commonly used for comparing the classification performances. In binary classification problems, depending on the problem type, there are well-known performance criteria (i.e. sensitivity, specificity and area under the ROC curve etc.) that are widely used to compare the classification models. However, in multinomial classification problems, such standardized criteria do not apply. Therefore, in our multi-class prediction problem, two different metrics were used to evaluate and compare the models' performances: 1) *Exact error rate* (*EER*), which is calculated by taking the ratio of the correctly classified samples to the total number of samples in the test sets, and 2*) One-Away-Class Error Rate* (*OACER*) [48], which not only considers the correctly classified classes but it also takes all of the other one-away misclassified classes into account. Therefore, in this criteria, one-away misclassified estimations are considered as correct. For example, for the observed rate '3', estimated rates '2' and '4' are treated as correct along with '3'. And then in the same manner with *EER*, the ratio of the number of correctly classified samples to the total number of samples in the test gives the *OACER* for the model. The mathematical details of the error rate calculation methods *EER* and *OACER* can be represented in Eqs (2.9)-(2.10), repectively.

$$EER = \frac{e}{n} \tag{2.9}$$

$$OACER = \frac{e+o}{n} \quad (2.10)$$

where e is the number of exact estimations, o is the number of the misclassified estimations that take place in the one away neighborhood of the exact rate, and n is the total number of samples.

2.11 Results and Discussion

As it has been discussed in the previous sections, machine learning classification models employed in this study aim to predict the star rates (out of 5) that were assigned by the reviewers, through analyzing the verbal comments they made for the associated service/product. The performance of the ML models is measured via using two accuracy metrics in the current study. The first one, *EER*, is a conservative metric in that it only considers the correctly classified cases (reviews). In tables 1, 2, and 3 the *EER*, and in tables 4, 5, and 6 the *OACER* of the testing models based on the 3 different independent variable sets are presented for each fold and data set separately. The rationale behind presenting them separately is to better visualize the success of each ML models on a finely-grained manner.

Indep. Variable Set	Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Composite Variable	**Random forest**	67.91	73.72	67.44	70.70	68.6	69.67
	C&RT	69.30	74.42	69.07	70.00	69.07	70.37
	Boosted trees	71.16	73.26	70.47	72.56	71.40	71.77
	Mean	69.45	73.8	68.99	71.08	69.69	70.60
SVD Concepts	**Random forest**	71.16	74.19	68.60	72.79	71.63	71.67
	C&RT	72.33	75.35	68.84	73.49	71.16	72.23
	Boosted trees	72.09	75.58	71.40	74.42	73.26	73.35
	Mean	71.86	75.04	69.61	73.56	72.01	72.41
Selected Features	**Random forest**	80.00	79.77	79.07	80.00	80.00	79.76
	C&RT	80.00	80.00	80.00	80.00	80.00	80.00
	Boosted trees	79.07	78.84	79.3	80.00	79.77	79.39
	Mean	79.69	79.53	79.45	80.00	79.92	79.72

Table 2.1: Testing Set - Prediction EER – E-reader Tablet

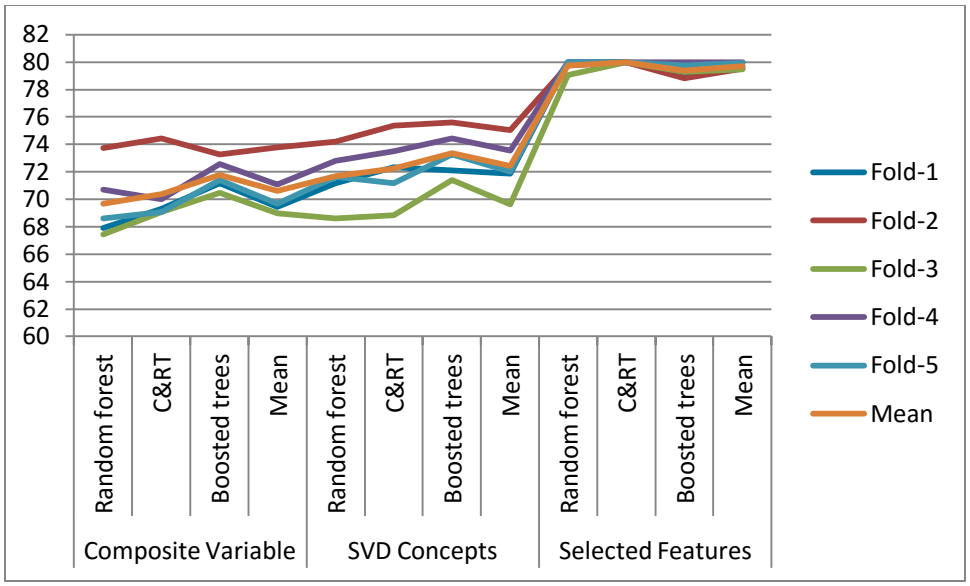


Figure 2.2: Testing Set - Prediction EER – E-reader Tablet

Indep. Variable Set	Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Composite Variable	**Random forest**	70	69.5	76	72.5	72	72
	C&RT	69.5	70.5	76	71	79.5	73.3
	Boosted trees	70	73.5	76	70	72	72.3
	Mean	69.83	71.16	76	71.16	74.5	72.53
SVD Concepts	**Random forest**	75	69.5	79.5	74	75	74.6
	C&RT	75	76.5	80	75.5	80	77.4
	Boosted trees	73	79.5	76	78	80	77.3
	Mean	74.33	75.16	78.5	75.83	78.33	76.43
Selected Features	**Random forest**	78.5	80	80	77.5	80	79.2
	C&RT	80	80	80	80	80	80
	Boosted trees	76.5	80	79	77.5	75.5	77.7
	Mean	78.33	80	79.66	78.33	78.5	78.96

Table 2.2: Testing Set - Prediction EER – Wireless Color Photo Printer

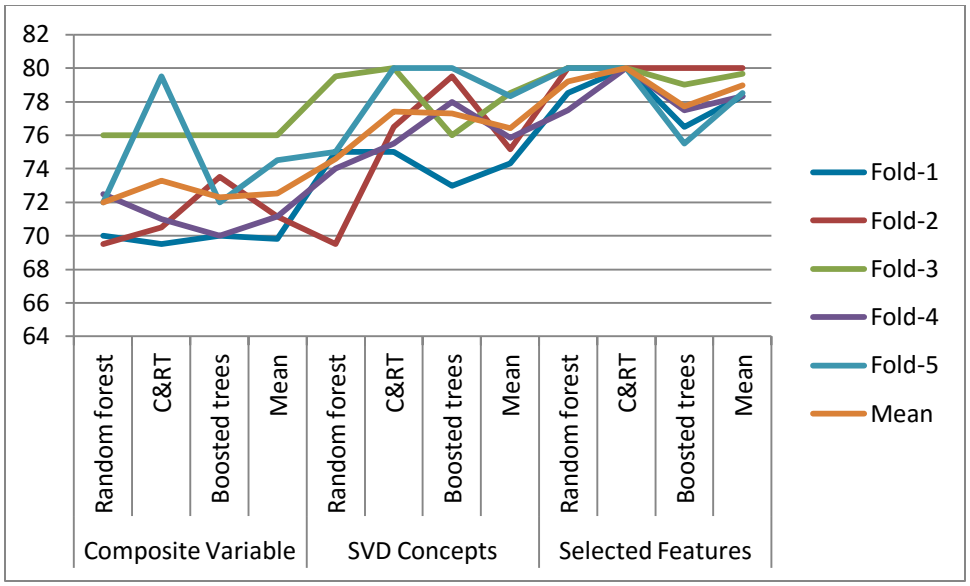


Figure 2.3: Testing Set - Prediction EER – Wireless Color Photo Printer

Indep. Variable Set	Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Composite Variable	**Random forest**	49.84	50.81	46.94	45.81	49.84	48.35
	C&RT	49.52	51.45	46.61	45	50.65	48.14
	Boosted trees	53.23	50.65	45.81	43.23	50	48.23
	Mean	50.86	50.97	46.45	44.68	50.16	48.24
SVD Concepts	**Random forest**	70.65	69.35	68.23	67.9	65.97	68.42
	C&RT	71.13	70.48	69.52	73.55	71.94	71.32
	Boosted trees	60.97	61.94	58.87	55.97	57.42	59.03
	Mean	67.58	67.25	65.54	65.8	65.11	66.25
Selected Features	**Random forest**	76.13	76.61	78.06	77.9	77.58	77.17
	C&RT	80	80	80	80	80	80
	Boosted trees	75.32	74.84	75.32	75.65	77.42	75.28
	Mean	77.15	77.15	77.79	77.85	78.33	77.48

Table 2.3: Testing Set - Prediction EER – Hotel Reviews

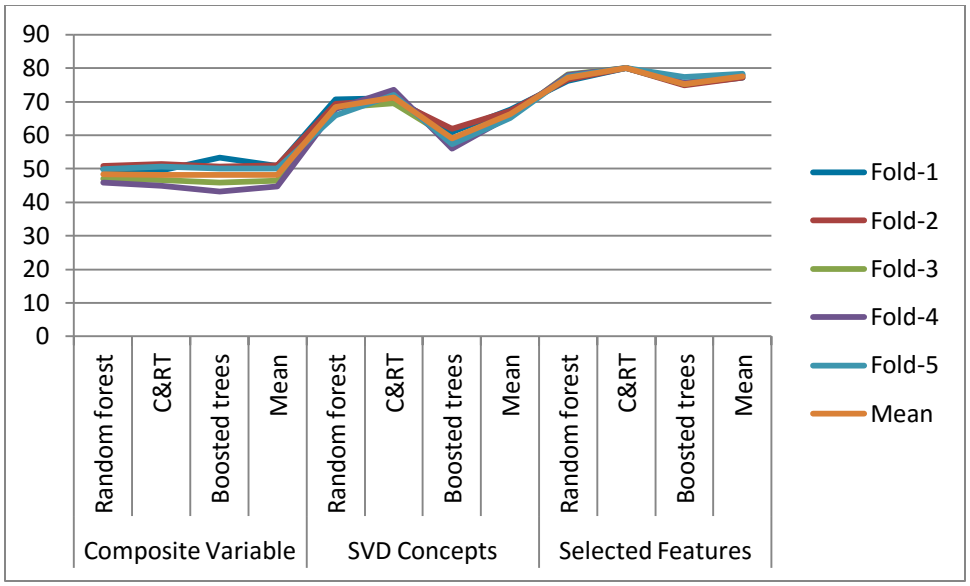


Figure 2.4: Testing Set - Prediction EER – Hotel Reviews

In our sentiment classification problem, there are 5-star rates to be classified by the ML models employed in this study. Therefore, correctly classifying the star rate of any review would be challenging since the chance for a random guess is only 20% (1/5 = 0.2). For this reason, in order to better compare the ML model performances, *OACER* can play a critical role in comparing the classification models in a more objective manner.

It should be noted that the performance of a classification model should be evaluated and judged by taking the expected error values into account. For our case, if the assignment (classification) was made on a random fashion, the expected error rate would be 80 %, since the outcome has 5 different classes (100% - 100⁄5% = 80%). Therefore, achieving an *EER* of less than 50% can be considered as a great success since it means that there is more than 200 % improvement over a random guess, in the case that the ML models are employed in such classification problem instead of making random guess. It is very clear that for the all data sets, for each fold, and for each error rate evaluation criteria the composite variable that we created based on the correlation between the selected terms and loc_txt categories, performs better than

both SVD concepts and selected features based on chi-square feature selection method.

The ML models based on the composite variable, has not only shown a great performance on the first metric that was used, but they have also shown a great performance on *OACER* results. The classification models employed in the proposed method have shown a similar error rate pattern (in terms of predictor sets: composite variable/SVD concepts/selected features) in both the first (*EER*) and the second evaluation criteria (*OACER*). The composite variable plays better classifier role than the other two classifier sets in both evaluation criteria.

Indep. Variable Set	Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Composite Variable	**Random forest**	30.4	32.5	28.6	31.1	31.1	30.74
	C&RT	30.6	30.6	30	31.8	31.3	30.86
	Boosted trees	30.2	32.7	27.9	32.7	41.8	33.06
	Mean	30.4	31.93	28.83	31.86	34.73	31.55
SVD Concepts	**Random forest**	31.3	35.1	30.9	33.2	37.4	33.58
	C&RT	32.09	34.6	31.1	37.2	38.6	34.71
	Boosted trees	37.4	35.5	33.7	36.04	40.2	36.56
	Mean	33.59	35.06	31.9	35.48	38.73	34.95
Selected Features	**Random forest**	80	59.5	59.06	80	57.9	67.29
	C&RT	80	80	80	80	80	80
	Boosted trees	57.6	56.5	56.7	57.9	56.2	56.98
	Mean	72.5	65.33	65.25	72.63	64.7	68.09

Table 2.4: Testing Set - Prediction OACER – E-reader Tablet

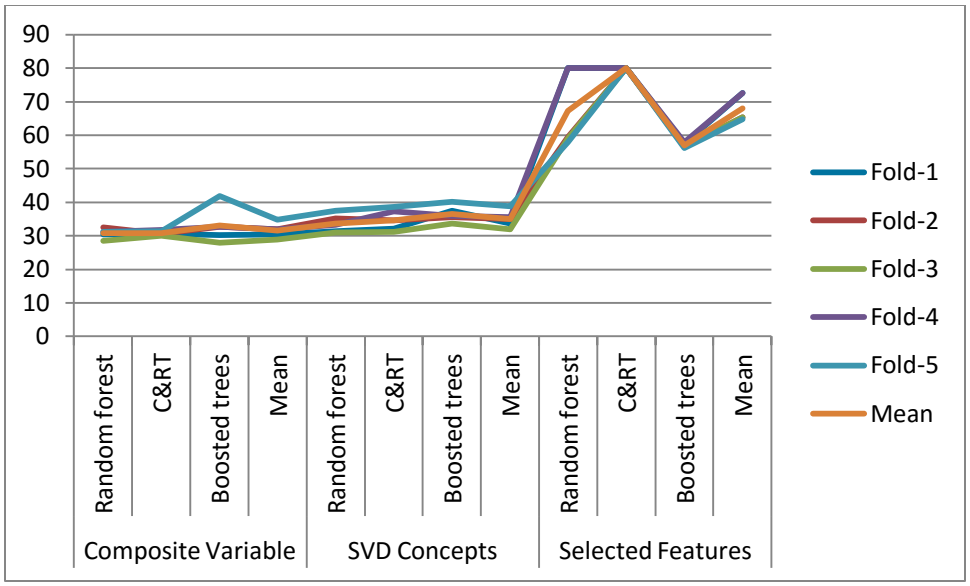


Figure 2.5: Testing Set - Prediction OACER – E-reader Tablet

Indep. Variable Set	Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Composite Variable	**Random forest**	30.5	35	37	31.5	31	33
	C&RT	30.5	35.5	35.5	30.5	30.5	32.5
	Boosted trees	36	43	35	28.5	30	34.5
	Mean	32.33	37.83	35.83	30.16	30.5	33.33
SVD Concepts	**Random forest**	36	38	38.5	34.5	37.5	36.9
	C&RT	38	38.5	49	39	60.5	45
	Boosted trees	45.5	42.5	42.5	51.5	54	47.2
	Mean	39.83	39.66	43.33	41.66	50.66	43.03
Selected Features	**Random forest**	78.5	60	80	54.5	80	70.6
	C&RT	80	80	80	80	80	80
	Boosted trees	54	58.5	44	54.5	54.5	53.1
	Mean	70.83	66.16	68	63	71.5	67.9

Table 2.5: Testing Set - Prediction OACER – Wireless Color Photo Printer

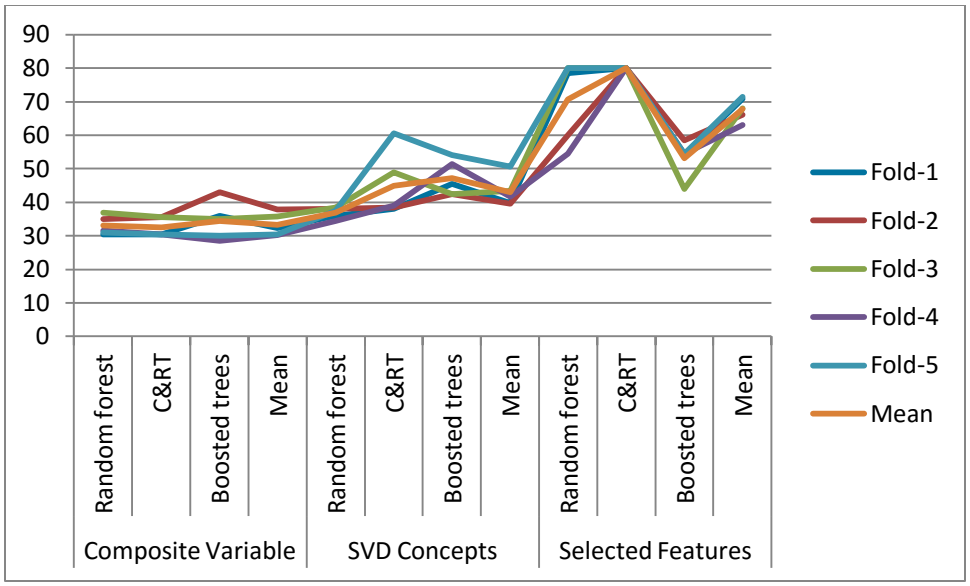


Figure 2.6: Testing Set - Prediction OACER – Wireless Color Photo Printer

Indep. Variable Set	Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Composite Variable	**Random forest**	18.38	25.16	17.25	19.19	20	20
	C&RT	22.09	24.83	17.09	19.19	25	21.64
	Boosted trees	21.61	24.35	23.22	24.19	24.83	23.64
	Mean	20.69	24.78	19.18	20.85	23.27	21.76
SVD Concepts	**Random forest**	39.67	38.22	32.9	35.8	33.38	35.99
	C&RT	36.93	38.54	45.16	39.19	46.12	41.18
	Boosted trees	29.67	32.9	35.96	32.25	32.25	32.6
	Mean	35.42	36.55	38	35.74	37.25	36.59
Selected Features	**Random forest**	54.51	53.87	57.41	56.77	54.83	55.48
	C&RT	60	60	60	80	80	68
	Boosted trees	49.51	50.32	53.54	52.9	52.74	51.8
	Mean	54.67	54.73	56.98	63.22	62.52	58.42

Table 2.6: Testing Set - Prediction OACER – Hotel Reviews

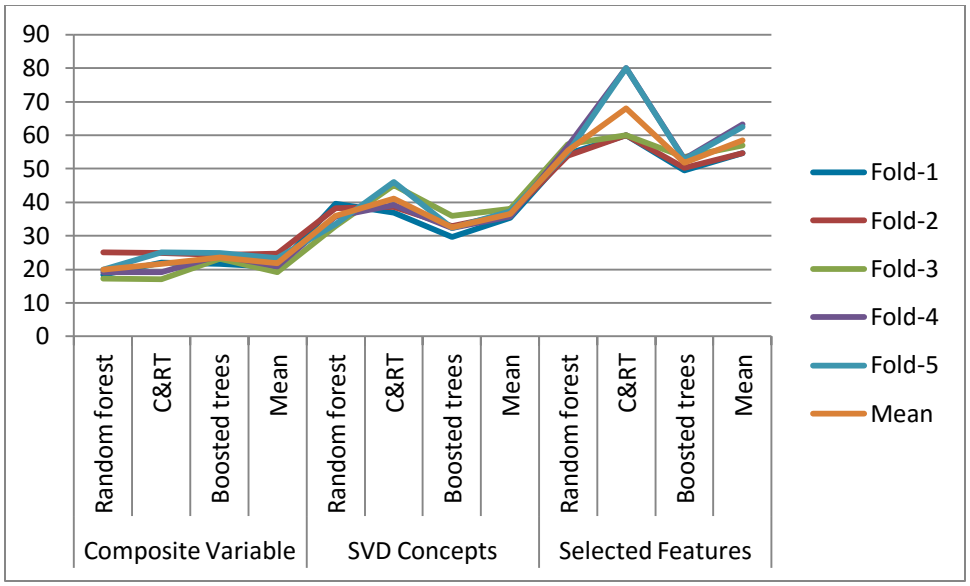


Figure 2.7: Testing Set - Prediction OACER – Hotel Reviews

In this study we use 3 different data sets from 3 different domains and at 3 different sizes. The data sets consist of 3200 hotel reviews, 2150 e-reader tablet reviews and 1000 wireless color photo printer reviews, along with their associated star rates. Since the models are built based on the training sets, the larger training data sets provide better trained models so that the models can make more accurate predictions on the testing sets. For showing this situation, we designed another experimental study. Since we have 1000 stratified reviews data for the wireless color photo printer, based on the random selections, we extracted 1000 stratified reviews data sets for the hotels and for the e-reader tablet separately from the existing data sets. After getting the new data sets for the hotels and e-reader tablet, the same modeling processes are repeated. The following experimental results prove the statement mentioned above; better predictive models can be produced with larger data sets.

		Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Printer	**Composite Variable**	**Random forest**	70	69.5	76	72.5	72	72
		C&RT	69.5	70.5	76	71	79.5	73.3
		Boosted trees	70	73.5	76	70	72	72.3

		Mean	69.83	71.16	76	71.16	74.5	**72.53**
	SVD Concepts	**Random forest**	75	69.5	79.5	74	75	74.6
		C&RT	75	76.5	80	75.5	80	77.4
		Boosted trees	73	79.5	76	78	80	77.3
		Mean	74.33	75.16	78.5	75.83	78.33	**76.43**
	Selected Features	**Random forest**	78.5	80	80	77.5	80	79.2
		C&RT	80	80	80	80	80	80
		Boosted trees	76.5	80	79	77.5	75.5	77.7
		Mean	78.33	80	79.66	78.33	78.5	**78.96**
E-Reader	**Composite Variable**	**Random forest**	65.5	74.5	77.5	72.5	72.5	72.5
		C&RT	68	74.5	79	71.5	73.5	73.3
		Boosted trees	69.5	73.5	74.5	73	73.5	72.8
		Mean	67.66	74.16	77	72.33	73.16	**72.86**
	SVD Concepts	**Random forest**	80	80	80	80	80	80
		C&RT	80	80	80	80	80	80
		Boosted trees	79.5	79.5	79.5	80	78.5	79.4
		Mean	79.83	79.83	79.83	80	79.5	**79.8**
	Selected Features	**Random forest**	80	74	80	77.5	74.5	77.2
		C&RT	76.5	75	78	76.5	78.5	76.9
		Boosted trees	79.5	74.5	77	74.5	78.5	76.8
		Mean	78.66	74.5	78.33	76.16	77.16	**76.96**
Hotels	**Composite Variable**	**Random forest**	58.5	66.5	67.5	65.5	61.5	63.9
		C&RT	63	66	66	65.5	63.5	64.8
		Boosted trees	60.5	68.5	65	65.5	60.5	64
		Mean	60.66	67	66.16	65.5	61.83	**64.23**
	SVD Concepts	**Random forest**	66.5	70	72.5	68.5	71.5	69.8
		C&RT	68.5	73.5	77.5	73.5	71.5	72.9
		Boosted trees	63	71.5	73	71	71	69.9
		Mean	66	71.66	74.33	71	71.33	**70.86**
	Selected Features	**Random forest**	80	77.5	77.5	78.5	80	78.7
		C&RT	80	80	80	80	80	80
		Boosted trees	76.5	76.5	78.5	76	79.5	77.4
		Mean	78.83	78	78.66	78.16	79.83	**78.7**

Table 2.7: Testing Set - Prediction EER – 1000 cases data sets

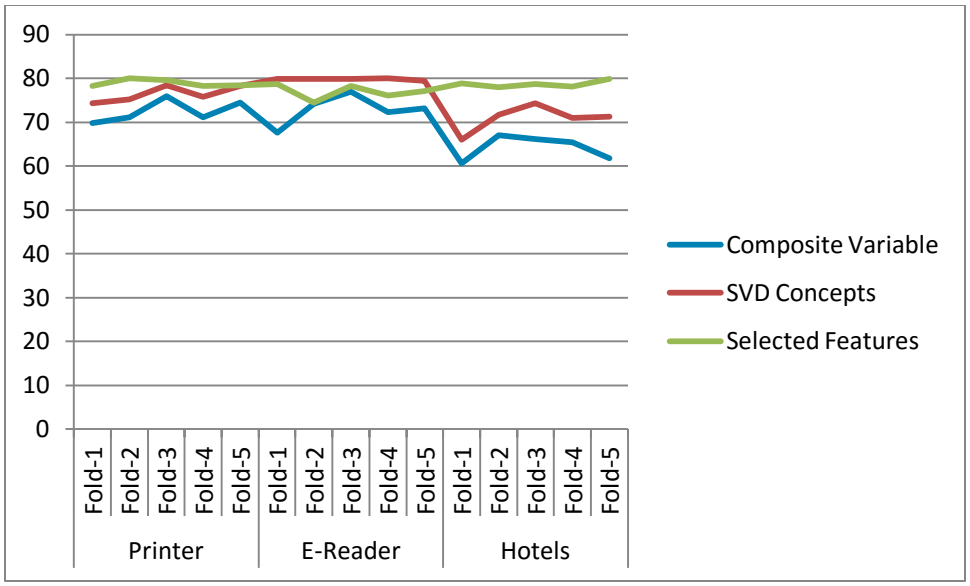


Figure 2.8: Testing Set - Prediction mean EER – 1000 cases data sets

		Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Printer	**Composite Variable**	**Random forest**	70	69.5	76	72.5	72	72
		C&RT	69.5	70.5	76	71	79.5	73.3
		Boosted trees	70	73.5	76	70	72	72.3
		Mean	69.83	71.16	76	71.16	74.5	72.53
	SVD Concepts	**Random forest**	75	69.5	79.5	74	75	74.6
		C&RT	75	76.5	80	75.5	80	77.4
		Boosted trees	73	79.5	76	78	80	77.3
		Mean	74.33	75.16	78.5	75.83	78.33	76.43
	Selected Features	**Random forest**	78.5	80	80	77.5	80	79.2
		C&RT	80	80	80	80	80	80
		Boosted trees	76.5	80	79	77.5	75.5	77.7
		Mean	78.33	80	79.66	78.33	78.5	78.964
E-Reader	**Composite Variable**	**Random forest**	67.91	73.72	67.44	70.7	68.6	69.674
		C&RT	69.3	74.42	69.07	70	69.07	70.372
		Boosted trees	71.16	73.26	70.47	72.56	71.4	71.77
		Mean	69.45	73.8	68.99	71.08	69.69	70.602
	SVD Concepts	**Random forest**	71.16	74.19	68.6	72.79	71.63	71.674
		C&RT	72.33	75.35	68.84	73.49	71.16	72.234
		Boosted trees	72.09	75.58	71.4	74.42	73.26	73.35
		Mean	71.86	75.04	69.61	73.56	72.01	72.416
	Selected Features	**Random forest**	80	79.77	79.07	80	80	79.768
		C&RT	80	80	80	80	80	80
		Boosted trees	79.07	78.84	79.3	80	79.77	79.396

		Mean	79.69	79.53	79.45	80	79.92	79.718
Hotels	**Composite Variable**	**Random forest**	49.84	50.81	46.94	45.81	49.84	48.648
		C&RT	49.52	51.45	46.61	45	50.65	48.646
		Boosted trees	53.23	50.65	45.81	43.23	50	48.584
		Mean	50.86	50.97	46.45	44.68	50.16	48.624
	SVD Concepts	**Random forest**	70.65	69.35	68.23	67.9	65.97	68.42
		C&RT	71.13	70.48	69.52	73.55	71.94	71.324
		Boosted trees	60.97	61.94	58.87	55.97	57.42	59.034
		Mean	67.58	67.25	65.54	65.8	65.11	66.256
	Selected Features	**Random forest**	76.13	76.61	78.06	77.9	77.58	77.256
		C&RT	80	80	80	80	80	80
		Boosted trees	75.32	74.84	75.32	75.65	77.42	75.71
		Mean	77.15	77.15	77.79	77.85	78.33	77.654

Table 2.8: Testing Set - Prediction EER – Original data sets

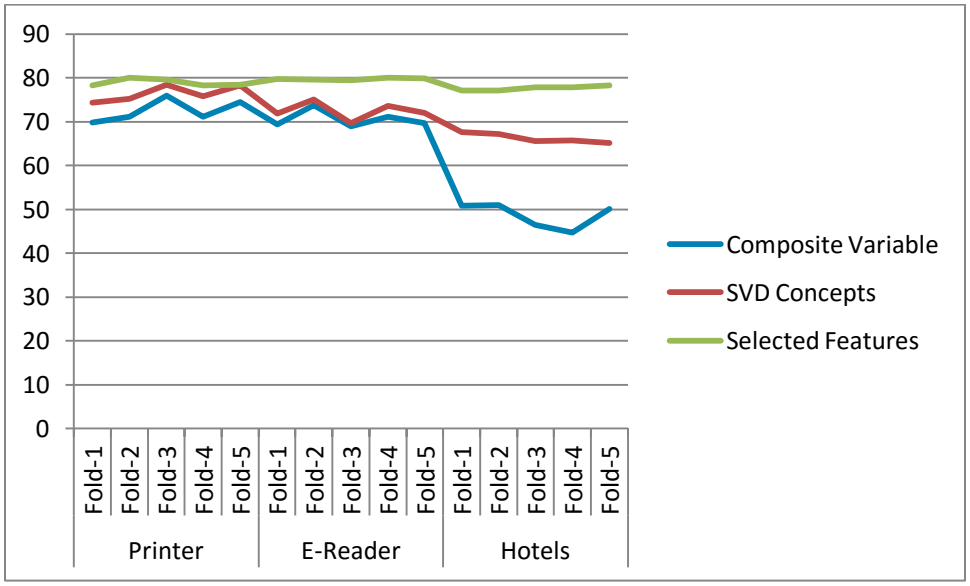


Figure 2.9: Testing Set - Prediction mean EER – Original data sets

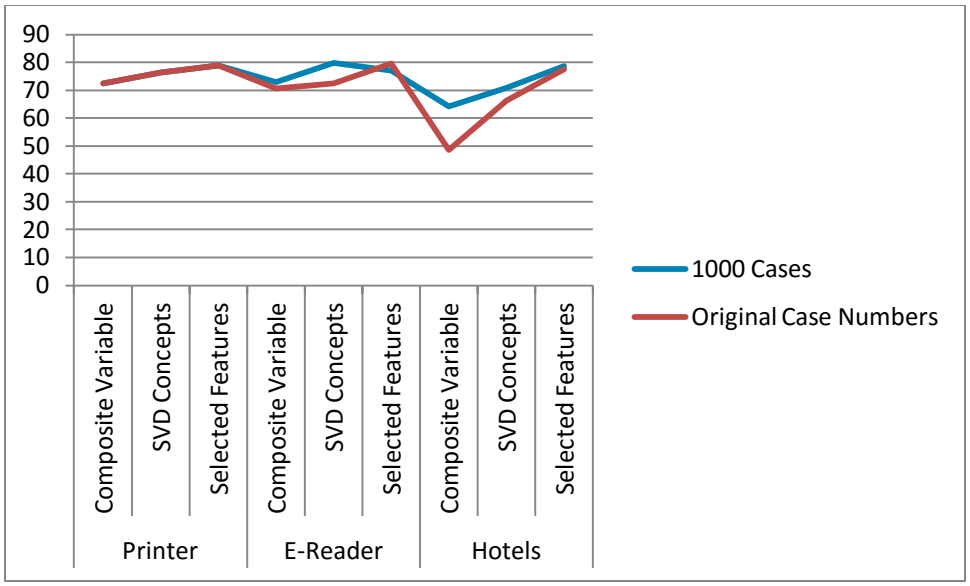


Figure 2.10: Testing Set - Prediction mean EER – Comparison

In each situation, composite variable performs better than the other predictor sets (Figure 2.8 and Figure 2.9). Also, when the sizes of e-reader and hotels data sets have decreased, the error rates of the models have increased (Figure 2.10). These experimental results prove that for better trained models we need larger training data sets.

2.12 Conclusion and Future Recommendation

The main goal of the proposed research methodology was to develop a composite variable that will perform very effective in machine learning based sentiment classification models to predict the star-rates that were assigned by the online reviewers. To achieve this goal, a composite variable is created by utilizing correlation scores and binary weights of the features (phrases), and for the comparison of the performance of the composite variable, SVD concepts are calculated and chi-square feature selection method is utilized to select the best set of predictors. Finally, by deploying separately these 3 sets of independent variables into tree-based machine learning algorithms by also incorporating 5-fold cross validation concept, the

independent variable sets are tested in the ML models. Also, for the convenience of the models, the same process is repeated for 3 different data sets. By analyzing the unstructured data through the methodology used in this study, the following research questions have been addressed:

1- Can ML models be utilized for the effective and accurate categorization of the unstructured data sets?
2- Every feature selection is an information loss at some level. Is it possible to create a composite variable in that all the extracted features would take place in some way, and it would prevent or minimize the information loss, and it could be used instead of them in the categorization of the data?
3- Each second very large dimensional text data sets are floating into the websites. Is it possible to categorize these textual data sets shared on the internet in an effective way by using ML models so that we can observe people's attributes for some specific subjects?

Therefore, the proposed method considers the tree-based machine learning algorithms to predict the star rates on a multinomial scale in three different data sets. Among these different conditions the composite variable has outperformed to SVD concepts and selected fetures in both of the two performance evaluation criteria. The findings, which have been obtained through such novel methodology, can provide valuable insights to the tourism and online retailer sectors for them to utilize in decision-making processes.

Chapter 3

Employing text analytics for automatic document classification: Hotel reviews as a case study

3.1 Employing text analytics for automatic document classification: Hotel reviews as a case study

Evaluating customer satisfaction is a key point to survive the high competition in the lodging industry [61]. Hotels employ different tools such as surveys to obtain customers' feedbacks about the services but customers are usually reluctant to give any comment to them [62]. Therefore hotels might be blind about their weakness and strength from the customer point of view. On the other hand, majority of people book hotels online [63]. People share their feedback on the hotels after their trips. These comments on the well-known hotel review platforms such as *TripAdvisor* are helpful and trustworthy for new customers to decide what lodging provider they choose [63-64]. Mauri and Minazzi [65] demonstrated a positive correlation between hotel purchasing intention and content of the online reviews. Several studies revealed a correlation between tourist satisfaction and desire to repeat the visit [66-68]. Hotel quality significantly affects tourist satisfaction [69] and itself is an important part of tourist destination management and marketing [70]. Therefore it is expected that hotels' service quality affects overall tourist industry's income in any tourist destinations. About 10% of global GDP and 10% of employment is related to tourism industry [71], therefore tourist industry is an important part of any sustainable economy. Alternatively, policy makers can perceive performance of hotels by analyzing the hotel reviews. The obtained information is beneficial for developing strategic roadmap of the tourist industry. Richter [72] demonstrated importance of

policy in the tourist industry and argued that this industry depends more on the administrative and political actions than business or economics.

Online hotel reviews usually consist of structured sections such as overall rating scale (e.g. 1-star to 5- stars) and an unstructured section containing customers' feedback about the hotel. Traditional statistical methods (e.g. descriptive statistics, regression, and etc.) are capable of investigating the structured part while text mining and NLP based models are required for analyzing the unstructured section. Sentiment analysis is developed for investigating texts' content and interpreting the sentimental orientation this technique is widely used in order to analyze unstructured web contents such as customers' reviews, news, weblogs and etc [73].

Regarding the methods utilized in sentiment classification, *semantic orientation (SO)* and *machine learning (ML)* are two major groups of sentiment analysis. The Semantic Orientation approach was developed in order to classify text corpuses into binary categories such as 'good' or 'bad'. This approach grammatically decomposes text to extract adverbs or adjectives, and then evaluates if they are associated with a 'good' or 'bad' meaning by utilizing domain-specific lexical resources, since a single word may have negative or positive connotations based on its text composition. In the later versions, connotations of consecutive words are taken into the account [7-8, 24, 74].

On the other hand, *ML* is a branch of *artificial intelligence (AI)* that is effectively used for data investigation in business analytics, health informatics and safety studies. These applications mainly appeared after 1970's when computers could perform fairly complex calculations [75]. Recently, IBM introduced the Watson Machine, which is a computer system that utilizes a complex ML algorithm to learn from text corpses, pictures etc. [76]. This machine demonstrated a wide variety of ML applications from winning the Jeopardy game to healthcare informatics.

Probabilistic association rules, classification and predictive modeling are major ML techniques in text mining [10, 12, 77]. ML learns from historical data, and then categorizes a text into multinomial classes. Therefore ML does not necessarily require a prior knowledge about any domain of data.

Semantic orientation on English texts requires development of a system that can recognize every single name, adverb, adjective and verb and delivers the adjectives' connotations. Therefore, practitioners utilize lexical resources such as WordNet® [19] for text mining. WordNet® is an English thesaurus that was developed at the Princeton University. The organization of WordNet is based on the lexical significances, rather than lexemes makes which make it different from a traditional thesaurus [78]. Soricut and Marcu [20] introduced a discourse parsing model that employs lexical and syntactic characteristics to discover structures of sentences. These models are able to recognize the discourse units and make the discourse parse-tree in sentence level analysis. In some early works such as Hatzivassiloglou and McKeown [6], practitioners applied a log-linear regression model with a preselected set of seed words to estimate the semantic orientation. Kamps et al. [21]. The second stream of sentiment classification methods is machine-learning. In several studies, machine-learning approaches outperform the semantic orientation approach. Pang and Lee [11] utilized maximum entropy classification, naïve Bayes, and support vector machines by considering bigrams, unigrams and a hybrid of them to categorize the movie reviews. They demonstrated that their proposed method outperform previous method in classifying unstructured web data. Chaovalit and Zhou [26] compared semantic orientation approach against machine-learning methods. The results revealed that machine-learning approaches yield a higher performance. In the following, a brief history on application of machine-learning approach in analyzing customers' reviews is presented.

Prabowo and Thelwall [28] investigated Social media (MySpace) movie reviews and product reviews by employing support vector machines, statistics based classifier (SBC), rule-based classifier (RBC) and general inquirer based classifier (GIBC). The results of the study show that a hybrid manner with a multiple classifiers has a better performance. Sarkar et al. [18] developed a novel approach for multinomial classification (i.e. extreme dislike, weak dislike, uncertain, weak like and extreme like) by combining k-means clustering algorithm (unsupervised method) and artificial neural networks (supervised learning algorithm). Travel Blogs are a kind of self-narrative diary that people explain their experiences during their travel adventures. Ye et al. [29] performed sentiment analysis on seven tourist destinations in Europe and USA by machine-learning techniques. In this study naïve Bayes (NB) model outperformed by two other supervised machine-learning algorithms of the character based n-gram model, and support vector machines. The accuracy of the mentioned method was over 80 percent. They also found a correlation between size of the training set and accuracy of the algorithms. Wang et al. [30] developed a novel latent rating regression (LRR) model to solve latent aspect rating analysis (LARA) for hotel reviews as the case study. They considered a set of the overall rating (measure from 1 star to 5 stars) and aspects of reviews to investigate relationships between reviewer's latent ratings on the different aspects. They demonstrated that their model can reveal different rating behavior. Bjørkelund et al. [31] compared performance of lexical-resource semantic orientation with machine-learning algorithms for sentiment analysis of hotel reviews with a multinomial scale (i.e. strong negative, weak negative, neutral, weak positive, and strong positive). They also performed spatial and temporal analysis on the reviews and visualized their data by using Google Maps. Current studies reveal that practitioners utilized lexical-based

resources, supervised learning and unsupervised learning algorithms for classifying customer reviews.

As can be understood from the existing relevant studies discussed in this section, the current study proposes a novel approach that creates three unique composite variables by employing a well-known clustering algorithm. These variables are then deployed into MLP based ANN model to classify the sentiment of the reviews (i.e. Barcelona, Istanbul and New York). The next section of this study discusses about the background information and the proposed methodology in detail.

3.2 Methodology

In this study, a hybrid sentiment classification methodology (as depicted in Figure 3.1) that is composed of four phases is proposed. The first phase of the study consists of four sequential steps. In the first step the dataset is split into *k* mutually exclusive folds in that *k-fold* cross validation concept will be employed in this study to decrease the uncertainty and thereby to increase the robustness of the classification algorithms employed in this study. In step 2, the text corpus is cleaned and artificial variables are embedded to each review in the dataset, to use in the following steps of the method. In the third step, feature extraction process is conducted and finally in the fourth step the binary-term document matrix is created. Phase 2 is divided into two parallel tasks such that each task represents different approach, which will be compared by using the prediction models. In phase 2.a, the most impactful variables are selected among the variable that were extracted in the previous phase, by employing Chi-square feature selection method. Alternatively in phase 2.b, k-means clustering algorithm by considering the correlations between the extracted features and the artificial variables that were created in Phase 1. In Phase 3, the potential predictors that were obtained by using two alternative approaches that have been

conducted in Phase 2.a and 2.b, are deployed into a powerful machine learning algorithm (MLP-based ANN), to classify the reviews based on the cities where the associated hotels are located. Finally in Phase 4, the multinomial classification results that were obtained through employing the two alternative approaches are compared. Additional detailed information on each of these phases and steps is provided in the following subsections.

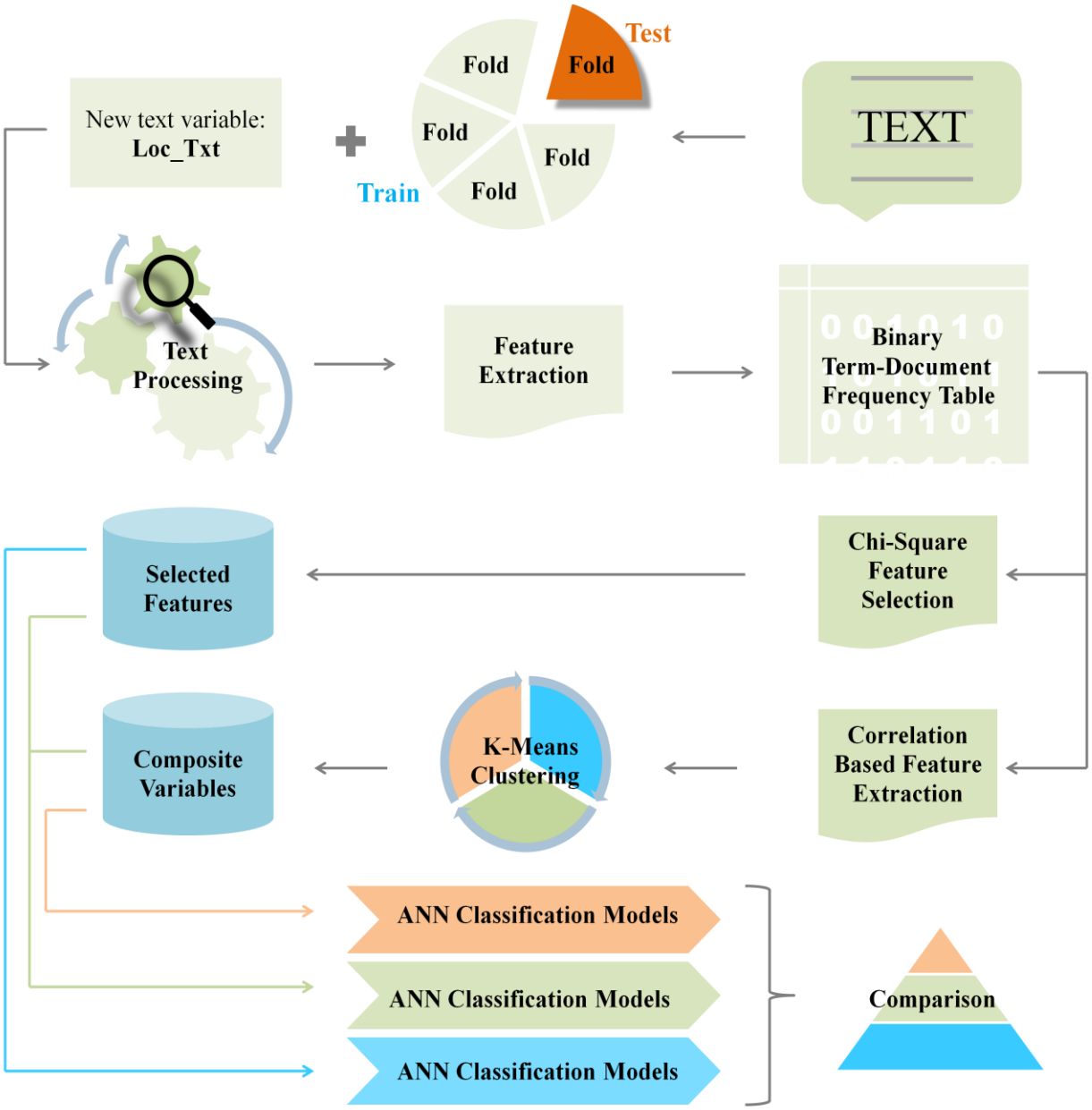


Figure 3.1: Overall Structure of the Proposed Method

3.3 Data Acquisition and Preparation

The customer reviews that were used in the current study was obtained from Tripadvisor.com, which is the world's largest travel site [32] that operates in 45 countries worldwide and has 340 million unique monthly visitors [33]. In this website, the customer reviews can be obtained from each hotel that are in their list. The reason why the users share their experiences with prospective travelers via such website is that they want to help to future travelers in making their decisions. Having the hotel name for all of the reviews (labeled data) gives us an advantage of being able to classify these reviews by using supervised learning (machine learning) algorithms in analyzing this large and complex unstructured data. In our current analysis, 1755 reviews have been collected about hotels that are located in New York (US), Barcelona (Spain) and Istanbul (Turkey) area. It should also be noted that, the dataset used in the current study is balanced such that the hotels that are selected for each city have equal amount from each star group. Specifically, there number of 3-star hotels selected from Istanbul, Barcelona and Newyork are equal. The rationale behind that is to minimize the biasness among the selected data points.

3.4 K-fold Cross Validation

The *k*-fold cross-validation approach is used to minimize the bias associated with the random sampling of the training and test data samples [34]. The entire dataset is randomly split into *k* mutually exclusive subsets of approximately equal size. The prediction model is tested *k* times by using the test sets. The estimation of the *k*-fold cross-validation for the overall performance criteria is calculated as the average of the *k* individual performances as follows [79]:

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \tag{3.1}$$

where *CV* stands for cross validation, *k* is the number of mutually exclusive subsets (folds) used, and *PM* is the performance measure used in the analysis. In this analysis, the stratified 5-fold cross validation approach is used to estimate the performance of the different classification models. The choice for *k*=5 is based on literature results [34, 79] which demonstrates that 5-folds provide an ideal balance between performance and the time required to run the folds.

3.5 Data Cleansing, Tokenization, Stemming, and Feature Extraction

Stop words that are frequently used in text files do not any meaningful value to the sentiment. Also, they may cause a heavier text corpus from the analytics perspective. Therefore, the stop words such as *the, a, is, at* etc. have been removed in our analysis. In addition, nonsense words as well as the typos have also been removed from that dataset used in the current study. Tokenization in text mining literature basically means to break a stream of text up into tokens. In such procedure, documents are broken into meaningful components such as words, phrases, sentences etc. The morphological roots or bases of the words can be identified by stemming. It is important to automatically identify the words that have same morphological structure. Also, these words should be treated accordingly since it would enable to reach more accurate prediction performances. Another critical component of sentiment analysis is *phrase definition*. It plays a crucial role since it helps discriminating the subtle structures that were caused by the change in the meaning of a phrase when the (immediate) neighbor words are taken into account. Also, longer phrases tend to be more informative in terms of identifying the polarity of the sentiments. To exemplify, while term “clean” is likely a positive sentiment, “not clean” or “not very clean” are less likely to appear in positive comments. For such reasons, models that employ

bag-of-unigrams or *bag-of-bigrams* would not be efficient in identifying the difference between the terms "*not clean*" and "*not very clean*". Also, the existing related literature [37-38] showed that "*sentiment classifiers combined with high order n-grams as features can achieve comparable, or better SA performance than state of the art on large-scale data sets*". In our proposed analysis, we utilize high order n-grams in classifying reviews.

3.6 K-means Clustering Algorithm

In the existing text mining literature, there are various types of document clustering techniques. Each one of these techniques uses a different similarity measure to represent the clusters. The main idea in the clustering is to classify the documents that have similar attributes or characteristics into the same group. The clustering approaches can be represented under three main titles such as a) text-based, b) link-based, and c) hybrid. Among these, text-based clustering works with similar idea used in libraries; the contents of the documents are used as the similarity measure and documents having similar contents are put in the same cluster, whereas the goal in link-based clustering is to group linked pages or documents and keep the link-path. Hybrid clustering is the combination of these two clustering techniques. The first one (text-based clustering) can be classified according to the clustering algorithms into three categories such as partitional, hierarchical, and graphical based clustering algorithms. In the current study, *k-means* algorithm has been employed, which can be considered as one of the most popular partitional clustering algorithm.

The main idea in k-means clustering is to group the observations into k clusters such that the observations in the same cluster have the maximum similarity with each other, and maximum diversity with the observations in the other clusters. General steps of the k-means algorithm can be understood in six consecutive steps such as;

1) Randomly assign k cluster centers.

2) Calculate the distance of the points to the each center (Euclidean distance is used in our study).

3) Assigning all the points into clusters (in the sense of minimum distance to the centers).

4) Calculate the new centers of the clusters.

5) Recalculate the distance each point to the new centers.

6) Repeat the steps 3, 4 and 5 until all the points stays stable on its previous place.

Let X be the set of all observations and C be the initial cluster centers. Then the k-means algorithm mainly based on the following formula:

$$\sum_{x \in X} \min_{c \in C} \sum_{i=1}^{k} \left| x_i - c_i \right|^2 \qquad (3.2)$$

In other words, *k-means* will create k clusters such that they have maximum variability with each other and minimum variability within the clusters. The best *k* value can be decided based in trial and error experiments.

3.7 Creating Composite Variables

Composite variable creating is a very common technique in data mining literature. In text mining composite variable creating is used to assign numerical weights for a document. In this study we use correlation values and k-means clustering algorithm for the creation of the composite variables. First of all, at the beginning of the process we add an artificial text variable into the data set. This variable is called as Loc_Txt and has 3 categories; Loc_NY_, Loc_BARC_, and Loc_IST_. We put them together into the text processing and produce a single binary term-document frequency matrix from them. Since the categories of Loc_Txt_ are all in

text format, they also appear on the matrix as terms. Since each one is written in a unique format, the frequency score for the corresponding city is 1 and for the other two cities it is 0. In the second step, we calculate the correlation between the terms Loc_NY_, Loc_BARC_, Loc_IST_ and the rest of the terms. This correlation table has the information of how the terms are scattered around the city names and so very informative. After then, by using the correlation matrix, we cluster al the terms in 3 clusters. As we mention above, k-means algorithm let us to define the number of clusters. We defined the number of clusters as 3 in the consideration of 3 cities. In other words, we grouped the terms around the city names. In the final step, for each document, we summed the frequencies of the terms taking place in the same cluster. These 3 summations create 3 new variables which are the composite variables that will be used in the ANN model for the estimation of the city categories.

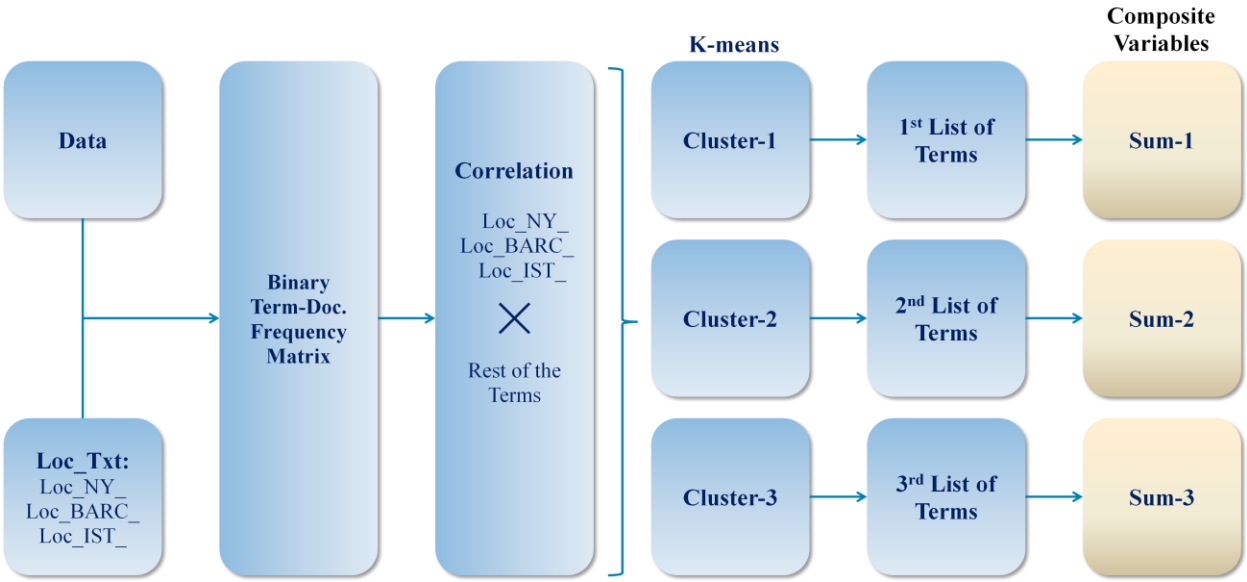
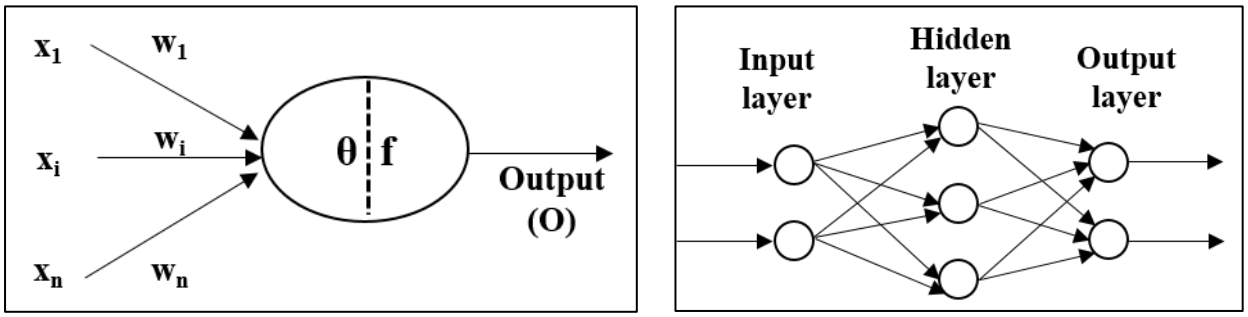


Figure 3.2: The algorithm for the creation of the composite variables

3.8 Multi-Layer Perceptron-based Artificial Neural Network (MLP-ANN)

ANNs are widely employed in a wide variety of computational data analytics problems that include classification, regression and pattern recognition [29-31]. A generic ANN is a computational system that consists of "a highly interconnected set of processing elements, called *neurons*, which process information as a response to external stimuli. An artificial neuron is a simplistic representation that emulates the signal integration and threshold firing behavior of biological neurons by means of mathematical equations" ([32], P.4). The information flow among each *neuron* takes place in an input-output manner, as shown in Figure 3.3.



(a) (b)

Figure 3.3: A Schematic of Information Flow for a) a Single Neuron, b) a Multilayered ANN

Based on figure 3.3-(a), the inputs received by a neuron can be represented by the input vector **x=** (x_1, x_2, … , x_n) where x_i is the information from the i^{th} input (i=1, 2, … , n), and weights connected to a neuron can be modeled via a weight vector **w=** (w_1, w_2, … , w_n), where w_i is the weight associated with i^{th} neuron. The neuron's signal output flow *0* can be calculated using:

$$O = f\left(\sum_{i=1}^{n} w_i.x_i\right), \tag{3.3}$$

where f represents an activation function of the weighted sum of the inputs associated with the n incoming signals. In our analysis, the sigmoid function is employed as the activation function for the ANN model. The function is represented in Eq. (3.4):

$$f(x) = \frac{1}{1+e^{-x}}. \tag{3.4}$$

Then, the output is converted to a binary representation through the following transformation:

$$O = \begin{cases} 1 & if \left(\sum_{i=1}^{n} w_i . x_i \right) \geq \theta \\ 0 & otherwise \end{cases}, \tag{3.5}$$

where θ denotes a threshold value that regulates the neuron's action potential "… by modulating the response of that neuron to a particular stimulus confining such response to a pre-defined range values" Sordo [32].

The Multi-layer Perceptron (MLP) learning model with a back-propagation algorithm has been employed in our ANN model due to its superior performance to the radial basis function (RBF) in the preliminary analysis. In a MLP network, there is an input, an output and a hidden layer (as depicted by figure 3.3-(b)). There are typically no restrictions on the number of hidden layers. The external stimuli is received by the input layer and propagated to the next layer. The weighted sum of incoming signals sent by the input units are received by the hidden layer(s) and processed by means of the sigmoid activation function in Eq. (3.4). Similarly, the units in the output layer receive and process the weighted sum of incoming signals using the activation function. Thus, the system propagates the information forward until the current output is

obtained for each existing input in the system. An error measure can be obtained by calculating the difference between the target value and the current output, as shown in Eq. (3.6). Finally, the system adjusts the current weights until the current output is satisfactory enough or the performance of the system cannot be improved any further.

$$E = \frac{1}{2}\sum_{j}(t_{pj} - o_{pj})^2 . \tag{3.6}$$

3.9 Feature Selection

The ultimate goal of feature selection is to select the most important set of feature among the entire feature set. In other words, instead of having a set of complex classifiers, a smaller subset is selected to simplify the models employed as well as to minimize the computation time. For an appropriate reliable and efficient text classification, there certain criteria should be met such as an appropriate data structure (to represent the text data), appropriate objective functions (avoid overfitting issue) and appropriate algorithms (to deal with the high dimensional matrices without losing so much information) etc. For such reasons, feature selection is one of the most crucial and challenging steps in that the variables that will be used as the predictors are selected through this process. By doing do, it simplifies and speeds up the run process of the learning algorithms. Such process is only applied on the features extracted from the train set [50].

3.9.1 Chi-Square Test based Feature Selection (CSFS)

In the field of Statistics, chi square is a well-known method that is used to check the independence of two events. Also, it is an important feature selection technique, which have been commonly used in the text mining literature. Therefore in our study, instead of events, the occurrences of the terms are checked and ranked with respect to the following measurement:

$$X^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \quad (3.7)$$

where D is document, e_t and e_c (term t and class c) are defined as e_t = 1 (if term t exists) and e_c = 0 (if t does not exist), N is the observed frequency in D and E is the expected frequency.

Therefore, chi-square measure represents the deviation between the expected E and observed N. If χ^2 value is larger than the threshold value, the hypothesis of independence is incorrect. In other words, the counts of expected and observed are close. Dependency of the two events indicates that the occurrence of the term is related with the occurrence of the class, and therefore it is an important feature for the class.

	Index	Net. name	Training perf.	Test perf.	Training algorithm	Error function	Hidden activation	Output activation
Fold-1	1	MLP 430-22-3	75.17730	75.65217	BFGS 15	Entropy	Exponential	Softmax
	2	MLP 430-8-3	86.31206	76.52174	BFGS 47	SOS	Tanh	Identity
	3	MLP 430-16-3	87.87234	77.39130	BFGS 27	Entropy	Tanh	Softmax
	4	MLP 430-14-3	81.84397	75.94203	BFGS 23	SOS	Logistic	Tanh
	5	MLP 430-10-3	85.88652	77.68116	BFGS 29	SOS	Logistic	Tanh
	6	MLP 430-16-3	78.43972	76.52174	BFGS 16	Entropy	Tanh	Softmax
	7	MLP 430-22-3	74.11348	76.52174	BFGS 43	SOS	Exponential	Exponential
	8	MLP 430-19-3	83.40426	73.04348	BFGS 43	SOS	Tanh	Exponential
	9	MLP 430-14-3	73.12057	75.65217	BFGS 12	Entropy	Exponential	Softmax
	10	MLP 430-15-3	82.83688	77.68116	BFGS 25	SOS	Exponential	Logistic
Fold-2	1	MLP 414-17-3	74.25532	69.27536	BFGS 24	Entropy	Exponential	Softmax
	2	MLP 414-11-3	76.59574	71.88406	BFGS 30	SOS	Logistic	Exponential
	3	MLP 414-15-3	73.75887	66.37681	BFGS 14	Entropy	Exponential	Softmax
	4	MLP 414-11-3	82.90780	73.04348	BFGS 24	SOS	Logistic	Logistic
	5	MLP 414-21-3	79.85816	74.78261	BFGS 26	SOS	Exponential	Identity
	6	MLP 414-12-3	84.25532	73.04348	BFGS 35	SOS	Tanh	Identity
	7	MLP 414-25-3	82.41135	76.81159	BFGS 92	SOS	Exponential	Tanh
	8	MLP 414-8-3	82.05674	74.20290	BFGS 16	Entropy	Tanh	Softmax
	9	MLP 414-19-3	82.41135	75.07246	BFGS 27	SOS	Tanh	Identity
	10	MLP 414-24-3	83.19149	75.36232	BFGS 27	SOS	Logistic	Tanh
Fold-3	1	MLP 424-12-3	78.43972	75.94203	BFGS 34	SOS	Exponential	Identity
	2	MLP 424-15-3	75.39007	75.36232	BFGS 38	SOS	Exponential	Exponential

	3	MLP 424-21-3	82.26950	77.97101	BFGS 19	Entropy	Logistic	Softmax
	4	MLP 424-10-3	84.68085	77.97101	BFGS 62	SOS	Identity	Tanh
	5	MLP 424-17-3	82.55319	76.23188	BFGS 30	SOS	Identity	Identity
	6	MLP 424-14-3	84.96454	77.97101	BFGS 46	SOS	Exponential	Tanh
	7	MLP 424-18-3	81.84397	77.97101	BFGS 24	Entropy	Logistic	Softmax
	8	MLP 424-25-3	82.05674	78.84058	BFGS 53	SOS	Logistic	Logistic
	9	MLP 424-25-3	81.98582	78.55072	BFGS 43	Entropy	Logistic	Softmax
	10	MLP 424-19-3	73.40426	73.04348	BFGS 10	Entropy	Exponential	Softmax
Fold-4	1	MLP 442-22-3	82.05674	74.49275	BFGS 14	Entropy	Logistic	Softmax
	2	MLP 442-20-3	77.80142	74.49275	BFGS 46	SOS	Exponential	Logistic
	3	MLP 442-19-3	74.53901	70.14493	BFGS 26	SOS	Exponential	Logistic
	4	MLP 442-25-3	82.12766	74.78261	BFGS 23	Entropy	Identity	Softmax
	5	MLP 442-15-3	70.42553	65.21739	BFGS 14	Entropy	Exponential	Softmax
	6	MLP 442-11-3	78.65248	75.65217	BFGS 19	SOS	Exponential	Tanh
	7	MLP 442-25-3	81.41844	76.23188	BFGS 20	Entropy	Tanh	Softmax
	8	MLP 442-12-3	71.70213	69.27536	BFGS 10	SOS	Exponential	Logistic
	9	MLP 442-21-3	84.89362	74.20290	BFGS 28	SOS	Logistic	Identity
	10	MLP 442-16-3	75.03546	68.69565	BFGS 17	Entropy	Exponential	Softmax
Fold-5	1	MLP 400-16-3	84.85507	72.53333	BFGS 29	SOS	Logistic	Logistic
	2	MLP 400-12-3	80.21739	71.46667	BFGS 69	SOS	Exponential	Exponential
	3	MLP 400-19-3	75.94203	69.06667	BFGS 16	Entropy	Exponential	Softmax
	4	MLP 400-16-3	70.65217	66.93333	BFGS 15	Entropy	Exponential	Softmax
	5	MLP 400-24-3	83.84058	75.46667	BFGS 16	Entropy	Logistic	Softmax
	6	MLP 400-17-3	74.13043	68.53333	BFGS 10	Entropy	Exponential	Softmax
	7	MLP 400-11-3	83.76812	74.66667	BFGS 22	Entropy	Tanh	Softmax
	8	MLP 400-23-3	75.86957	71.46667	BFGS 18	Entropy	Exponential	Softmax
	9	MLP 400-15-3	83.04348	74.93333	BFGS 20	Entropy	Tanh	Softmax
	10	MLP 400-24-3	81.95652	74.93333	BFGS 18	Entropy	Tanh	Softmax

Table 3.1: Neural network classification models based on the selected features

By using the formula given in the literature review part, we calculate chi-square value of each feature and then the values are sorted from the largest to the smallest. The dependent variable of the feature selection model is City_Category variable. City_Category consists of 3 categories based on the 3 cities. All the features having P-value<.01 are treated as significant and put into the selected features set. To show and validate the ability of the selected features to classify the dependent variable, we are going to utilize ANN classification models.

Since we have a binary table, all the selected features are categorical. The features are used as the independent variables (predictors), and City_Category as the dependent variable in the ANN model. By using the Train_Test variable, we define the train and test sets. An ANN has many parameters such as number of hidden layers, error function, hidden activation function, output activation function and so on. We run the model for 10 different combinations of the parameters. You can find the results on the Table 3.1 for each fold.

3.9.2 Correlation based feature extraction (Creating composite variables)

In this study we do not use a regular correlation feature selection. Indeed, the technique we apply here is composite features creating technique by using all the extracted features and using them as if they are selected features. As we mention this technique is based on correlation. Since we want to find the best features classifying the documents by city, we are going to use the correlation between the features and the cities. However, the correlations between the features and 'City' variable do not provide the correlation for each city separately. When we grouped the documents by 'City', we would not get any correlation because city category would be the same. For overcoming this problem, we added a new variable to the train set called 'Loc_Text'. This variable consists of the corresponding city names in a unique format: loc_ny_, loc_barc_, and loc_ist_. After merging the 'Customer Reviews' with 'Loc_Text', put them together into the text processing so that we have the new format city names in the term-document table. Since we use a unique format for each city, the correlation between the unique city names and the other terms provides very important information for the seeing the scatter of terms around each city. For example, if term x has larger correlation with loc_ny_ than it has with loc_barc_ and loc_ist_, then this means that the occurrence of x in the reviews for the hotels located in New York is larger its occurrence in the others. After this step, we use the correlation table for the clustering

to group all the terms around the city names. K-means clustering algorithm gives us an opportunity for defining the number of clusters. By considering the number of city categories, we defined the number of clusters as 3. Finally, since in a review, the increase of occurrence of the terms from the same cluster increases the possibility that review is from the city corresponding the cluster, we are going to sum the terms in the same cluster. These summations create 3 new continuous composite variables. We are going to use them as selected features and put them into the ANN classification model. Again, we run the ANN models for 10 different combinations of the parameters. You can find the results in the Table 3.2 for each fold.

	Index	Net. name	Training perf.	Test perf.	Training algorithm	Error function	Hidden activation	Output activation
Fold-1	1	MLP 3-5-3	90.69597	66.56977	BFGS 37	Entropy	Exponential	Softmax
	2	MLP 3-10-3	88.71795	65.98837	BFGS 15	Entropy	Exponential	Softmax
	3	MLP 3-9-3	90.03663	66.56977	BFGS 19	SOS	Identity	Identity
	4	MLP 3-6-3	91.06227	66.27907	BFGS 42	Entropy	Tanh	Softmax
	5	MLP 3-4-3	90.25641	66.86047	BFGS 97	SOS	Exponential	Exponential
	6	MLP 3-7-3	90.47619	66.56977	BFGS 28	SOS	Identity	Logistic
	7	MLP 3-9-3	91.06227	65.69767	BFGS 28	Entropy	Identity	Softmax
	8	MLP 3-8-3	90.03663	66.86047	BFGS 28	SOS	Logistic	Identity
	9	MLP 3-7-3	90.69597	66.56977	BFGS 54	SOS	Exponential	Tanh
	10	MLP 3-9-3	90.62271	65.98837	BFGS 19	SOS	Identity	Logistic
Fold-2	1	MLP 3-8-3	91.83526	68.14159	BFGS 84	SOS	Exponential	Tanh
	2	MLP 3-7-3	91.83526	69.02655	BFGS 61	SOS	Exponential	Tanh
	3	MLP 3-4-3	91.83526	68.43658	BFGS 32	SOS	Identity	Tanh
	4	MLP 3-4-3	91.97977	69.91150	BFGS 144	SOS	Exponential	Tanh
	5	MLP 3-5-3	91.32948	69.32153	BFGS 30	SOS	Logistic	Identity
	6	MLP 3-9-3	89.45087	69.32153	BFGS 28	SOS	Logistic	Identity
	7	MLP 3-6-3	89.23410	68.73156	BFGS 16	SOS	Exponential	Tanh
	8	MLP 3-4-3	92.26879	67.25664	BFGS 26	Entropy	Identity	Softmax
	9	MLP 3-10-3	91.90751	67.84661	BFGS 22	SOS	Exponential	Logistic
	10	MLP 3-8-3	91.61850	68.14159	BFGS 26	Entropy	Identity	Softmax
Fold-3	1	MLP 3-9-3	92.94545	65.38462	BFGS 64	SOS	Exponential	Tanh
	2	MLP 3-3-3	92.65455	65.68047	BFGS 37	SOS	Exponential	Exponential
	3	MLP 3-8-3	93.01818	64.49704	BFGS 32	Entropy	Tanh	Softmax
	4	MLP 3-3-3	93.16364	65.68047	BFGS 26	Entropy	Logistic	Softmax

	5	MLP 3-9-3	93.89091	66.27219	BFGS 75	Entropy	Logistic	Softmax
	6	MLP 3-10-3	93.09091	66.86391	BFGS 25	Entropy	Logistic	Softmax
	7	MLP 3-8-3	91.70909	65.68047	BFGS 11	Entropy	Identity	Softmax
	8	MLP 3-3-3	92.94545	66.27219	BFGS 36	SOS	Exponential	Exponential
	9	MLP 3-9-3	93.09091	65.97633	BFGS 52	Entropy	Logistic	Softmax
	10	MLP 3-4-3	92.80000	64.20118	BFGS 27	SOS	Tanh	Exponential
Fold-4	1	MLP 3-3-3	91.49856	64.58333	BFGS 24	Entropy	Logistic	Softmax
	2	MLP 3-8-3	90.99424	63.09524	BFGS 15	SOS	Logistic	Logistic
	3	MLP 3-10-3	91.21037	65.17857	BFGS 35	Entropy	Logistic	Softmax
	4	MLP 3-6-3	91.42651	65.17857	BFGS 36	SOS	Tanh	Logistic
	5	MLP 3-10-3	91.49856	64.58333	BFGS 30	Entropy	Identity	Softmax
	6	MLP 3-4-3	90.99424	63.69048	BFGS 22	Entropy	Identity	Softmax
	7	MLP 3-5-3	90.85014	66.66667	BFGS 37	SOS	Logistic	Tanh
	8	MLP 3-10-3	91.21037	63.69048	BFGS 27	Entropy	Identity	Softmax
	9	MLP 3-3-3	90.20173	65.77381	BFGS 22	Entropy	Logistic	Softmax
	10	MLP 3-7-3	90.63401	66.36905	BFGS 41	SOS	Logistic	Exponential
Fold-5	1	MLP 3-7-3	91.38827	68.57923	BFGS 51	SOS	Exponential	Logistic
	2	MLP 3-6-3	90.94284	68.03279	BFGS 27	Entropy	Identity	Softmax
	3	MLP 3-8-3	91.75947	68.85246	BFGS 36	Entropy	Tanh	Softmax
	4	MLP 3-10-3	91.09131	68.85246	BFGS 19	SOS	Tanh	Logistic
	5	MLP 3-10-3	91.98218	69.94536	BFGS 55	SOS	Logistic	Logistic
	6	MLP 3-3-3	91.16555	68.30601	BFGS 24	Entropy	Identity	Softmax
	7	MLP 3-7-3	91.31403	68.57923	BFGS 42	SOS	Identity	Tanh
	8	MLP 3-6-3	90.86860	68.57923	BFGS 18	Entropy	Tanh	Softmax
	9	MLP 3-5-3	89.60653	66.66667	BFGS 19	Entropy	Exponential	Softmax
	10	MLP 3-3-3	91.16555	68.30601	BFGS 23	Entropy	Exponential	Softmax

Table 3.2: Neural network classification models based on the composite variables

3.10 Combination of the features

In this step we bring together the features selected by the chi-square measure and features obtained in the correlation based process, and then for 10 different combinations of parameters we repeat the same ANN process. You can find the results for each fold on the Table 3.3.

	Index	**Net. name**	**Training perf.**	**Test perf.**	**Training algorithm**	**Error function**	**Hidden activation**	**Output activation**
Fold-1	1	MLP 433-18-3	91.72161	79.94186	BFGS 23	SOS	Identity	Logistic

	2	MLP 433-19-3	88.86447	77.90698	BFGS 24	Entropy	Tanh	Softmax
	3	MLP 433-14-3	88.42491	79.94186	BFGS 23	SOS	Identity	Logistic
	4	MLP 433-16-3	91.86813	78.77907	BFGS 25	Entropy	Identity	Softmax
	5	MLP 433-25-3	89.45055	79.94186	BFGS 93	SOS	Logistic	Logistic
	6	MLP 433-8-3	95.09158	79.06977	BFGS 29	Entropy	Tanh	Softmax
	7	MLP 433-14-3	87.47253	79.65116	BFGS 53	SOS	Exponential	Tanh
	8	MLP 433-11-3	84.68864	74.70930	BFGS 102	SOS	Exponential	Exponential
	9	MLP 433-22-3	91.06227	77.90698	BFGS 38	SOS	Logistic	Identity
	10	MLP 433-23-3	86.81319	77.61628	BFGS 24	Entropy	Tanh	Softmax
Fold-2	1	MLP 417-18-3	89.16185	75.51622	BFGS 25	Entropy	Identity	Softmax
	2	MLP 417-9-3	91.61850	76.99115	BFGS 32	Entropy	Identity	Softmax
	3	MLP 417-8-3	86.99422	76.69617	BFGS 43	SOS	Exponential	Tanh
	4	MLP 417-10-3	97.03757	77.58112	BFGS 57	SOS	Logistic	Tanh
	5	MLP 417-18-3	94.43642	77.28614	BFGS 43	SOS	Identity	Tanh
	6	MLP 417-13-3	96.38728	76.69617	BFGS 58	SOS	Tanh	Identity
	7	MLP 417-21-3	95.52023	76.99115	BFGS 35	Entropy	Logistic	Softmax
	8	MLP 417-22-3	94.43642	79.64602	BFGS 37	SOS	Identity	Tanh
	9	MLP 417-15-3	94.72543	74.63127	BFGS 48	SOS	Logistic	Exponential
	10	MLP 417-24-3	94.79769	74.92625	BFGS 28	Entropy	Logistic	Softmax
Fold-3	1	MLP 427-10-3	93.09091	80.76923	BFGS 39	SOS	Logistic	Tanh
	2	MLP 427-24-3	96.43636	81.65680	BFGS 213	SOS	Identity	Logistic
	3	MLP 427-8-3	84.80000	78.69822	BFGS 32	SOS	Exponential	Logistic
	4	MLP 427-14-3	85.52727	79.28994	BFGS 18	Entropy	Tanh	Softmax
	5	MLP 427-10-3	97.45455	80.47337	BFGS 54	SOS	Identity	Logistic
	6	MLP 427-12-3	95.34545	81.06509	BFGS 28	Entropy	Identity	Softmax
	7	MLP 427-12-3	97.45455	80.76923	BFGS 42	Entropy	Logistic	Softmax
	8	MLP 427-25-3	93.23636	79.28994	BFGS 129	SOS	Logistic	Logistic
	9	MLP 427-25-3	92.65455	80.17751	BFGS 125	SOS	Logistic	Logistic
	10	MLP 427-11-3	95.85455	81.65680	BFGS 25	Entropy	Tanh	Softmax
Fold-4	1	MLP 445-25-3	86.31124	75.00000	BFGS 55	Entropy	Logistic	Softmax
	2	MLP 445-24-3	86.31124	75.59524	BFGS 51	Entropy	Exponential	Softmax
	3	MLP 445-20-3	93.87608	74.10714	BFGS 32	SOS	Logistic	Logistic
	4	MLP 445-16-3	93.29971	76.78571	BFGS 64	SOS	Identity	Exponential
	5	MLP 445-10-3	88.25648	72.91667	BFGS 21	Entropy	Tanh	Softmax
	6	MLP 445-12-3	94.16427	76.78571	BFGS 44	SOS	Exponential	Identity
	7	MLP 445-20-3	94.66859	72.91667	BFGS 27	Entropy	Tanh	Softmax
	8	MLP 445-9-3	94.09222	73.21429	BFGS 34	Entropy	Tanh	Softmax
	9	MLP 445-24-3	86.45533	74.40476	BFGS 50	SOS	Logistic	Logistic
	10	MLP 445-18-3	95.74928	75.29762	BFGS 23	Entropy	Logistic	Softmax
Fold-5	1	MLP 403-16-3	92.57610	77.59563	BFGS 44	SOS	Tanh	Identity
	2	MLP 403-15-3	98.36674	77.04918	BFGS 77	Entropy	Tanh	Softmax

3	MLP 403-11-3	97.77283	75.95628	BFGS 42	Entropy	Tanh	Softmax
4	MLP 403-20-3	94.35783	78.41530	BFGS 97	SOS	Identity	Exponential
5	MLP 403-24-3	94.20935	78.41530	BFGS 125	SOS	Identity	Exponential
6	MLP 403-12-3	94.13512	78.68852	BFGS 180	SOS	Identity	Exponential
7	MLP 403-15-3	98.81218	76.50273	BFGS 40	Entropy	Tanh	Softmax
8	MLP 403-8-3	94.72903	77.04918	BFGS 39	Entropy	Logistic	Softmax
9	MLP 403-17-3	97.77283	78.14208	BFGS 49	Entropy	Identity	Softmax
10	MLP 403-20-3	97.47587	80.05464	BFGS 69	SOS	Logistic	Tanh

Table 3.3: Neural network classification models based on the union of the sets of independent variables

3.11 Results

Both feature selection models are set on the train set. So we can say that the ANN models based on the correlation based features fit better to the data then the ANN models based on the chi-square based features. However, when we check the performance on the models on the test set, this time the ANN models based on the chi-square based features give better results. On the other hand, how long a statistical model's running process takes is an important aspect in assessing the model. To finish the run of a model having millions of cases and thousands of variables may take days. From this point of view, correlation based models are really time-friendly because chi-square based models have 211 input features in average, but correlation based models have only 3 input features in average.

In the third step we bring together all the features selected from the two models and put them all in the ANN models. This time the success of correlation based features on the train and the success of chi-square based features on the test contribute together and we get better results for both train and test. This time we have models that are both fitting to the train very good and have really high classification performance on the test.

	Correlation	**Chi-Square**	**All**

	Training perf.	Test perf.	Training perf.	Test perf.	Training perf.	Test perf.
Fold-1 Mean	90.36630	66.39535	80.90071	76.26087	89.54579	78.54651
Fold-2 Mean	91.32948	68.61357	80.17021	72.98551	93.51156	76.69617
Fold-3 Mean	92.93091	65.65089	80.75887	76.98551	93.18545	80.38462
Fold-4 Mean	91.05187	64.88095	77.86525	72.31884	91.31844	74.70238
Fold-5 Mean	91.12843	68.46995	79.42754	72.00000	96.02079	77.78689
Overall Mean	91.36140	66.80214	79.82451	74.11014	92.71641	77.62331

Table 3.4: The comparison of the neural network classification models' performances

Chapter 4

Text Mining Algorithm: Grading Written Exam Papers

4.1 Text Mining Algorithm: Grading Written Exam Papers

In this study we develop a Turkish text mining algorithm which is grading written exam papers automatically via text mining techniques. The general idea of predictive model studies is to build a model on the training set and then apply it on the testing set. We have used a similar idea for this study. We have built the algorithm on the answer key prepared by the grader and then applied it on the answer papers of the students. And so, the grades have been recalculated by the algorithm and then compared with the real grades. The comparison is the main measurement for the indication of success of the algorithm.

For this study, we have used the exam papers of the literature course of the Ankara Demetevler Anadolu Imam Hatip Lisesi located in Turkey. The class that the exam papers were obtained consists of 21 students. For the privacy, the students' name/id/class number/term will

not be shared. The student id numbers given in the extracted data are not the real student id numbers, they were assigned arbitrarily by the researcher.

The main idea in this study is to build a text mining algorithm in Turkish which is going to grade exam papers in Turkish. So we need to prepare dictionaries for the categories. The categories list consists of the answers, sub-answers of the answers and sections of the sub-answers. In the final list we get 15 categories.

For each category we prepare a dictionary separately. A dictionary consists of synonym words list, homonym words list, words list from the same stem, all the possible suffix combinations of the words in Turkish, and all the possible combinations of the phrases in Turkish word order based on the Turkish grammar. The details are given in the further parts.

Before starting creation of dictionaries, first we should know the grading technique of the grader. Some graders just focus on the correct answer. Even an answer has some incorrect parts, if the requested answer is given, the student gets a full credit for the question. However, some graders cut off grades for the incorrect parts of the answer even if the student give and correct answer in the full answer.

The first grading technique which is only focusing on the correct answers is also the grading technique used for the exam papers in the data set. In this technique corresponding positive grades are assigned for all the student answers taking place in the dictionary but for the student answers that are not taking place in the dictionary, no action is required. However, in the second technique, while giving the corresponding positive grades for the answers taking place in the dictionary, negative grades are assigned for the student answers that are not taking place in the dictionary.

One of the most challenging issues in this study is the grader's personal notions during the grading. Even if some student answers do not match the answers given in the answer key, the grader may give positive grades for his/her personal notion. It is almost impossible to build an algorithm for calculating the grades given by the grader's personal notion.

In the creation of the algorithm we have some challenges. The most important challenges are given as following:

C1: Incorrect answer with correct words.

C2: Correct answer with incorrect words.

C3: Words having the same stem but different meanings with different suffixes.

C4: Answers with non-synonym words taking place in the answer key.

C5: Answers with synonym words but not taking place in the answer key.

C6: Homonym words.

C7: Open-ended questions.

For the each challenge we develop a special method which is overcoming the problem.

4.2 Turkish Grammar

Turkish is an agglutinative language and each word can have very different suffixes defining the meaning and tense. **Suffixes** are added to a **stem**. There are two different suffixes in Turkish: Constructive and inflectional suffixes. Constructive suffixes are used to create a new word from an existent word. Inflectional suffixes are used to define the word's grammatical role in the sentence. The inflectional suffixes –ler/-lar are used to make the words plural. Also there are some irregular plural words without having these two suffixes. For example, halk (toplum)/people. In Turkish, the third person singular pronoun is "o" and it is corresponding to

she/he/it in English. In other words, in Turkish there is no third person singular pronoun to define the gender of the subject [80].

As we mention above, Turkish is an agglutinative language and new words are created by adding suffixes to a word stem. For understanding how to add a new suffix to word stem, one should know the Turkish vowel and consonant harmony rules of Turkish. If we know the Turkish vowels and consonants harmony rules, by using the almost all the possible suffixes [81], we can get all the possible stem-suffix matching.

In a Turkish sentence there are nine parts:

1. noun
2. pronoun
3. adjective
4. verb
5. interjection
6. adverb
7. postposition
8. particle
9. conjunction

Suffixes define the word's tense and its grammatical meaning in the sentence. For example;

Git/Go

Git**ti**/Went

Düş**mek**/To fall/Verb

Düş**ünce**/Consideration/Noun

Therefore, during the creation of dictionary, for the each stem extracted from the answer key, all the possible stem-suffix combinations should be considered.

4.2.1 Turkish Characters

The Turkish Alphabet does not include Q, X, W letters and it includes Ç, Ğ, İ, Ö, Ş, Ü. This is the only difference between English and Turkish alphabets. The rest of the alphabets are exactly the same. In the text mining algorithm's natural language processing step, this difference should be defined and all the non-Turkish characters should be filtered from the documents.

4.2.2 Word orders in Turkish phrases

In a Turkish sentence, the regular word order is that (adjectives and adverbs) the modifier precedes (noun and verb) the modified. The most general order of a Turkish transitive sentence is subject-object-verb or the other five combinations of them. In Turkish, the word order is very important because it may change the meaning of the sentence, moreover even the sentence may have completely the opposite meaning. The following sentences show how the order changes meaning of the sentence.

Yanlis, ifade dogru degil/No, expression is not true

Dogru, ifade yanlis degil/True,expression is not false

Ifade yanlis, dogru degil/Expression is false, not true

Ifade dogru, yanlis degil/Expression is true, not false

4.2.3 Turkish vowel and consonant harmony chain rules

In Turkish, the suffixes are added to a word stem based on the harmony chain rules. The harmony chain rules are branching off 1. The vowel harmony chain rule, and 2. The consonant harmony chain rule.

The vowel harmony chain rule:

There are eight vowels in Turkish alphabet: A E O Ö U Ü I İ. The first vowel harmony rule is that in Turkish a vowel following another vowel is not allowed. There are some irregular words like saat/watch, and they should be listed in a different category. For the creation a new word from the existing word stem, one should follow two certain vowel harmony rules [82-83]:

a. *The hard vowel harmony chain*:

The hard vowels are A O U I. The general idea in the hard vowel harmony chain is that a word stem having the hard vowels can be extended with suffixes having hard vowels.

Ok**u**l-d**a**n/From school

K**a**p**ı**-y**a**/To door

There are some irregular word formats that are not appropriate to the harmony rule.

S**aa**t-l**e**r/Watches

R**o**l-d**e**n/From role

b. *The thin vowel harmony chain*:

The thin vowels are E Ö Ü İ. The general idea in the thin vowel harmony chain is that a word stem having the thin vowels can be extended with suffixes having thin vowels.

K**i**m-**e**/To whom

G**e**l-m**e**/Do not come

There are some irregular word formats that are not appropriate to the harmony rule.

M**e**sl**e**k-t**a**ş/Colleague

Y**a**r**ı**n-k**i**/Of tomorrow

The consonant harmony chain rule:

In addition to the vowel harmony rules, there are some consonant harmony rules which are affecting the way suffixes are attached to the word stem. There are two consonant harmony cases: 1- The last consonant of the stem, 2- The first consonant of the suffix. The most challenging consonants in this case are p, ç, t and k. For a word stem which is ending with a consonant, the suffix may start with a vowel or consonant. If the stem has only one vowel/syllable and the suffix start with a vowel, then the stem usually does not change; Suç/Suça-Crime/To crime. However, if the stem has more than one vowel/syllable, and the suffix ends one the consonants p, ç, t, and k, then usually the consonant changes; Araç/Araca-Gadget/To gadget, Arap/Arabın-Arab/Of Arab, Kağıt/Kağıda-Paper/To paper, Açık/Açığa-Open/To open.

P becomes B,

Ç becomes C,

T becomes D,

K becomes Ğ.

There are some exceptions that should be listed in a different dictionary for the algorithm; Aç/Açık-Open/Open.

If the stem ends one of the consonants p, ç, t, k, f, h, s, and ş, then the suffix should start with one of c or d. In this case, the first consonant letter of the suffix changes [84].

C becomes Ç,

D becomes T.

Yap/Yaptı (Yapdı)-Do/Did.

4.3 Stemming

Stemming is the process of reducing extracted words to their stem by removing the suffixes from the original words. A stem is the smallest meaningful part of the words and does not have to have the same dictionary mean with the origin word. There are many computer based stemming algorithms. SAS Enterprise miner, Statistica, R statistical packages have text mining tools and stemming algorithm. Also, there are some online stemmer tools. For example, NLTK 2.0.4 **stem** package is very popular stemmer and the NLTK stemmer supports the following languages [85]: Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Porter, Portuguese, Romanian, Russian, Spanish, Swedish.

Unfortunately, there is no significant stemmer in Turkish Language. So here we develop a stemmer in Turkish. The stemming algorithm is working based on the following steps:

1- Tokenization: All the punctuations are removed and the text is broken into tokens. Here each token is a single word.

2- All the different words are listed.

3- Words having the same stem are gathered in the same cluster and the cluster is labeled with the stem.

Göz-Göz**lük**-Göz**lükçü**-Göz**lükçülük** + Göz**e**-Göz**lüklü**-Göz**lükçüye**-Göz**lükçülükde** + …

Ayak-Ayak**kabı**-Ayak**kabıcı**-Ayak**kabıcılık** + Ayağ**a**-Ayak**kabılı**-Ayak**kabıcıya**-Ayak**kabıcılıkda** + …

4- For the each stem, all the possible variation taking place in the official Big Turkish Dictionary (BTD) published by the Turkish Language Association (TLA) are added into the cluster.

4.3.1 Synonyms

Synonym words/phrases have exactly or nearly the same means and synonym words/phrases are called as synonymous. For example begin and start words are synonymous. Synonymous words can be any part of a sentence such as nouns, adverbs, verbs, adjectives or prepositions and they need not to have the same meaning in different sentences.

Sanat/Eser-Art

Halk/Toplum-People

Düsünce/Fikir-Idea

Bilgilendirmek/Haberdar etmek-To inform

Nesilden nesile akmak/Gelecek nesillere aktarmak-From generation to generation

Cosku veren/Duygulari kamçilayan-Dramatic

Zaman alir/Bir anda olmaz/Yavas yavas olur-In the course of time

After splitting a sentence into tokens, for each token (word/phrase) all the existing synonyms and related words are obtained from the TLA-BTD and for each one of the obtained synonym and related word/phrase, all the possible different suffix combinations are obtained by using the 'almost all Turkish suffixes' list.

Example 4.1: Tek dişi kalmış canavar/Single-fanged monster

Tek-tekim-tekimiz-tekin-tekiniz-tekler-teklerimiz-tekleriniz-teki-…

bir-biri-bire-birimiz-biriniz-birinin-birisi-birin-birine-birini-birimi-birisini-birisine-…

sadece-sade-sadem-saden-sademin-sadenin-sademiz-sadenizin-sademi-sadeni-sadenin-…

yalnız-yalnızca-yalnıza-yalnızsın-yalnızım-yalnızız-yalnızsınız-yalnızlar-yalın-yalınca-…

biricik-biri-bir-biricigim-biricigin-birinin-birimin-birli-birine-…

yegane-yeganem-yeganemiz-yeganen-yeganeniz-yeganemiz-yeganemizin-yeganemin-…

Dişi-dişim-dişimin-dişimiz-dişimizin-dişlerimiz-dişlerimizin-dişleriniz-dişleri-dişlerin-dişte-…

Çark-çarkı-çarkın-çarkını-çarkının-çarkına-çarka-çarketme-çarketmek-çarketmem-çarketmez-…

Testere-testeresi-testeresini-testereli-testeremi-testereme-testeresine-testeresiz-…

Tarak-taraklı-taraksız-tarakla-taraklamak-taraksızca-tarağı-tarağım-tarağın-…

Kalmış-kalmışım-kalmışsın-kalmışız-kalmışsınız-kalmışlar-kalmışsa-kalmışsanız-kalmışa-…

Koruma-korumak-korumalı-korumada-korumaksa-korumakla-korumadı-…

Sürdürme-sürdürmek-sürdürmeli-sürdürmedi-sürdürmedim-sürdürmedin-sürdürmemeli-…

Canavar-Canavara-Canavarı-Canavarın-Canavarca-Canavarda-Canavarmı-Canavardan-…

Hayvan-hayvanı-hayvanın-hayvanca-hayvanda-hayvana-hayvansa-hayvanmı-hayvansız-…

Köpek-köpeksi-köpekce-köpekli-köpeğin-köpeğe-köpekten-köpekde-köpeksiz-köpeğine-...

Kurt-kurtu-kurtlu-kurtla-kurtun-kurtta-kurttaki-kurtcu-kurtca-kurtum-kurttan-…

Acımasız-acımasızsın-acımasızsan-acımasızca-acımasızmı-acımasızım-acımasızın-…

Kötü ruhlu-kötü ruhlunun-kötü ruhluyu-kötü ruhlusu-kötünün ruhlu-kötünün ruhluna-…

Zalim-zalime-zalimin-zalimce-zalimde-zalimine-zalimane-zalimdi-zalimse-…

Çirkin-çirkinse-çirkinsen-çirkinsem-çirkinseniz-çirkinsiniz-çirkinmi-çirkince-…

In Turkish some words can have different means in different sentences even if they have the same stem and the same suffixes. Such words are called as homonym words.

Ayvalı reçeli çok severim/I like quinces jar so much/Fruit

Bu yaz **Ayvalı**'ya gidecegim/I will go to the Ayvalı (quinces) this summer/City name

This is an important challenge for the algorithm. For overcoming this issue, we need a sub-algorithm. The general steps of the sub-algorithm are as following:

i. First of all, we need to know what is the dictionary definition of the word in the sentence.

ii. If the word has more than one meaning then we need to define which one was used in the sentence. In other words, if the word has homonymous words in the answer key, then we need to define what are the corresponding meanings of the words in the dictionary, separately.

iii. We need to prepare dictionaries for each meaning of the word, separately.

For defining the meaning of the word in a phrase, we need to get the all synonymous of the words taking place in the phrase and we need to get all the meaningful combinations of the words. The combinations that are not defined in the TLA BTD are eliminated from the list and then the remaining list is accepted as the definition/meaning of the word. This process is similar to the process given in *example 4.1*.

If the word takes place multiple times in the document, the same lists are created for the other phrases that the word takes place. In the final step, if the intersection of word lists is empty then they are homonymous. The following example gives the overall steps of this process.

Example 4.2:

a. [1]Sulu şaka/prank

Şaka-şakacı-şakalı-şakam-şakan-şakaca-şakacık-şakana-şakama-şakanın-şakamın-şakaya-…

Latife-latifeli-latifem-latifen-latifesi-latifene-latifeme-latifeye-latifede-latifeden-latifemi-…

Koşuk-koşuğa-koşuğu-koşuğun-koşukda-koşukta-koşuksa-koşuk-koşuksu-koşukçu-…

Güldürü-güldürücü-güldürünün-güldürülü-güldürüyor-güldürür-güldürüş-güldürüp-…

Komiklik-komikti-komikdi-komikçe-komikmi-komiksin-komiğim-komikse-komiksen-…

Takılmak-takılmalı-takılması-takılmaya-takılmana-takılmam-takılmanı-takılmada-…

Güldüren-güldüreni-güldürenci-güldürenin-güldürenim-güldürene-güldürense-…

b. [2]Sulu yemek/Watery Food

Yemek-yemekli-yemekte-yemekde-yemekmi-yemekse-yemeği-yemeğim-yemeğin-…

Aş-aşa-aşı-aşım-aşın-aşçı-aşçın-aşçım-aşıma-aşına-aşının-aşımın-aşda-aşta-…

Çorba-çorbacı-çorbam-çorban-çorbanda-çorbamda-çorbamın-çorbama-çorbana-çorbası-…

Gıda-gıdacı-gıdam-gıdan-gıdama-gıdana-gıdanın-gıdamın-gıdalı-gıdaya-gıdada-…

Yiyecek-yiyecekle-yiyeceği-yiyeceğim-yiyeceğin-yiyeceğine-yiyeceğinin-yiyecekde-…

Taam-taama-taamı-taamın-taamım-taamda-taamana-taamının-taamının-taamsız-…

Besin-besine-besini-besinin-besinine-besinde-besinsiz-besinim-besinime-besininiz-…

Kahvaltı-kahvaltılı-kahvaltılık-kahvaltıda-kahvaltıya-kahvaltısı-kahvaltım-kahvaltın-…

Atıştırma-atıştırmak-atıştırmalı-atıştırmalık-atıştırmaya-atıştırmam-atıştırmadı-…

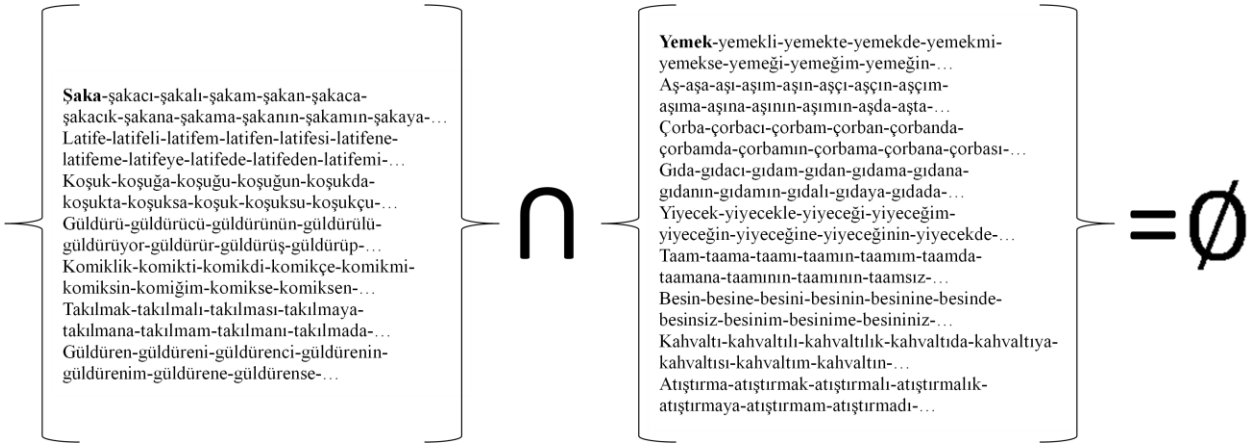


Figure 4.1: Intersection of the stemming lists

Then,

$^{1}\text{Sulu} \neq {}^{2}\text{Sulu}$

And so, we need to prepare two different synonym lists for the two ‘sulu’ homonym words. This process can be repeated for the all homonyms of ‘sulu’ and this whole process should be repeated for the all words in the dictionaries.

In Turkish some words have even the same stems, they have different meanings with different suffixes.

Gözlemek/**Göz**lük = To wait/Glasses

Gözlemek
Beklemek
İzlemek
Gözetlemek
Bakmak
Tarassut
Tecessüs
Korumak
Kollamak
İntizar
Gözlük
Gözene
Gözecik
Çerçeve
Siper
Nazar boncuğu
Tohum
İnce bulgur

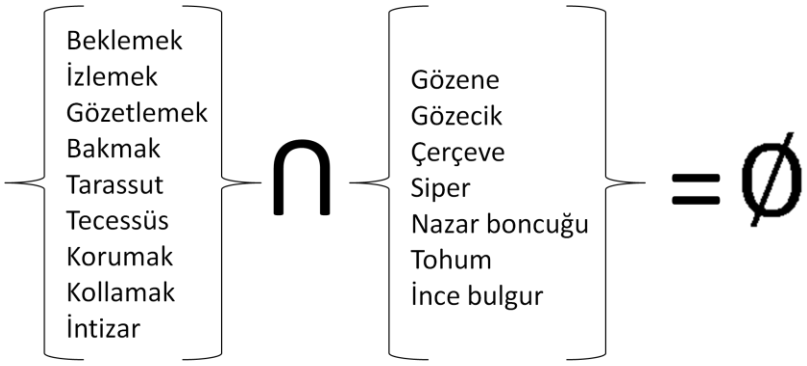


Figure 4.2: Intersection of the synonyms and related words lists

For identifying such situations, we need to create two different dictionaries for the two words. First we get all the possible phrase combinations from the original phrases. And then we get all the synonym and homonym words of the neighbor words located in the original phrases. Before passing to the final step, we need to produce new lists consist of the related words to the original words. The words in the related words list do not have to have the same stem or to be synonyms, but they can be used in the related subjects to the original one. For example; Yaz-kış-sonbahar-ilkbahar-mevsimler/Summer-winter-spring-fall-seasons, these words are related to each other and so can be used in the similar topics.

After getting the two related words lists for the two words, the original words are replaced by the related ones and then the existence of the phrases are checked in the TLA BTD. Undefined combinations are deleted from the list. And thus the dictionaries are ready for the two original words. Now if the original word stem is not in the intersection of the word combinations for the two dictionaries, then it means that these words have different means even if they have the same stem, and so we need to prepare two different stem clusters for them.

5- Each stem cluster is filed in a separated text file and separately uploaded to the text miner tool and since all the different versions of the stem with different suffixes are taking place in the list, all the versions are forwarded to the stem and so the tool will automatically match all the different versions with the stem and they all will be presented under a single column in the term-document frequency table.

Of course the algorithm is based on the data set gathered for this study and so it is very limited comparing to the overall Turkish language. But even developing a stemmer for such a small data is very time consuming. However the presented algorithm can be generalized for the whole Turkish language.

düs
düsün
düsünce
düsünce
düsünce
düsünceler
düsüncelerini

***

düsüncesini
güzel
güzellik
güzellikler
güzelliklerden

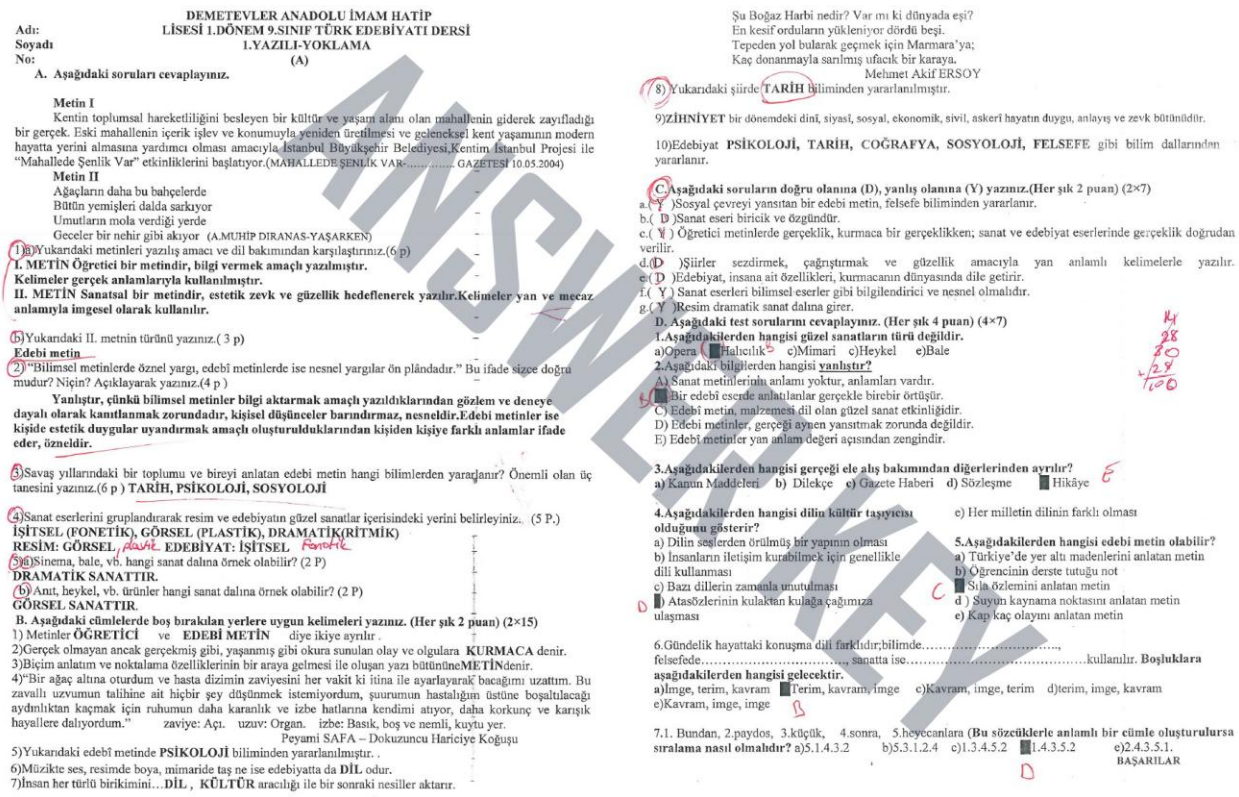
***

his
hissi
hissiyat

After getting the word stem lists for each word separately, the same process is repeated for the phrases, but this time the existent word stems are used.

özelliklerindendir
en önem özel
en önemli özellik
en önemli özelliklerindendir
en önemli ortak
en önem parça
en önemli parça
en önemli parçasi
en önemli parçasidir
en önem var
en önemli varlik
en önemli varligi
en önemli varligidir

4.4 Case Study



DEMETEVLER ANADOLU İMAM HATİP
LİSESİ 1.DÖNEM 9.SINIF TÜRK EDEBİYATI DERSİ
1.YAZILI-YOKLAMA
(A)

Adı:
Soyadı
No:

A. Aşağıdaki soruları cevaplayınız.

Metin I
Kentin toplumsal hareketliliğini besleyen bir kültür ve yaşam alanı olan mahallenin giderek zayıfladığı bir gerçek. Eski mahallenin içerik işlev ve konumuyla yeniden üretilmesi ve gelenceksel kent yaşamının modern hayatta yerini almasına yardımcı olması amacıyla İstanbul Büyükşehir Belediyesi,Kentim İstanbul Projesi ile "Mahallede Şenlik Var" etkinliklerini başlatıyor.(MAHALLEDE ŞENLİK VAR-..... GAZETESİ 10.05.2004)

Metin II
Ağaçların daha bu bahçelerde
Bütün yemişleri dalda sarkıyor
Umutların mola verdiği yerde
Geceler bir nehir gibi akıyor (A.MUHİP DIRANAS-YAŞARKEN)

1)a)Yukarıdaki metinleri yazılış amacı ve dil bakımından karşılaştırınız.(6 p)
I. METİN Öğretici bir metindir, bilgi vermek amaçlı yazılmıştır. Kelimeler gerçek anlamlarıyla kullanılmıştır.
II. METİN Sanatsal bir metindir, estetik zevk ve güzellik hedeflenerek yazılır.Kelimeler yan ve mecaz anlamıyla imgesel olarak kullanılır.

b)Yukarıdaki II. metnin türünü yazınız.(3 p)
Edebi metin

2) "Bilimsel metinlerde öznel yargı, edebî metinlerde ise nesnel yargılar ön plândadır." Bu ifade sizce doğru mudur? Niçin? Açıklayarak yazınız.(4 p)
Yanlıştır, çünkü bilimsel metinler bilgi aktarmak amaçlı yazıldıklarından gözlem ve deneye dayalı olarak kanıtlanmak zorundadır, kişisel düşünceler barındırmaz, nesneldir.Edebi metinler ise kişide estetik duygular uyandırmak amaçlı oluşturulduklarından kişiden kişiye farklı anlamlar ifade eder, öznеldir.

3)Savaş yıllarındaki bir toplumu ve bireyi anlatan edebi metin hangi bilimlerden yararlanır? Önemli olan üç tanesini yazınız.(6 p) **TARİH, PSİKOLOJİ, SOSYOLOJİ**

4)Sanat eserlerini gruplandırarak resim ve edebiyatın güzel sanatlar içerisindeki yerini belirleyiniz. (5 P.)
İŞİTSEL (FONETİK), GÖRSEL (PLASTİK), DRAMATİK(RİTMİK)
RESİM: GÖRSEL , EDEBİYAT: İŞİTSEL

5)6)Sinema, bale, vb. hangi sanat dalına örnek olabilir? (2 P)
DRAMATİK SANATTIR.
b) Anıt, heykel, vb. ürünler hangi sanat dalına örnek olabilir? (2 P)
GÖRSEL SANATTIR.

B. Aşağıdaki cümlelerde boş bırakılan yerlere uygun kelimeleri yazınız. (Her şık 2 puan) (2×15)
1) Metinler **ÖĞRETİCİ** ve **EDEBİ METİN** diye ikiye ayrılır .
2)Gerçek olmayan ancak gerçekmiş gibi, yaşanmış gibi okura sunulan olay ve olgulara **KURMACA** denir.
3)Biçim anlatım ve noktalama özelliklerinin bir araya gelmesi ile oluşan yazı bütününe**METİN**denir.
4)"Bir ağaç altına oturdum ve hasta dizimin zaviyesini her vakit ki itina ile ayarlayarak bacağımı uzattım. Bu zavallı uzvumun talihine ait hiçbir şey düşünmek istemiyordum, şuurumun hastalığım üstüne boşaltılacağı aydınlıktan kaçmak için ruhumun daha karanlık ve izbe hatlarına kendimi atıyor, daha korkunç ve karışık hayallere dalıyordum." zaviye: Açı. uzuv: Organ. izbe: Basık, boş ve nemli, kuytu yer.
Peyami SAFA – Dokuzuncu Hariciye Koğuşu
5)Yukarıdaki edebî metinde **PSİKOLOJİ** biliminden yararlanılmıştır. .
6)Müzikte ses, resimde boya, mimaride taş ne ise edebiyatta da **DİL** odur.
7)İnsan her türlü birikimini…**DİL , KÜLTÜR** aracılığı ile bir sonraki nesiller aktarır.

Şu Boğaz Harbi nedir? Var mı ki dünyada eşi?
En kesif orduların yükleniyor dördü beşi.
Tepeden yol bularak geçmek için Marmara'ya;
Kaç donanmayla sarılmış ufacık bir karaya.
Mehmet Akif ERSOY

8) Yukarıdaki şiirde **TARİH** biliminden yararlanılmıştır.

9)**ZİHNİYET** bir dönemdeki dinî, siyasî, sosyal, ekonomik, sivil, askerî hayatın duygu, anlayış ve zevk bütünüdür.

10)Edebiyat **PSİKOLOJİ, TARİH, COĞRAFYA, SOSYOLOJİ, FELSEFE** gibi bilim dallarından yararlanır.

C. Aşağıdaki soruların doğru olanına (D), yanlış olanına (Y) yazınız.(Her şık 2 puan) (2×7)
a.(Y)Sosyal çevreyi yansıtan bir edebi metin, felsefe biliminden yararlanır.
b.(D)Sanat eseri biricik ve özgündür.
c.(Y) Öğretici metinlerde gerçeklik, kurmaca bir gerçeklikken; sanat ve edebiyat eserlerinde gerçeklik doğrudan verilir.
d.(D)Şiirler sezdirmek, çağrıştırmak ve güzellik amacıyla yan anlamlı kelimelerle yazılır.
e.(D)Edebiyat, insana ait özellikleri, kurmacanın dünyasında dile getirir.
f.(Y) Sanat eserleri bilimsel eserler gibi bilgilendirici ve nesnel olmalıdır.
g.(Y)Resim dramatik sanat dalına girer.

D. Aşağıdaki test sorularını cevaplayınız. (Her şık 4 puan) (4×7)
1.Aşağıdakilerden hangisi güzel sanatların türü değildir.
a)Opera b)Halıcılık c)Mimari c)Heykel e)Bale
2.Aşağıdaki bilgilerden hangisi yanlıştır?
A) Sanat metinlerinin anlamı yoktur, anlamları vardır.
B) Bir edebî eserde anlatılanlar gerçekle birebir örtüşür.
C) Edebî metin, malzemesi dil olan güzel sanat etkinliğidir.
D) Edebi metinler, gerçeği aynen yansıtmak zorunda değildir.
E) Edebî metinler yan anlam değeri açısından zengindir.

3.Aşağıdakilerden hangisi gerçeği ele alış bakımından diğerlerinden ayrılır?
a) Kanun Maddeleri b) Dilekçe c) Gazete Haberi d) Sözleşme e) Hikâye

4.Aşağıdakilerden hangisi dilin kültür taşıyıcısı olduğunu gösterir?
a) Dilin seslerden örülmüş bir yapının olması
b) İnsanların iletişim kurabilmek için genellikle dili kullanması
c) Bazı dillerin zamanla unutulması
d) Atasözlerinin kulaktan kulağa çağımıza ulaşması
e) Her milletin dilinin farklı olması

5.Aşağıdakilerden hangisi edebi metin olabilir?
a) Türkiye'de yer altı madenlerini anlatan metin
b) Öğrencinin derste tuttuğu not
c) Sıla özlemini anlatan metin
d) Suyun kaynama noktasını anlatan metin
e) Kaç kaç olayını anlatan metin

6.Gündelik hayattaki konuşma dili farklıdır;bilimde……,
felsefede……, sanatta ise……kullanılır. **Boşluklara aşağıdakilerden hangisi gelecektir.**
a)İmge, terim, kavram b)Terim, kavram, imge c)Kavram, imge, terim d)terim, imge, kavram
e)Kavram, imge, imge

7.1. Bundan, 2.paydos, 3.küçük, 4.sonra, 5.heyecanlara **(Bu sözcüklerle anlamlı bir cümle oluşturulursa sıralama nasıl olmalıdır?** a)5.1.4.3.2 b)5.3.1.2.4 c)1.3.4.5.2 d)1.4.3.5.2 e)2.4.3.5.1.

BAŞARILAR

Figure 4.3: Answer key

In this case study, the algorithms developed throughout the research are applied on the real exam papers. The exam papers were obtained from Ankara Demetevler Anatolian Vocational Religious High School. In the automatic grading literature there are many computer based tools which are used for multiple-choice or true/false questions but there is no any automatic grading tool for questions needing written answers. In this study we develop a text mining tool for such questions and therefore we only focus on the questions needing written answers during the creation of the tool. The data set contains student answer papers and answer keys prepared by the graders. In figure 4.3 an original copy one of the answer keys is represented. For each student we got 3 different exam papers. First of all for each student we bring the 3 student answer papers and 3 corresponding answer keys together. And then clean all the combinations from the multiple-choice and true/false questions.

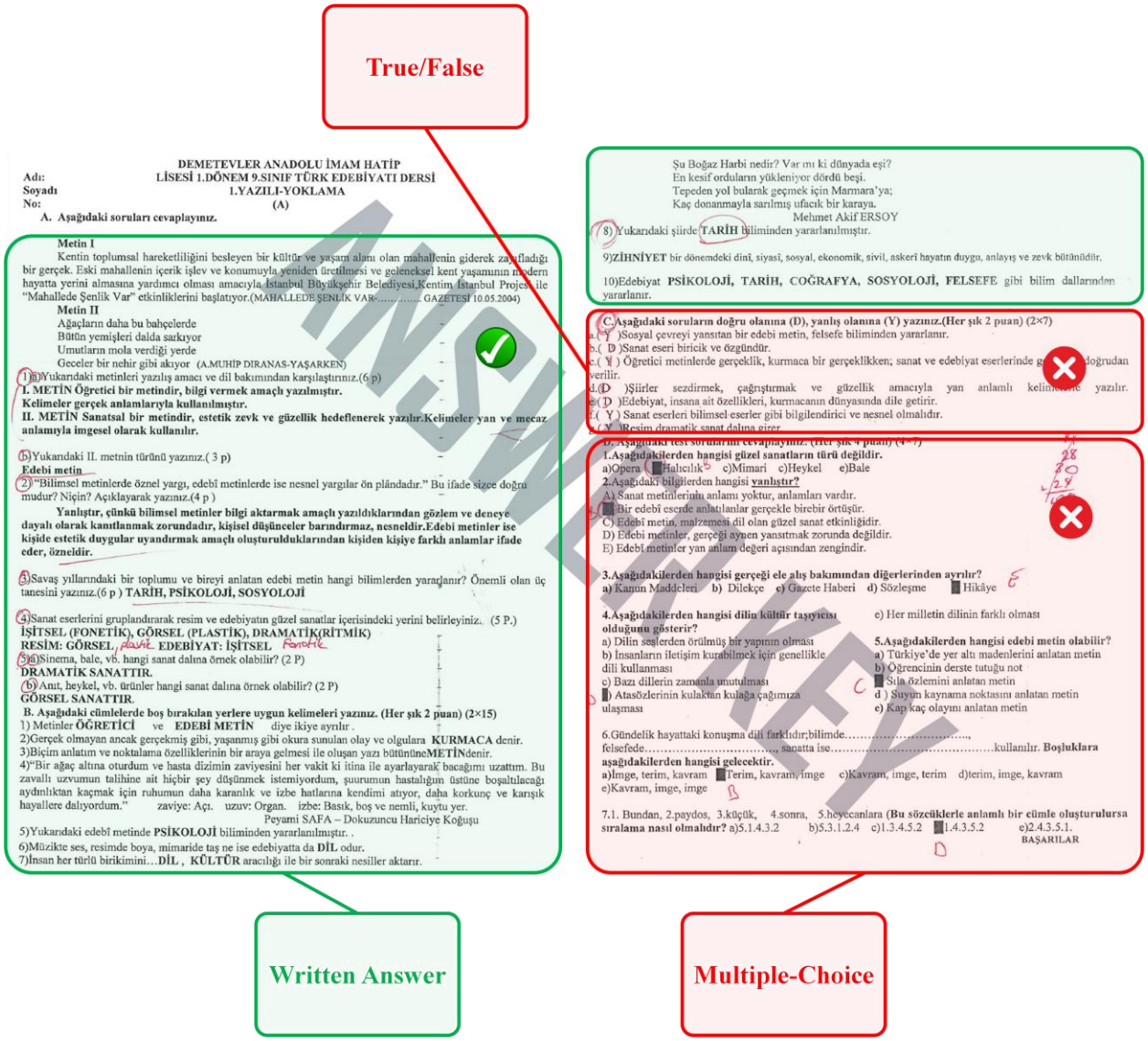


Figure 4.4: Categories of the answer key

Figure 4.5 is an original copy of the combination prepared for each student. As we mention before, this combination is only based on the questions needing written answers. The next step is to categorize the answer key combination. Each question is a main category. Also, new sub-categories are extracted from the main categories. The sub-categories correspond to the sub-answers in a main answer. Depending on the content of the sub-category, smaller sub-categories may be extracted from the first sub-category. In other words, in the final step of the

categorization we want to have sub-categories such that any further categorization cannot be performed on them. After completing the categorization, the credit of the question/main category is divided by the sub-categories correspondingly.

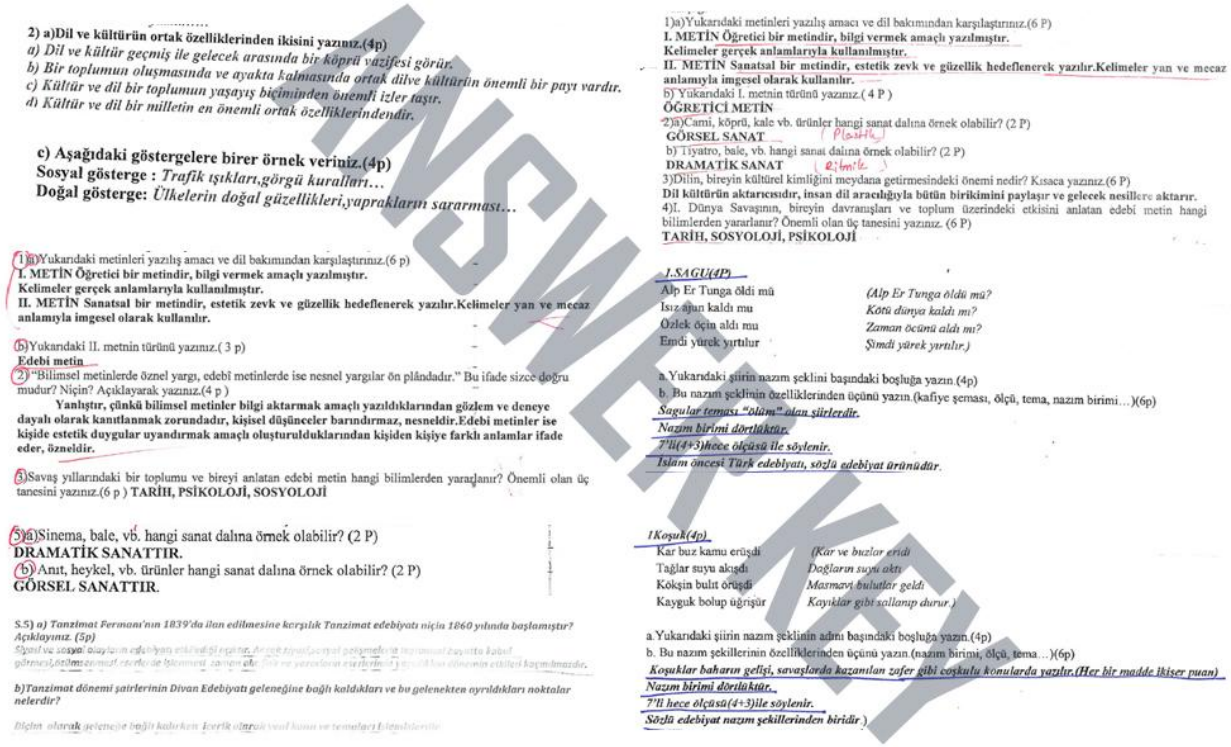


Figure 4.5: Selected categories of the answer keys

For each category we repeat the same process. Since it takes very large space to represent the processes of the all categories with graphics, we only present the process of a randomly selected category. At the end of the study the final results of the all categories are given in a single table.

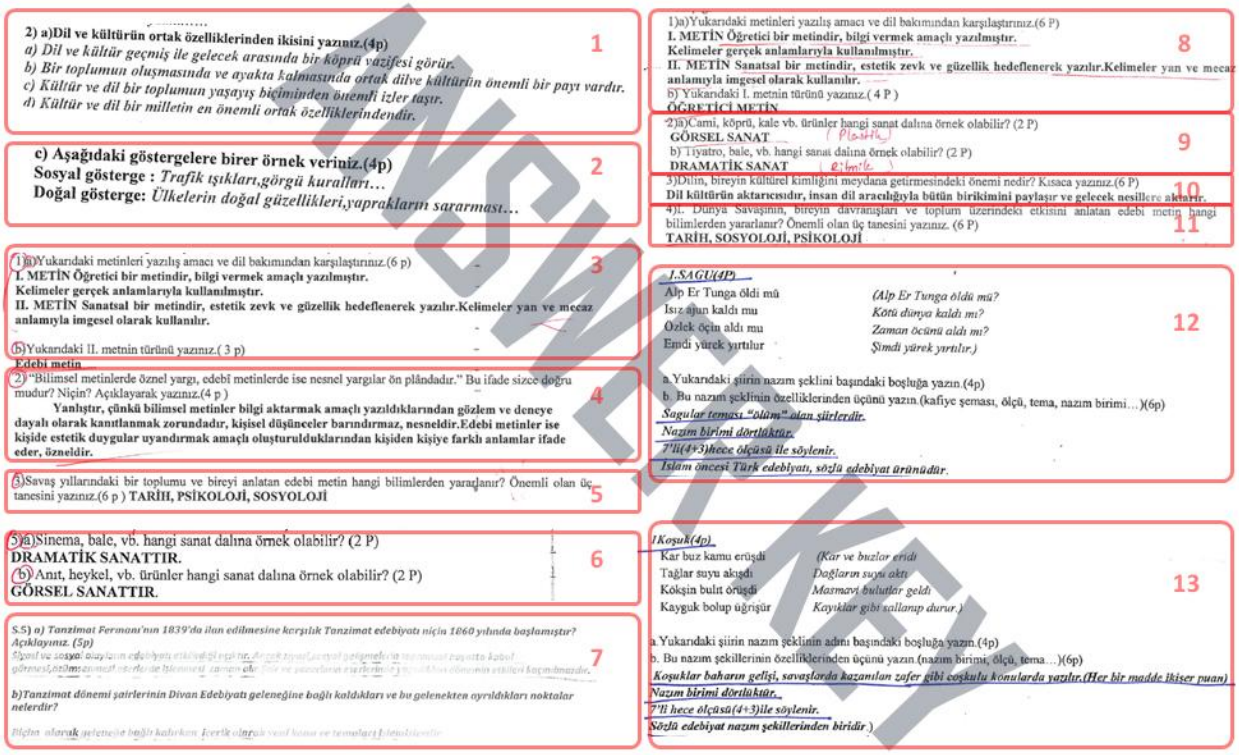
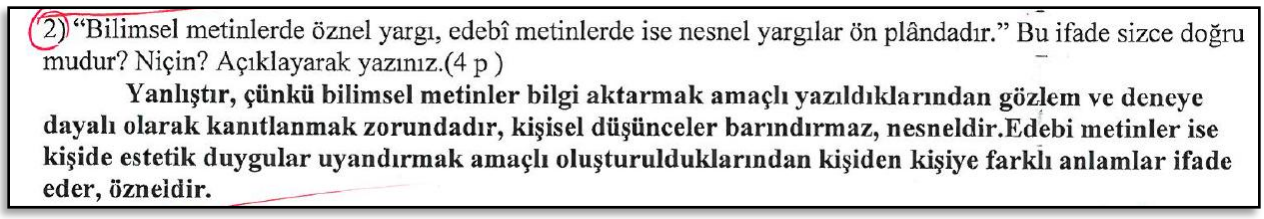


Figure 4.6: Categorized answer key

The 4th category was selected randomly. We share the details of the process set on the 4th category. Figure 4.7 is an original copy of the 4th category. First the scanned image of the category is converted into text file. In figure 4.8 you can find the converted version and the English translation of the category. Also, in each step of the algorithm creation process we give the translations of the Turkish parts. The reasons that we give the English translation of the category are to make the non-Turkish readers to have more details about the text, and to have more idea about the process of algorithm creation. Other than these reasons, the English translations do not have any contribution for the creation and application of the algorithms.


2) "Bilimsel metinlerde öznel yargı, edebî metinlerde ise nesnel yargılar ön plândadır." Bu ifade sizce doğru mudur? Niçin? Açıklayarak yazınız.(4 p)

Yanlıştır, çünkü bilimsel metinler bilgi aktarmak amaçlı yazıldıklarından gözlem ve deneye dayalı olarak kanıtlanmak zorundadır, kişisel düşünceler barındırmaz, nesneldir.Edebi metinler ise kişide estetik duygular uyandırmak amaçlı oluşturulduklarından kişiden kişiye farklı anlamlar ifade eder, özneldir.

Figure 4.7: Selected category

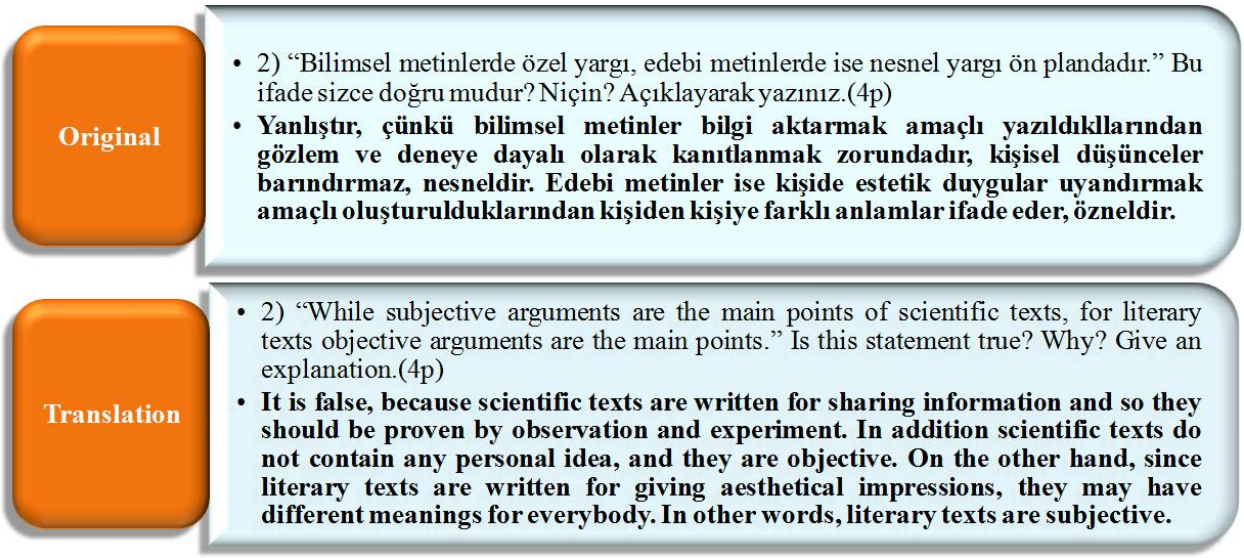


Figure 4.8: Original and translated versions of the selected category

- Question-2 is a main category and it has two sub-categories

Each one of the main questions is a main category. So the first categorization is based on the main questions. The 2nd question is a main category and so we show the applications of the algorithms on the 2nd question.

- For each sub-category, the corresponding sub-answer is extracted

The applications of the algorithms on the categories are very similar and we repeat the same processes for the all categories. First we identify the sub-categories in the main categories. If there are smaller sub-categories in the indentified sub-categories, then the categorization

continues. The main goal of doing categorization is to divide the each main category into as small parts as possible so that each part cannot be graded with any smaller grade. In other words, by doing categorization we want to get the most detailed matching between the grades and answers.

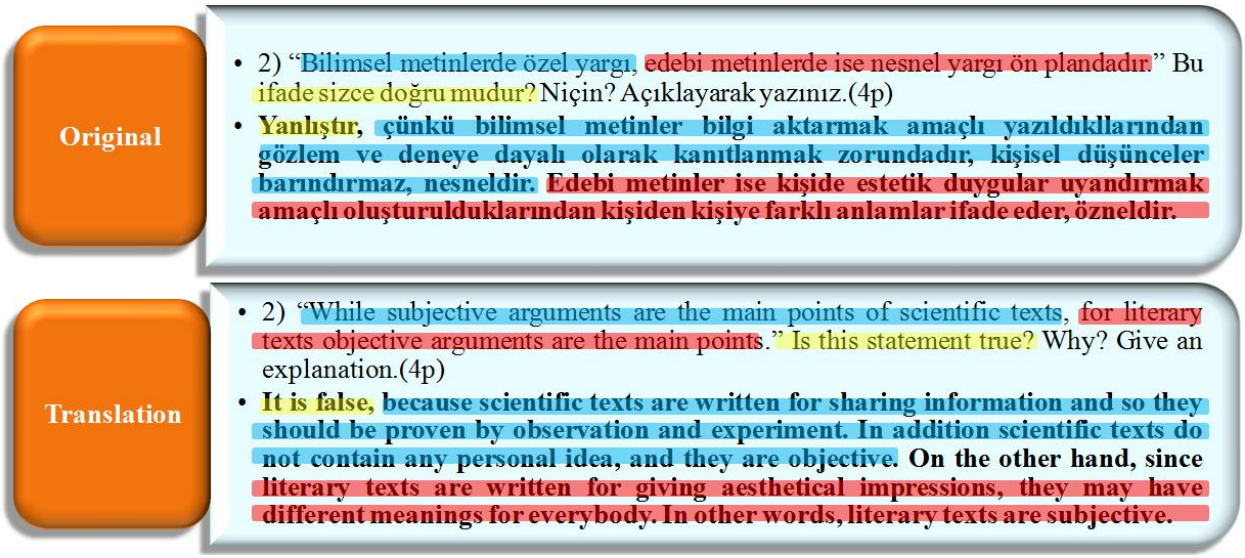


Figure 4.9: Sub-categories of the selected category

The method that we use for identifying the sub-categories is to trace the relation between the questions and answers, and the relation between the words within the answers. First we identify the sub-questions and sub-answers correspondingly. Each sub-answer is assigned as a subcategory.

In this question we identified two sub-categories. As we mentioned before, we repeat the same process for each sub-category. So we only give the process of the first category. The results of the second category will be given with the other sub-categories in the final table.

Figure 4.10: Sub-categories of the selected category

Now we have the sub-categories as smallest as possible. So we can start for the application of the algorithms on the category. First we need to extract the significant feature from the category. Figure 4.11 has the list of the extracted features.

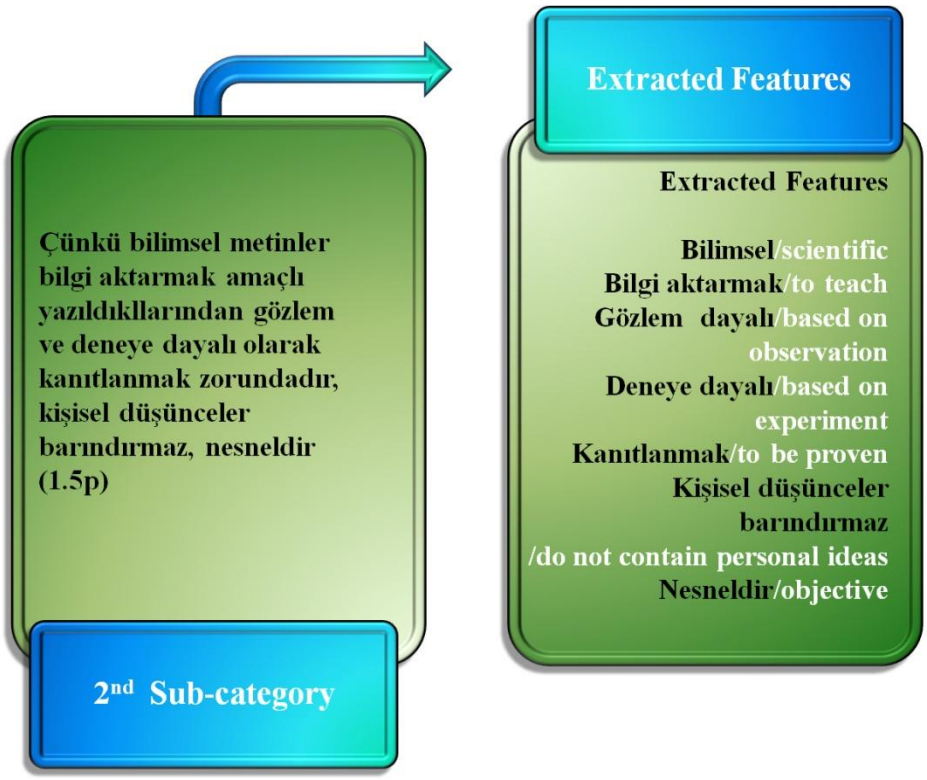


Figure 4.11: Extracted features of the selected sub-categories

The synonyms and related words/phrases of the extracted features in the Turkish Language Association – Big Turkish Dictionary (TLA – BTD) are added in the list and then the extended list (Figure 4.12) is uploaded into the text miner tool. The next is to prepare the stemming list.

Synonym and related words list of the extracted features:

- **False/yanlış = yanlış**/Doğru değil
- **Scientific/bilimsel = bilim**/Bilimle ilgili/Bilgiye dayalı/Bilime dayalı/Bilgisel/ Bilgi/İlim/İlimsel/İlime dayalı/İlimle ilgili
- **To teach/bilgi aktarmak = Bilgi aktar**/İlim aktar/bilimsel aktar/ilimsel aktar

- **Based on observation/gözleme dayalı = gözlem/**gözleme bağlı/gözetime dayalı/gözetime bağlı/izlemeye bağlı/izlemeye dayalı/gözleme dayanan/izle dayanan/gözetime dayanan/izlemeye dayanan
- **Based on experiment/deneye dayalı = deney/**deneye dayalı/deneye bağlı/deneye dayanan
- **To be proven/kanıtlanmak = kanıt/**ispat/delil
- **Does not contain personal ideas/kişisel düşünceler barındırmaz = kişi düşünce barındırma/**şahsi düşünce yok/şahsi düşünce bulunmaz/şahsi düşünce barındırmaz/ şahsi fikir yok/şahsi fikir bulunmaz/şahsi fikir barındırmaz/kişisel düşünce yok/kişisel düşünce bulunmaz/kişisel düşünce barındırmaz/kişisel değil/şahsi değil
- **Objective/nesneldir = nesne**

Synonym List Uploaded in the Text Miner Tool:

Yanlış	bilgi aktarmak	gözetime dayanan	şahsi düşünce yok
Doğru değil	Bilgi aktar	izlemeye dayanan	şahsi düşünce bulunmaz
Bilimsel	İlim aktar	deneye dayalı	şahsi düşünce barındırmaz
Bilim	bilimsel aktar	deney	şahsi fikir yok
Bilimle ilgili	ilimsel aktar	deneye dayalı	şahsi fikir bulunmaz
Bilgiye dayalı	gözleme dayalı	deneye bağlı	şahsi fikir barındırmaz
Bilime dayalı	gözlem	deneye dayanan	kişisel düşünce yok
Bilime dayanan	gözleme bağlı	kanıtlanmak	kişisel düşünce bulunmaz
Bilgisel	gözetime dayalı	kanıt	kişisel düşünce barındırmaz
Bilgi	gözetime bağlı	Ispat	kişisel değil
İlim	izlemeye bağlı	delil	şahsi değil
İlimsel	izlemeye dayalı	kişisel düşünceler	nesneldir
İlime dayalı	gözleme dayanan	barındırmaz	nesne
İlimle ilgili	izle dayanan	kişi düşünce barındırma	

Figure 4.12: Synonym words list extracted from the selected sub-category

For the preparation of the stemming list, we use the ‘almost all Turkish suffixes’ list. For the all terms in the synonyms and related words/phrases list, all the possible term+suffix matchings are added into the stemming list. In other words, all the terms in the figure 4.12 are

used as stems and matched with the suffixes from the ‘almost all Turkish suffixes’ list. For saving the time we have some reduction on the suffix list. The reduction is applied based on the Turkish grammar rules. Some of the grammar rules that are used for the reduction are vowel/consonant harmony rules, first-second-third singular/plural person(s) rules and so on.

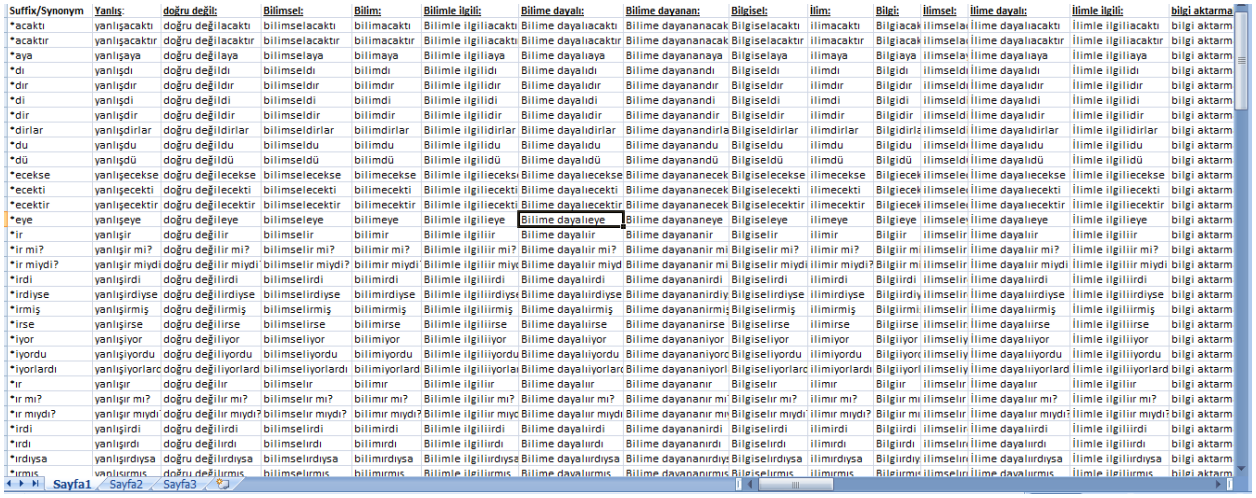
Based on the synonym list, the stemming lists will be prepared. In preparation of the stemming list, we utilize the ‘almost all Turkish suffixes’ list.

- The question and the answer require 3rd person singular answer
- The correct answer word has -a -ı vowels and it ends with –ş consonant
- The synonym of the correct answer has -o -u –e –ı vowels and it ends with –l consonant
- The incorrect answer word has -o -u vowels and it ends with –u vowel
- The synonym of the incorrect answer has -a –ı –e -i vowels and it ends with –l consonant

The suffix list will be reduced based on the above given informations. The suffixes are from the list “almost all Turkish suffixes”. Based on the given informations above, the suffixes list is reduced. The following is a part of the reduced list.

-acaktı	-ir miydi?	-iyor	-memisti	-mismisse	-muşdu	-rdu	-ur	-üyordu
-acaktır	-irdi	-iyordu	-memuşdu	-misse	-muşmussa	-rdü	-ur mu?	-üyorlardı
-aya	-irdiyse	-iyorlardı	-memuş	-misti	-musse	-rlar	-ur mıydı?	-yacaktı
-dı	-irmiş	-malı	-meu	-miyecektir	-u	-rler	-urdu	-yacaktır
-dır	-irse	-mamalı	-memüş	-miyor	-muyor	-rmıs	-urduysa	-yaya
-di	-iyor	-maya	-memüşdü	-miyor mu?	-müs	-rmiş	-urmuş	-yecekse
-dir	-iyordu	-maz	-memüştü	-mis	-müşdü	-rsa	-uyor	-yecekti
-dirlar	-iyorlardı	-maz mi?	-meye	-misdi	-müşdür	-rse	-uyordu	-yecektir
-du	-ır	-mazdi	-mez	-misdir	-müşmüşse	-sin	-uyorlardı	-yeye
-dü	-ır mı?	-meli	-mez mi?	-mismissa	-müşsa	-sın	-ür	-yor
-ecekse	-ır mıydı?	-memeli	-mezdi	-missa	-müştü	-ti	-ür mi?	-yordu
-ecekti	-irdi	-memis	-mezler	-misti	-müyor	-tir	-ür miydi?	-yorlardı
-ecektir	-ırdı	-memişdi	mi?	-miyacaktir	-r	-tı	-ürdü	
-eye	-ırdıysa	-memişti	-mis	-miyor	-r mi?	-tır	-ürdüyse	
-ir	-ırmış	-memiş	-misdi	-muş	-rdi	-tu	-ürmüş	
-ir mi?	-ırsa	-memisdi	-misdir	-muşdur	-rdı	-tü	-üyor	

Figure 4.13: Reduced suffix list



Suffix/Synonym	Yanlış:	doğru değil:	Bilimsel:	Bilim:	Bilimle ilgili:	Bilime dayalı:	Bilime dayanan:	Bilgisel:	İlim:	Bilgi:	İlimsel:	İlime dayalı:	İlimle ilgili:	bilgi aktarma
*acaktı	yanlışacaktı	doğru değilacaktı	bilimselacaktı	bilimacaktı	Bilimle ilgiliacaktı	Bilime dayalıacaktı	Bilime dayananacak	Bilgiselacaktı	ilimacaktı	Bilgiacak	ilimselac	İlime dayalıacaktı	İlimle ilgiliacaktı	bilgi aktarm
*acaktır	yanlışacaktır	doğru değilacaktır	bilimselacaktır	bilimacaktır	Bilimle ilgiliacaktı	Bilime dayalıacaktır	Bilime dayananacak	Bilgiselacaktır	ilimacaktır	Bilgiacak	ilimselac	İlime dayalıacaktır	İlimle ilgiliacaktır	bilgi aktarm
*aya	yanlışaya	doğru değilaya	bilimselaya	bilimaya	Bilimle ilgiliaya	Bilime dayalıaya	Bilime dayananaya	Bilgiselaya	ilimaya	Bilgiaya	ilimselay	İlime dayalıaya	İlimle ilgiliaya	bilgi aktarm
*dı	yanlışdı	doğru değildı	bilimseldı	bilimdı	Bilimle ilgilidı	Bilime dayalıdı	Bilime dayandı	Bilgiseldı	ilimdı	Bilgidı	ilimseldi	İlime dayalıdı	İlimle ilgilidı	bilgi aktarm
*dır	yanlışdır	doğru değildır	bilimseldır	bilimdır	Bilimle ilgilidır	Bilime dayalıdır	Bilime dayanandır	Bilgiseldır	ilimdır	Bilgidır	ilimseldi	İlime dayalıdır	İlimle ilgilidır	bilgi aktarm
*di	yanlışdi	doğru değildi	bilimseldi	bilimdi	Bilimle ilgilidi	Bilime dayalıdi	Bilime dayanandi	Bilgiseldi	ilimdi	Bilgidi	ilimseldi	İlime dayalıdi	İlimle ilgilidi	bilgi aktarm
*dir	yanlışdir	doğru değildir	bilimseldir	bilimdir	Bilimle ilgilidir	Bilime dayalıdir	Bilime dayanandir	Bilgiseldir	ilimdir	Bilgidir	ilimseldi	İlime dayalıdir	İlimle ilgilidir	bilgi aktarm
*dirlar	yanlışdirlar	doğru değildirlar	bilimseldirlar	bilimdirlar	Bilimle ilgilidirlar	Bilime dayalıdirlar	Bilime dayanandirla	Bilgiseldirlar	ilimdirlar	Bilgidirla	ilimseldi	İlime dayalıdirlar	İlimle ilgilidirlar	bilgi aktarm
*du	yanlışdu	doğru değildu	bilimseldu	bilimdu	Bilimle ilgilidu	Bilime dayalıdu	Bilime dayanandu	Bilgiseldu	ilimdu	Bilgidu	ilimseldi	İlime dayalıdu	İlimle ilgilidu	bilgi aktarm
*dü	yanlışdü	doğru değildü	bilimseldü	bilimdü	Bilimle ilgilidü	Bilime dayalıdü	Bilime dayandü	Bilgiseldü	ilimdü	Bilgidü	ilimseldi	İlime dayalıdü	İlimle ilgilidü	bilgi aktarm
*ecekse	yanlışecekse	doğru değilecekse	bilimselecekse	bilimecekse	Bilimle ilgilieksı	Bilime dayalıecekse	Bilime dayananecek	Bilgiselecekse	ilimecekse	Bilgiecek	ilimselec	İlime dayalıecekse	İlimle ilgiliecekse	bilgi aktarm
*ecekti	yanlışecekti	doğru değilecekti	bilimselecekti	bilimecekti	Bilimle ilgiliecekti	Bilime dayalıecekti	Bilime dayananecek	Bilgiselecekti	ilimecekti	Bilgiecek	ilimselec	İlime dayalıecekti	İlimle ilgiliecekti	bilgi aktarm
*ecektir	yanlışecektir	doğru değilecektir	bilimselecektir	bilimecektir	Bilimle ilgiliecekti	Bilime dayalıecektir	Bilime dayananecek	Bilgiselecektir	ilimecektir	Bilgiecek	ilimselec	İlime dayalıecektir	İlimle ilgiliecektir	bilgi aktarm
*eye	yanlışeye	doğru değileye	bilimseleye	bilimeye	Bilimle ilgilieye	Bilime dayalıeye	Bilime dayananeye	Bilgiseleye	ilimeye	Bilgieye	ilimselei	İlime dayalıeye	İlimle ilgilieye	bilgi aktarm
*ir	yanlışir	doğru değilir	bilimselir	bilimir	Bilimle ilgilir	Bilime dayalır	Bilime dayananir	Bilgiselir	ilimir	Bilgiir	ilimselir	İlime dayalır	İlimle ilgilir	bilgi aktarm
*ir mi?	yanlışir mi?	doğru değilir mi?	bilimselir mi?	bilimir mi?	Bilimle ilgilir mi?	Bilime dayalır mi?	Bilime dayananir mi	Bilgiselir mi?	ilimir mi?	Bilgiir mi	ilimselir	İlime dayalır mi?	İlimle ilgilir mi?	bilgi aktarm
*ir miydi?	yanlışir miydi	doğru değilir miydi	bilimselir miydi?	bilimir miydi	Bilimle ilgilir miyc	Bilime dayalır miyd	Bilime dayananir mi	Bilgiselir miydi	ilimir miydi?	Bilgiir mi	ilimselir	İlime dayalır miydi	İlimle ilgilir miydi	bilgi aktarm
*irdi	yanlışirdi	doğru değilirdi	bilimselirdi	bilimirdi	Bilimle ilgilirdi	Bilime dayalırdi	Bilime dayananirdi	Bilgiselirdi	ilimirdi	Bilgiirdi	ilimselir	İlime dayalırdi	İlimle ilgiliirdi	bilgi aktarm
*irdiyse	yanlışirdiyse	doğru değilirdiyse	bilimselirdiyse	bilimirdiyse	Bilimle ilgiliirdiyse	Bilime dayalırdiyse	Bilime dayananirdiy	Bilgiselirdiyse	ilimirdiyse	Bilgiirdiy	ilimselir	İlime dayalırdiyse	İlimle ilgiliirdiyse	bilgi aktarm
*irmiş	yanlışirmiş	doğru değilirmiş	bilimselirmiş	bilimirmiş	Bilimle ilgilirmiş	Bilime dayalırmış	Bilime dayananirmiş	Bilgiselirmiş	ilimirmiş	Bilgiirmi	ilimselir	İlime dayalırmış	İlimle ilgiliirmiş	bilgi aktarm
*irse	yanlışirse	doğru değilirse	bilimselirse	bilimirse	Bilimle ilgiliirse	Bilime dayalırse	Bilime dayananirse	Bilgiselirse	ilimirse	Bilgiirse	ilimselir	İlime dayalırse	İlimle ilgiliirse	bilgi aktarm
*iyor	yanlışiyor	doğru değiliyor	bilimseliyor	bilimiyor	Bilimle ilgiliyor	Bilime dayaliiyor	Bilime dayananiyor	Bilgiseliyor	ilimiyor	Bilgiiyor	ilimseliy	İlime dayaliiyor	İlimle ilgiliiyor	bilgi aktarm
*iyordu	yanlışiyordu	doğru değiliyordu	bilimseliyordu	bilimiyordu	Bilimle ilgiliiyordu	Bilime dayaliiyordu	Bilime dayananiyorc	Bilgiseliyordu	ilimiyordu	Bilgiiyorc	ilimseliy	İlime dayaliiyordu	İlimle ilgiliiyordu	bilgi aktarm
*iyorlardı	yanlışiyorlarc	doğru değiliyorlard	bilimseliyorlardı	bilimiyorlard	Bilimle ilgiliiyorlaı	Bilime dayaliiyorlarc	Bilime dayananiyorl	Bilgiseliyorlarc	ilimiyorlardı	Bilgiiyorl	ilimseliy	İlime dayaliiyorlard	İlimle ilgiliiyorlard	bilgi aktarm
*ır	yanlışır	doğru değilir	bilimselır	bilimır	Bilimle ilgilir	Bilime dayalır	Bilime dayananır	Bilgiselır	ilimır	Bilgiır	ilimselır	İlime dayalır	İlimle ilgilir	bilgi aktarm
*ır mı?	yanlışır mı?	doğru değilir mı?	bilimselır mı?	bilimır mı?	Bilimle ilgilir mı?	Bilime dayalır mı?	Bilime dayananır mı	Bilgiselır mı?	ilimır mı?	Bilgiır mı	ilimselır	İlime dayalır mı?	İlimle ilgilir mı?	bilgi aktarm
*ır mıydı?	yanlışır mıydı	doğru değilir mıydı?	bilimselır mıydı?	bilimır mıydı?	Bilimle ilgilir mıyc	Bilime dayalır mıydı	Bilime dayananır mı	Bilgiselır mıydı	ilimır mıydı?	Bilgiır mı	ilimselır	İlime dayalır mıydı?	İlimle ilgilir mıydı?	bilgi aktarm
*irdi	yanlışirdi	doğru değilirdi	bilimselirdi	bilimirdi	Bilimle ilgiliirdi	Bilime dayalıirdi	Bilime dayananirdi	Bilgiselirdi	ilimirdi	Bilgiirdi	ilimselir	İlime dayalıirdi	İlimle ilgiliirdi	bilgi aktarm
*ırdı	yanlışırdı	doğru değilırdı	bilimselırdı	bilimırdı	Bilimle ilgilirdi	Bilime dayalırdı	Bilime dayananırdı	Bilgiselırdı	ilimırdı	Bilgiırdı	ilimselır	İlime dayalırdı	İlimle ilgilirdi	bilgi aktarm
*ırdıysa	yanlışırdıysa	doğru değilırdıysa	bilimselırdıysa	bilimırdıysa	Bilimle ilgilirdıysa	Bilime dayalırdıysa	Bilime dayananırdıys	Bilgiselırdıysa	ilimırdıysa	Bilgiırdıy	ilimselır	İlime dayalırdıysa	İlimle ilgilirdıysa	bilgi aktarm
*ırmış	yanlışırmış	doğru değilırmış	bilimselırmış	bilimırmış	Bilimle ilgilirmiş	Bilime dayalırmış	Bilime dayananırmış	Bilgiselırmış	ilimırmış	Bilgiırmı	ilimselır	İlime dayalırmış	İlimle ilgilirmiş	bilgi aktarm

Sayfa1 Sayfa2 Sayfa3

Figure 4.14: Stem+suffix matchings

Figure 4.14 is a screen captured picture of the list of all stem+suffix matchings. However some the matchings do not take place in the TLA-BTD. It means these matchings do not have meaning or not used in daily life. From the matchings list only the meaningful ones are selected and uploaded into the text miner tool.

Yanlış	Bilgiye dayalıdır	İlimle ilgilidir	gözlemmişdir	gözetime dayanan	deneye dayananırdıysa	barındırmazmışdı	barındırmazmıştı	bulunmazmışse
yanlışdı	--	İlimle ilgilidirler	gözlemmişmişse	gözetime dayanandır	deneye dayananırmış	kişisel düşünceler	şahsi fikir yok	kişisel düşünce
yanlışdır	Bilime dayanan	bilgi aktarmak	gözlemmişse	gözetime dayanandı	deneye dayananırsa	barındırmazmışdır	şahsi fikir yokdu	bulunmazmıştı
yanlışdırlar	Bilime dayanandır	bilgi aktarmakdı	gözlemmişti	gözetime dayanandırlar	deneye dayanansın	kişisel düşünceler	şahsi fikir yokmuş	kişisel düşünce
yanlışmış	Bilime dayanandı	bilgi aktarmakdır	gözleme bağlı	gözetime dayananır	kanıtlanmak	barındırmazmışmışse	şahsi fikir yokmuşdur	bulunmazmıştı
yanlışmışdı	Bilime dayanandırlar	bilgi aktarmakmış	gözleme bağlıdır	gözetime dayananır mı?	kanıtlanmakdı	kişisel düşünceler	şahsi fikir yokmuşdu	kişisel düşünce
yanlışmışdır	Bilime dayananır	bilgi aktarmakmışdı	gözleme bağlıdırlar	gözetime dayananır mıydı?	kanıtlanmakdır	barındırmazmışse	şahsi fikir yokmuşmuşsa	barındırmaz
yanlışmıştı	Bilime dayananır mı?	bilgi aktarmakmışdır	gözleme bağlıyacaktı	gözetime dayananırdı	kanıtlanmakmış	kişisel düşünceler	şahsi fikir yokmuşsa	kişisel düşünce
yanlışmışmışsa	Bilime dayananır mıydı?	bilgi aktarmakmıştı	gözleme bağlıyacaktır	gözetime dayananırdıysa	kanıtlanmakmışdı	barındırmazmıştı	şahsi fikir yoktu	barındırmazdı
yanlışmışsa	Bilime dayananırdı	bilgi aktarmakmışmışsa	gözleme bağlıya	gözetime dayananırmış	kanıtlanmakmışdır	kişisel düşünceler	şahsi fikir bulunmaz	kişisel düşünce
yanlışsın	Bilime dayananırdıysa	bilgi aktarmakmışsa	gözleme bağlıyor	gözetime dayananırsa	kanıtlanmakmıştı	barındırmazmıştı	şahsi fikir bulunmazdı	barındırmazdır
yanlışsın	Bilime dayananırmış	Bilgi aktar	gözleme bağlıyordu	gözetime dayanansın	kanıtlanmakmışmışsa	şahsi düşünce yok	şahsi fikir bulunmazdır	kişisel düşünce
yanlıştı	Bilime dayananırsa	Bilgi aktardı	gözleme bağlıyorlardı	izlemeye dayanan	kanıtlanmakmışsa	şahsi düşünce yokdu	şahsi fikir bulunmazdırlar	barındırmazdırlar
yanlıştır	Bilime dayanansın	Bilgi aktardır	gözetime dayalı	izlemeye dayanandır	kanıtlanmakmıştı	şahsi düşünce yokmuş	şahsi fikir bulunmazmış	kişisel düşünce
yanlıştı	--	Bilgi aktarmış	gözetime dayalıdır	izlemeye dayanandı	kanıtlanmaktı	şahsi düşünce yokmuşdur	şahsi fikir bulunmazmışdı	barındırmazmış
-	Bilgisel	Bilgi aktarmışdı	gözetime bağlı	izlemeye dayanandırlar	kanıtlanmaktır	şahsi düşünce yokmuşdu	şahsi fikir bulunmazmışdır	kişisel düşünce
doğru değil	Bilgiseldi	Bilgi aktarmışdır	gözetime bağlıdır	izlemeye dayananır	kanıt	şahsi düşünce	şahsi fikir	barındırmazmışdı
doğru değildi	Bilgiseldir	Bilgi aktarmıştı	gözetime bağlıdırlar	izlemeye dayananır mı?	kanıtdı	yokmuşmuşsa	bulunmazmışmışse	kişisel düşünce
doğru değildir	Bilgiseldirler	Bilgi aktarmışmışsa	gözetime bağlıyacaktı	izlemeye dayananır mıydı?	kanıtdır	şahsi düşünce yokmuşsa	şahsi fikir bulunmazmışse	barındırmazmışdır
doğru değildirler	Bilgiselmişti	Bilgi aktarmışsa	gözetime bağlıyacaktır	izlemeye dayananırdı	kanıtmış	şahsi düşünce yoktu	şahsi fikir bulunmazmıştı	kişisel düşünce
doğru değilmiş	Bilgiselmiş	İlim aktar	gözetime bağlıya	izlemeye dayananırdıysa	kanıtmışdı	şahsi düşünce bulunmaz	şahsi fikir bulunmazmıştı	barındırmazmışmışse
doğru değilmişdi	Bilgiselmişdi	İlim aktardı	gözetime bağlıyor	izlemeye dayananırmış	kanıtmışdır	şahsi düşünce bulunmazdı	şahsi fikir barındırmaz	kişisel düşünce
doğru değilmişdir	Bilgiselmişdir	İlima aktardır	gözetime bağlıyordu	izlemeye dayananırsa	kanıtmışmışsa	şahsi düşünce bulunmazdır	şahsi fikir barındırmazdı	barındırmazmışse
doğru değilmişmişse	Bilgiselmişmişse	İlim aktarmış	gözetime bağlıyorlardı	izlemeye dayanansın	kanıtmışsa	şahsi düşünce	şahsi fikir barındırmazdır	kişisel düşünce
doğru değilmişse	Bilgiselmişse	İlim aktarmışdı	izlemeye bağlı	deneye dayalı	kanıtmıştı	bulunmazdırlar	şahsi fikir barındırmazdırlar	barındırmazmıştı
doğru değilmişti	-----	İlim aktarmışdır	izlemeye bağlıdır	deneye dayalıdır	kanıttı	şahsi düşünce	şahsi fikir barındırmazmış	kişisel düşünce
doğru değilmişdi	İlim	İlim aktarmıştı	izlemeye bağlıdırlar	deney	kanıttır	bulunmazmış	şahsi fikir barındırmazmışdı	barındırmazmıştı
doğru değilmişdir	ilimdi	İlim aktarmışmışsa	izlemeye bağlıyacaktı	deneydi	ispat	şahsi düşünce	şahsi fikir	kişisel değil
Bilimsel	ilimdir	İlim aktarmışsa	izlemeye bağlıyacaktır	deneydir	ispatdı	bulunmazmışdı	barındırmazmışdır	kişisel değildi
bilimseldi	ilimdirler	bilimsel aktar	izlemeye bağlıya	deneydirler	ispatdır	şahsi düşünce	şahsi fikir	kişisel değildir
bilimseldir	ilimmiş	bilimsel aktardı	izlemeye bağlıyor	deneyecekse	ispatmış	bulunmazmışdır	barındırmazmışmışse	kişisel değildirler
bilimseldirler	ilimmişdi	bilimsel aktardır	izlemeye bağlıyordu	deneyecekti	ispatmışdı	şahsi düşünce	şahsi fikir	kişisel değilmiş
bilimselmis	ilimmişdir	bilimsel aktarmış	izlemeye bağlıyorlardı	deneyecektir	ispatmışdır	bulunmazmışmışse	barındırmazmışse	kişisel değilmişdi
bilimselmişdi	ilimmişmişse	bilimsel aktarmışdı	izlemeye dayalı	deneye	ispatmışmışsa	şahsi düşünce	şahsi fikir barındırmazmıştı	kişisel değilmişdir
bilimselmişdir	ilimmişse	bilimsel aktarmışdır	izlemeye dayalıdır	deneymiş	ispatmışsa	bulunmazmışse	şahsi fikir barındırmazmıştı	kişisel değilmişmişse
bilimselmişse	ilimmişti	bilimsel aktarmıştı	gözleme dayanan	deneymişdi	ispatmıştı	şahsi düşünce	kişisel düşünce yok	kişisel değilmişse
bilimselmişti	--	bilimsel aktarmışmışsa	gözleme dayanandır	deneymişdir	ispattı	bulunmazmıştı	kişisel düşünce yokdu	kişisel değilmişti
Bilim	Bilgi	bilimsel aktarmışsa	gözleme dayanandı	deneymişmişse	ispattır	şahsi düşünce	kişisel düşünce yokmuş	kişisel değilmişti
bilimdi	Bilgidir	ilimsel aktar	gözleme dayanandırlar	deneymişse	delil	bulunmazmıştı	kişisel düşünce yokmuşdur	şahsi değil
bilimdir	Bilgidirler	ilimsel aktardı	gözleme dayananır	deneymişti	delildi	şahsi düşünce barındırmaz	kişisel düşünce yokmuşdu	şahsi değildi
bilimdirler	Bilgiye	ilimsel aktardır	gözleme dayananır mı?	deneye dayalı	delildir	şahsi düşünce	kişisel düşünce	şahsi değildir
bilimmiş	-----	ilimsel aktarmış	gözleme dayananır mıydı?	deneye dayalıdır	delildirler	barındırmazdı	yokmuşmuşsa	şahsi değildirler
bilimmişdi	İlimsel	ilimsel aktarmışdı	gözleme dayananırdı	deneye bağlı	delilmiş	şahsi düşünce	kişisel düşünce yokmuşsa	şahsi değilmiş
bilimmişdir	ilimseldi	ilimsel aktarmışdır	gözleme dayananırdıysa	deneye bağlıdır	delilmişdi	barındırmazdır	kişisel düşünce yoktu	şahsi değilmişdi
bilimmişmişse	ilimseldir	ilimsel aktarmıştı	gözleme dayananırmış	deneye bağlıdırlar	delilmişdir	şahsi düşünce	kişisel düşünce bulunmaz	şahsi değilmişdir
bilimmişse	ilimseldirler	ilimsel aktarmışmışsa	gözleme dayananırsa	deneye bağlıyacaktı	delilmişmişse	barındırmazdırlar	kişisel düşünce bulunmazdı	şahsi değilmişmişse
bilimmişti	ilimselmiş	ilimsel aktarmışsa	gözleme dayanansın	deneye bağlıyacaktır	delilmişse	şahsi düşünce	kişisel düşünce	şahsi değilmişse
---	ilimselmişdi	gözleme dayalı	izle dayanan	deneye bağlıya	delilmişti	barındırmazmış	bulunmazdır	şahsi değilmişti
Bilimle ilgili	ilimselmişdir	gözleme dayalıdır	izle dayanandır	deneye bağlıyor	kişisel düşünceler	şahsi düşünce	kişisel düşünce	şahsi değilmişti
Bilimle ilgilidir	ilimselmişmişse	gözlem	izle dayanandı	deneye bağlıyordu	barındırmaz	barındırmazmışdı	bulunmazdırlar	nesnel
Bilimle ilgiliyecekse	ilimselmişse	gözlemdi	izle dayanandırlar	deneye bağlıyorlardı	kişisel düşünceler	şahsi düşünce	kişisel düşünce	nesneldi
Bilimle ilgiliyecekti	ilimselmişti	gözlemdir	izle dayananır	deneye dayanan	barındırmazdı	barındırmazmışdır	bulunmazmış	nesneldir
Bilimle ilgiliyecektir	İlime dayalı	gözlemdirler	izle dayananır mı?	deneye dayanandır	kişisel düşünceler	şahsi düşünce	kişisel düşünce	nesneldirler
Bilimle ilgiliye	İlime dayalıdı	gözlemeye	izle dayananır mıydı?	deneye dayanandı	barındırmazdır	barındırmazmışmışse	bulunmazmışdı	nesne
--	İlime dayalıdır	gözlemiyor	izle dayananırdı	deneye dayanandırlar	kişisel düşünceler	şahsi düşünce	kişisel düşünce	nesnedir
Bilime dayalı	İlime dayalıyor	gözlemiyordu	izle dayananırdıysa	deneye dayananır	barındırmazdırlar	barındırmazmışse	bulunmazmışdır	nesnedirler
Bilime dayalıdır	İlime dayalıyordu	gözlemiyorlardı	izle dayananırmış	deneye dayananır mı?	kişisel düşünceler	şahsi düşünce	kişisel düşünce	nesneye
-----	İlime dayalıyorlardı	gözlemmiş	izle dayananırsa	deneye dayananır mıydı?	barındırmazmış	barındırmazmıştı	bulunmazmışmışse	
Bilgiye dayalı	İlimle ilgili	gözlemmişdi	izle dayanansın	deneye dayananırdı	kişisel düşünceler	şahsi düşünce	kişisel düşünce	

Figure 4.15: Selected stem+suffix matchings

This is the stemming list and will be used in grouping the related words because the list has been created based on the features/terms extracted from the answer key. After preparing the stemming list, all the filtration parameters are defined and then the text is taken into the text processing. At the end of this process we extract numerical information from the answer key. This numerical information is used in the grading of the student answer papers. The main grading measurement is the matching rate between the answer key and student answer paper. Therefore we repeat the similar process for the student answer paper.

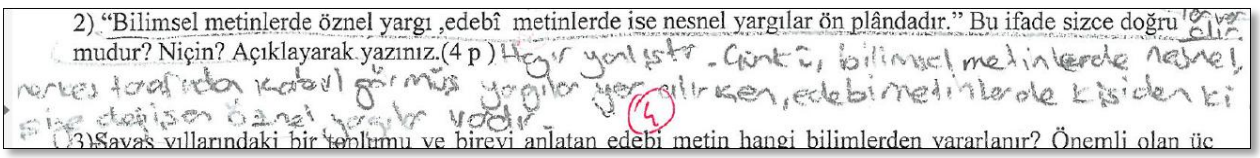


Figure 4.16: Student answer

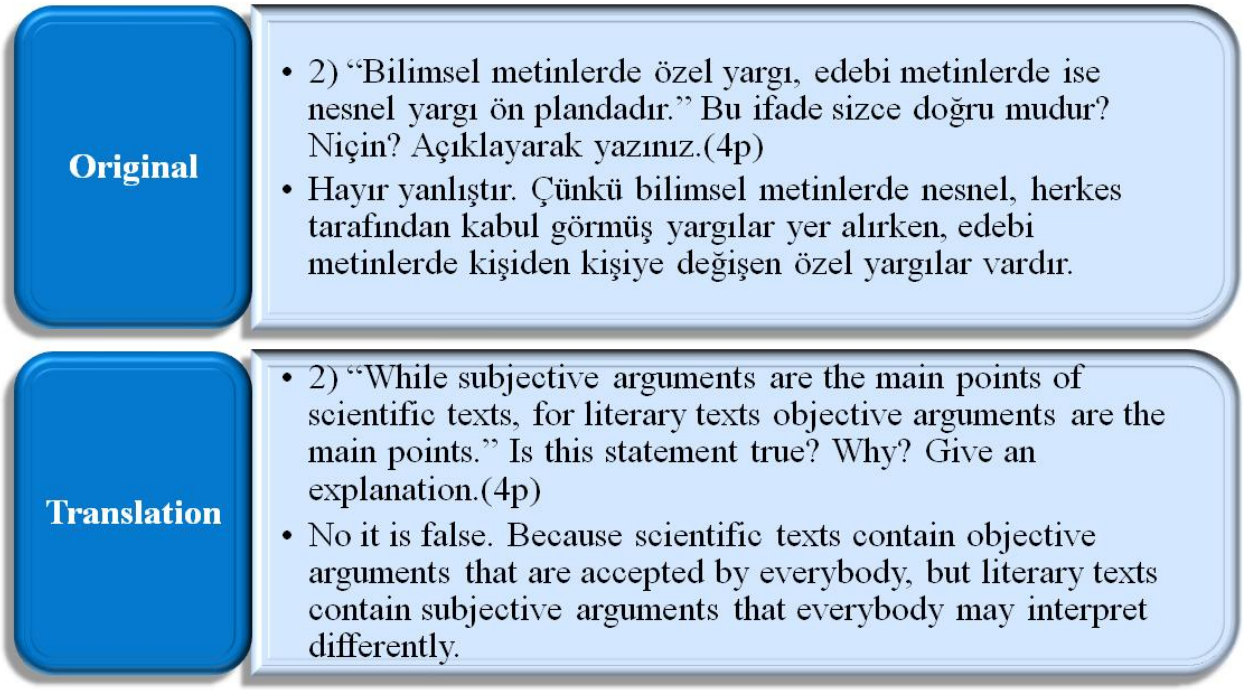


Figure 4.17: Original and translated versions of the student answer

First of all the original handwriting of the student is imported in a text file. And then, sub-categories of the student answer are identified by the sub-categories of the question that are identified before. As we did for the answer key, we only share the details of the process for the 1st category for saving some space in the thesis but you can see all the results of the sub-categories and main categories together in the final table.

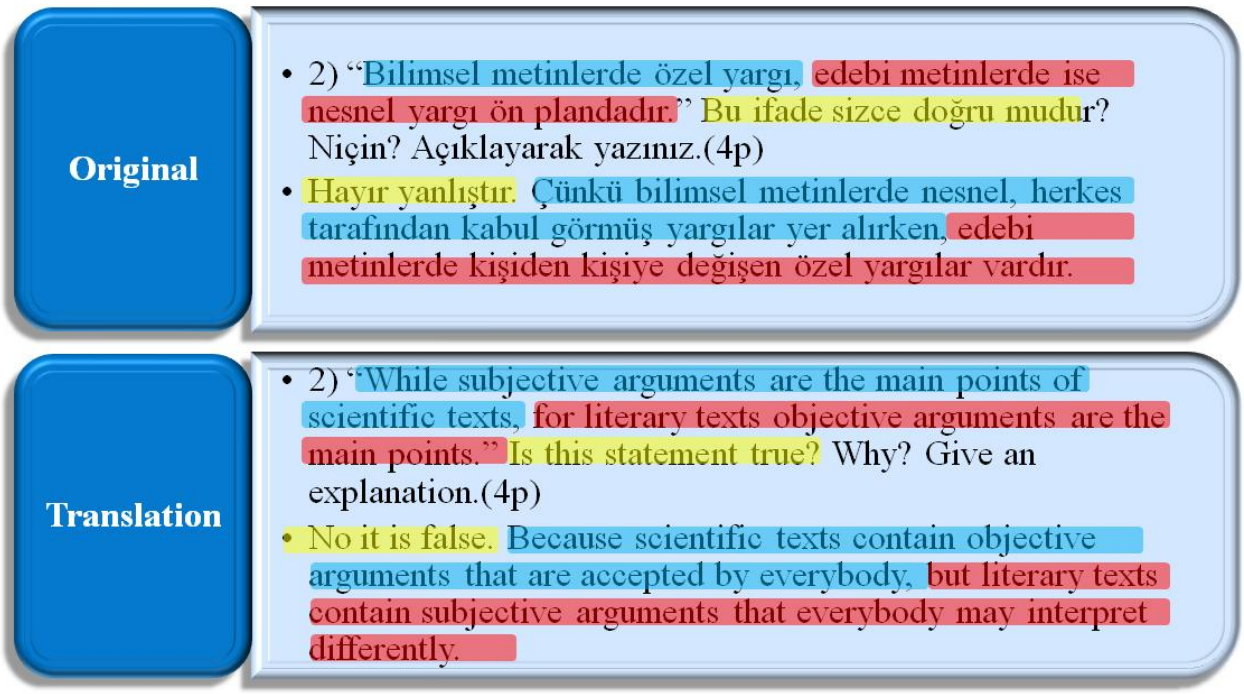


Figure 4.18: Selected sub-categories of the student answer

From the 1st sub-category all the significant features are extracted as a list. And then, all the synonyms and related words of the features that are taking place in the TLA-BTD are added into the list. The final list is called as the synonymous list. After uploading the list into text miner tool, a new list is prepared with the 'almost all Turkish suffixes" list. The terms of the synonymous are used as the stems and for each term all the possible stem+suffix matchings are obtained with the 'almost all Turkish suffixes'. All the meaningful matchings are selected. The final list is used as the stemming list and uploaded into the text miner tool. After defining all the filtration parameters, the text processing is run. At the end of this process, some numerical information is obtained.

The grading method of the exam is just based on the correct answers. In other words, the grader gives positive credits for the correct answers but does not give negative credits for the incorrect answers. Therefore, since the sub-categories are the possible smallest parts, just a single

term in the intersection of the answer key stemming list and student answer stemming list is enough for the student to get a full credit for that sub-category.

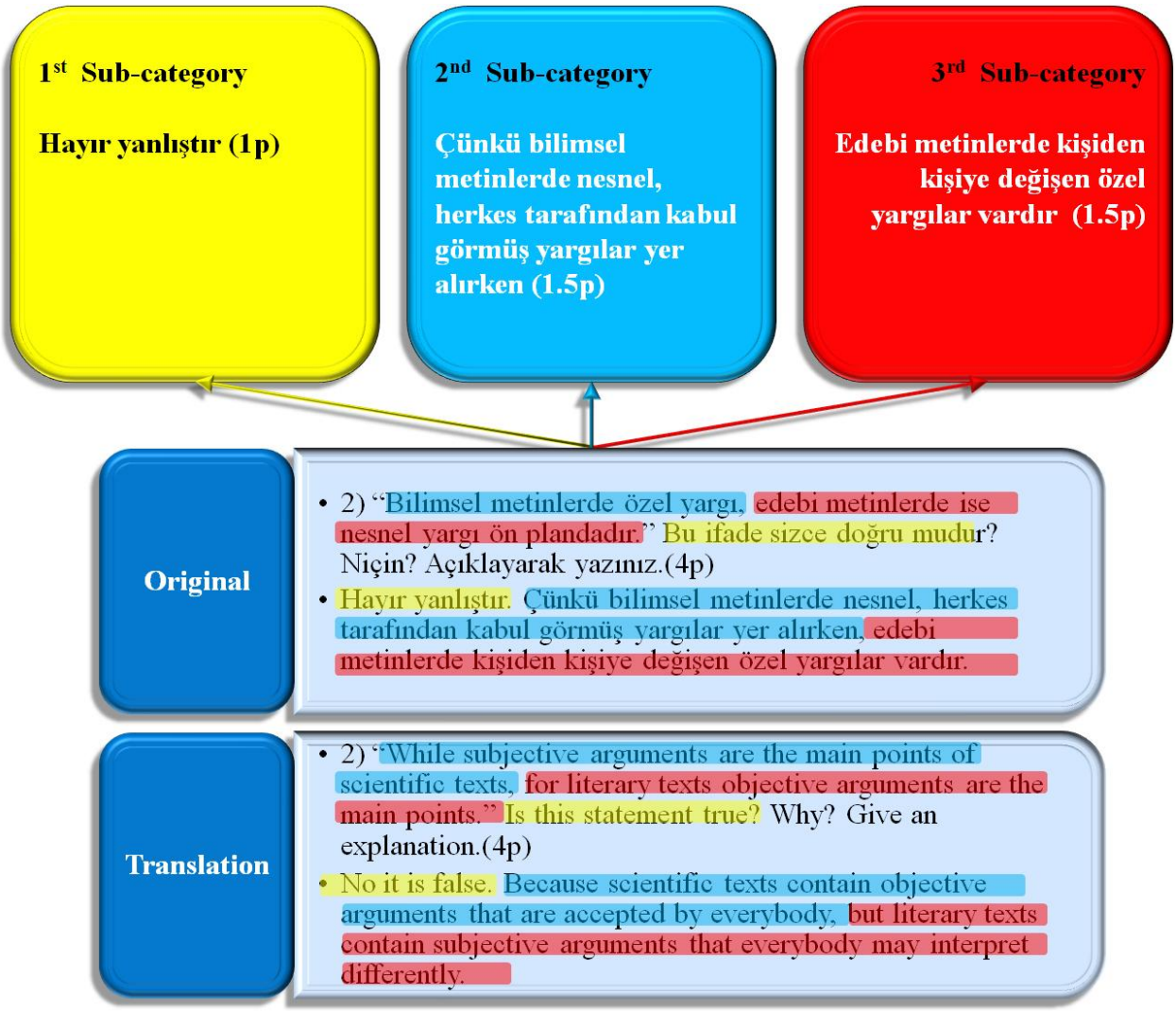


Figure 4.19: Sub-categories of the student answer

Extracted Features/Terms From the Student's Answer for the corresponding question/category:

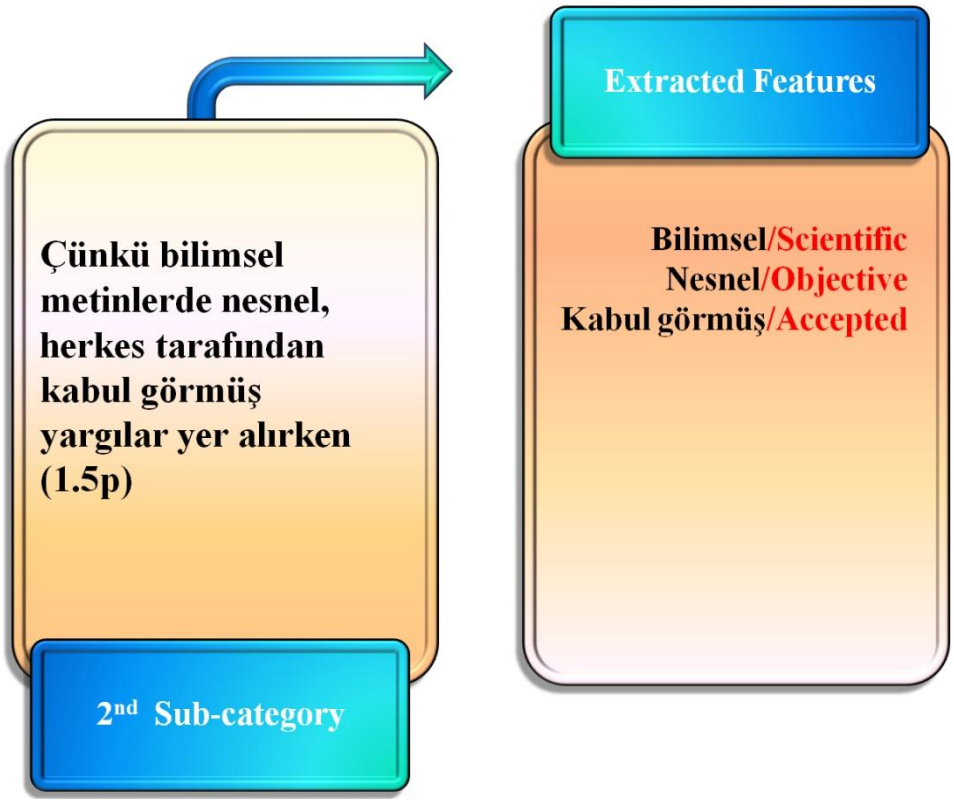


Figure 4.20: Extracted features from the seleted sub-category

In the next step we will get the intersection of the above terms list and the dictionary prepared for the category. Based on the grading method, just the correct answers will be counted and the incorrect answers will not be taken into account. Therefore a single element in the intersection set would be enough for the full credit (2p) for the category.

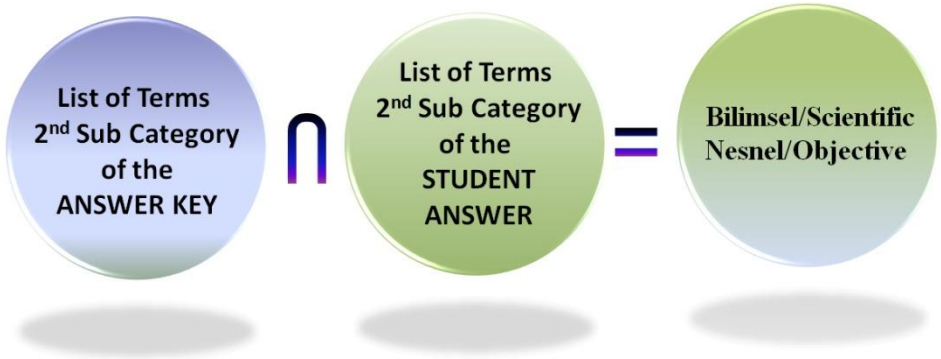


Figure 4.21: Intersection of the stemming lists

So, the student gets a full credit (2p) for the answer. The same process is repeated for the all categories and then the students overall grade is calculated based on the written answered questions. And in the final step, the original grade and the grade calculated by the text mining tool are compared. The comparison is the main measurement for evaluating the success of the algorithms.

In addition, during the processes of some categories we have got some issues such as homonym words, different words having the same stem and so on. For such issues, we give some applications of the algorithms for the case study. For the homonym terms, the corresponding algorithm is applied.

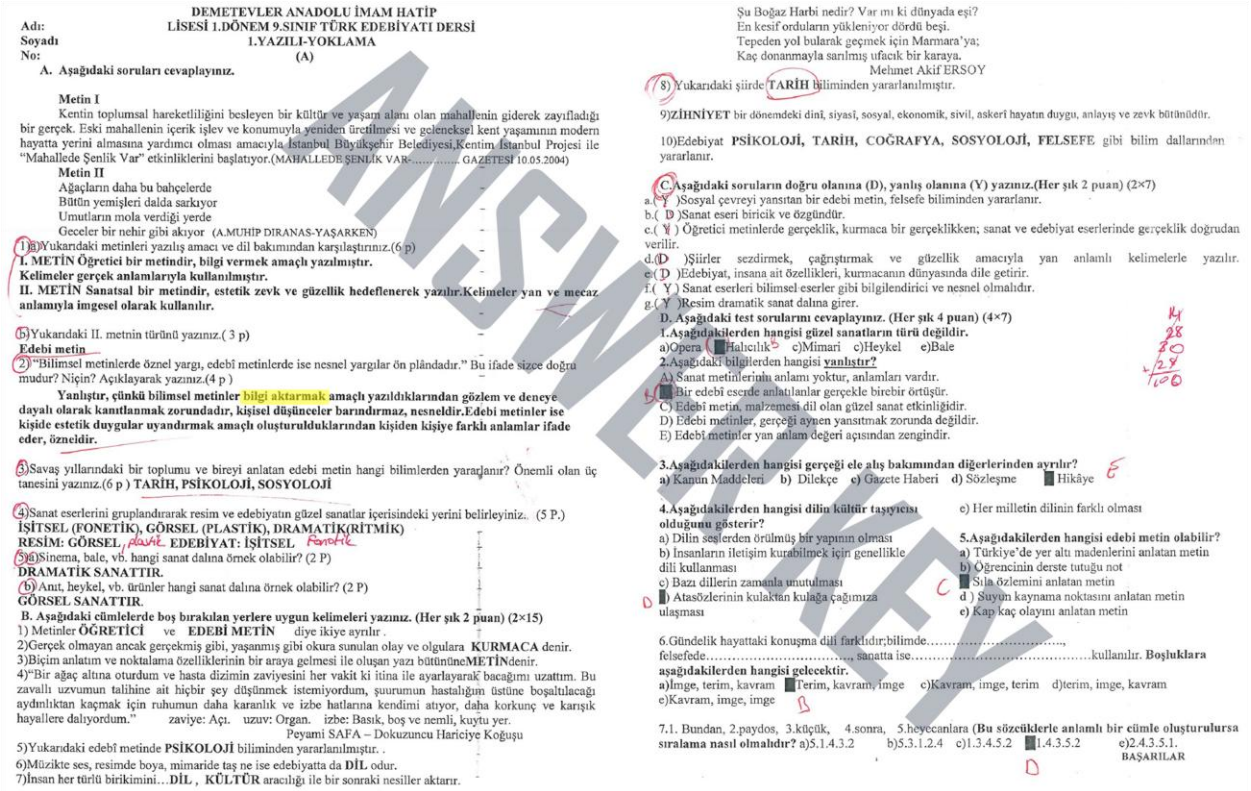


Figure 4.22: Answer key

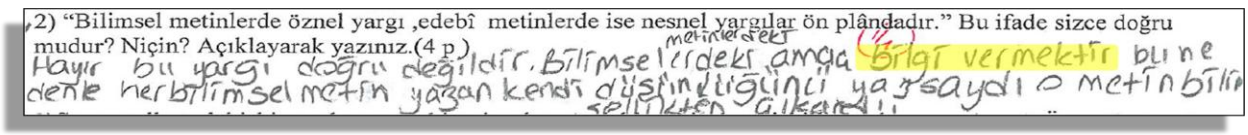


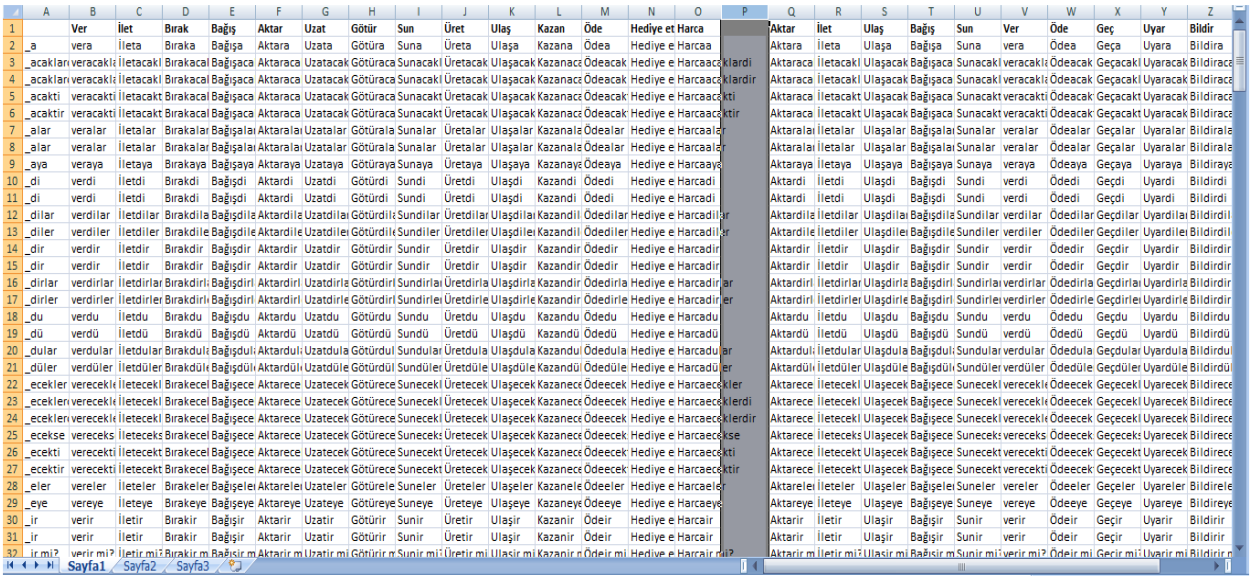
Figure 4.23: Student answer

All the significant features are extracted from the 4th main category (question) which is selected randomly. In the extracted features list, we have ‘bilgi aktarmak’ term. In the extracted features list of the student answer for the 4th main category, we have ‘bilgi vermek’ term. In the both terms ‘bilgi’ word is used in common. By using the following algorithm we identify if the ‘bilgi’ word is used with the same meaning in the both terms or they are homonym words.

Figure 4.24: Synonym and related words

In this case we have two phrases: ‘bilgi vermek’ and ‘bilgi aktarmak’. First of all we get all the synonym and related words of ‘vermek/to give’ and ‘aktarmak/to forward’ words. And

then as given in figure 4.25, we get all the stem+suffix matchings as we did in the previous example, and meaningful matchings are selected in a separated list (figure 4.26).



	Ver	İlet	Bırak	Bağış	Aktar	Uzat	Götür	Sun	Üret	Ulaş	Kazan	Öde	Hediye et	Harca
1														
2	_a	vera	İleta	Bıraka	Bağışa	Aktara	Uzata	Götüra	Suna	Üreta	Ulaşa	Kazana	Ödea	Hediye e	Harcaa
3	_acakları	veracakl	İletacakl	Bırakacal	Bağışaca	Aktaraca	Uzatacak	Götüraca	Sunacakl	Üretacak	Ulaşacak	Kazanacı	Ödeacak	Hediye e	Harcaacı
4	_acakları	veracakl	İletacakl	Bırakacal	Bağışaca	Aktaraca	Uzatacak	Götüraca	Sunacakl	Üretacak	Ulaşacak	Kazanacı	Ödeacak	Hediye e	Harcaacı
5	_acakti	veracakti	İletacakt	Bırakacal	Bağışaca	Aktaraca	Uzatacak	Götüraca	Sunacakt	Üretacak	Ulaşacak	Kazanacı	Ödeacak	Hediye e	Harcaacı
6	_acaktir	veracakti	İletacakt	Bırakacal	Bağışaca	Aktaraca	Uzatacak	Götüraca	Sunacakt	Üretacak	Ulaşacak	Kazanacı	Ödeacak	Hediye e	Harcaacı
7	_alar	veralar	İletalar	Bırakalar	Bağışalar	Aktaralar	Uzatalar	Götürala	Sunalar	Üretalar	Ulaşalar	Kazanala	Ödealar	Hediye e	Harcaaler
8	_alar	veralar	İletalar	Bırakalar	Bağışalar	Aktaralar	Uzatalar	Götürala	Sunalar	Üretalar	Ulaşalar	Kazanala	Ödealar	Hediye e	Harcaaler
9	_aya	veraya	İletaya	Bırakaya	Bağışaya	Aktaraya	Uzataya	Götüraya	Sunaya	Üretaya	Ulaşaya	Kazanaya	Ödeaya	Hediye e	Harcaaya
10	_di	verdi	İletdi	Bırakdi	Bağışdı	Aktardı	Uzatdı	Götürdı	Sundı	Üretdi	Ulaşdı	Kazandı	Ödedi	Hediye e	Harcadı
11	_di	verdi	İletdi	Bırakdi	Bağışdı	Aktardı	Uzatdı	Götürdı	Sundı	Üretdi	Ulaşdı	Kazandı	Ödedi	Hediye e	Harcadı
12	_dilar	verdilar	İletdilar	Bırakdıla	Bağışdıla	Aktardıla	Uzatdıla	Götürdıla	Sundılar	Üretdilar	Ulaşdıla	Kazandıl	Ödedilar	Hediye e	Harcadıl
13	_diler	verdiler	İletdiler	Bırakdıle	Bağışdıle	Aktardıle	Uzatdıle	Götürdıle	Sundıler	Üretdiler	Ulaşdıle	Kazandıl	Ödediler	Hediye e	Harcadıl
14	_dir	verdir	İletdir	Bırakdır	Bağışdır	Aktardır	Uzatdır	Götürdır	Sundır	Üretdir	Ulaşdır	Kazandır	Ödedir	Hediye e	Harcadır
15	_dir	verdir	İletdir	Bırakdır	Bağışdır	Aktardır	Uzatdır	Götürdır	Sundır	Üretdir	Ulaşdır	Kazandır	Ödedir	Hediye e	Harcadır
16	_dirlar	verdirlar	İletdirlar	Bırakdırla	Bağışdırl	Aktardırl	Uzatdırla	Götürdırl	Sundırla	Üretdirla	Ulaşdırla	Kazandır	Ödedirla	Hediye e	Harcadır
17	_dirler	verdirler	İletdirler	Bırakdırle	Bağışdırl	Aktardırl	Uzatdırle	Götürdırl	Sundırle	Üretdirle	Ulaşdırle	Kazandır	Ödedirle	Hediye e	Harcadır
18	_du	verdu	İletdu	Bırakdu	Bağışdu	Aktardu	Uzatdu	Götürdu	Sundu	Üretdu	Ulaşdu	Kazandu	Ödedu	Hediye e	Harcadu
19	_dü	verdü	İletdü	Bırakdü	Bağışdü	Aktardü	Uzatdü	Götürdü	Sundü	Üretdü	Ulaşdü	Kazandü	Ödedü	Hediye e	Harcadü
20	_dular	verdular	İletdular	Bırakdula	Bağışdul	Aktardula	Uzatdula	Götürdul	Sundular	Üretdula	Ulaşdula	Kazandu	Ödedula	Hediye e	Harcadu
21	_düler	verdüler	İletdüler	Bırakdüle	Bağışdül	Aktardüle	Uzatdüle	Götürdül	Sundüler	Üretdüle	Ulaşdüle	Kazandü	Ödedüle	Hediye e	Harcadü
22	_ecekler	verecekl	İletecekl	Bırakecel	Bağışece	Aktarece	Uzatecek	Götürece	Sunecekl	Üretecek	Ulaşecek	Kazaneca	Ödeecek	Hediye e	Harcaecı
23	_ecekleri	verecekl	İletecekl	Bırakecel	Bağışece	Aktarece	Uzatecek	Götürece	Sunecekl	Üretecek	Ulaşecek	Kazaneca	Ödeecek	Hediye e	Harcaecı
24	_ecekleri	verecekl	İletecekl	Bırakecel	Bağışece	Aktarece	Uzatecek	Götürece	Sunecekl	Üretecek	Ulaşecek	Kazaneca	Ödeecek	Hediye e	Harcaecı
25	_ecekse	vereceks	İleteceks	Bırakecel	Bağışece	Aktarece	Uzatecek	Götürece	Sunecekl	Üretecek	Ulaşecek	Kazaneca	Ödeecek	Hediye e	Harcaecı
26	_ecekti	verecekti	İletecekt	Bırakecel	Bağışece	Aktarece	Uzatecek	Götürece	Sunecekt	Üretecek	Ulaşecek	Kazaneca	Ödeecek	Hediye e	Harcaecı
27	_ecektir	verecekti	İletecekt	Bırakecel	Bağışece	Aktarece	Uzatecek	Götürece	Sunecekt	Üretecek	Ulaşecek	Kazaneca	Ödeecek	Hediye e	Harcaecı
28	_eler	vereler	İleteler	Bırakeler	Bağışeler	Aktareler	Uzateler	Götürele	Suneler	Üreteler	Ulaşeler	Kazanela	Ödeeler	Hediye e	Harcaeler
29	_eye	vereye	İleteye	Bırakeye	Bağışeye	Aktareye	Uzateye	Götüreye	Suneye	Üreteye	Ulaşeye	Kazaneya	Ödeeye	Hediye e	Harcaeya
30	_ir	verir	İletir	Bırakır	Bağışır	Aktarır	Uzatır	Götürır	Sunır	Üretir	Ulaşır	Kazanır	Ödeir	Hediye e	Harcaır
31	_ir	verir	İletir	Bırakır	Bağışır	Aktarır	Uzatır	Götürır	Sunır	Üretir	Ulaşır	Kazanır	Ödeir	Hediye e	Harcaır
32	ir mi?	verir mi?	İletir mi?	Bırakır m	Bağışır m	Aktarır mı	Uzatır mı	Götürır m	Sunır mı?	Üretir mi	Ulaşır mı	Kazanır n	Ödeir mi	Hediye e	Harcaır mı?

	Aktar	İlet	Ulaş	Bağış	Sun	Ver	Öde	Geç	Uyar	Bildir
1										
2	Aktara	İleta	Ulaşa	Bağışa	Suna	vera	Ödea	Geça	Uyara	Bildira
3	Aktaraca	İletacakl	Ulaşacak	Bağışaca	Sunacakl	veracakl	Ödeacak	Geçacakl	Uyaracak	Bildiraca
4	Aktaraca	İletacakl	Ulaşacak	Bağışaca	Sunacakl	veracakl	Ödeacak	Geçacakl	Uyaracak	Bildiraca
5	Aktaraca	İletacakt	Ulaşacak	Bağışaca	Sunacakt	veracakti	Ödeacak	Geçacakt	Uyaracak	Bildiraca
6	Aktaraca	İletacakt	Ulaşacak	Bağışaca	Sunacakt	veracakti	Ödeacak	Geçacakt	Uyaracak	Bildiraca
7	Aktaralar	İletalar	Ulaşalar	Bağışalar	Sunalar	veralar	Ödealar	Geçalar	Uyaralar	Bildirala
8	Aktaralar	İletalar	Ulaşalar	Bağışalar	Sunalar	veralar	Ödealar	Geçalar	Uyaralar	Bildirala
9	Aktaraya	İletaya	Ulaşaya	Bağışaya	Sunaya	veraya	Ödeaya	Geçaya	Uyaraya	Bildiraya
10	Aktardı	İletdi	Ulaşdı	Bağışdı	Sundı	verdi	Ödedi	Geçdi	Uyardı	Bildirdi
11	Aktardı	İletdi	Ulaşdı	Bağışdı	Sundı	verdi	Ödedi	Geçdi	Uyardı	Bildirdi
12	Aktardıla	İletdilar	Ulaşdıla	Bağışdıla	Sundılar	verdilar	Ödedilar	Geçdilar	Uyardıla	Bildirdil
13	Aktardıle	İletdiler	Ulaşdıle	Bağışdıle	Sundıler	verdiler	Ödediler	Geçdiler	Uyardıle	Bildirdil
14	Aktardır	İletdir	Ulaşdır	Bağışdır	Sundır	verdir	Ödedir	Geçdir	Uyardır	Bildirdir
15	Aktardır	İletdir	Ulaşdır	Bağışdır	Sundır	verdir	Ödedir	Geçdir	Uyardır	Bildirdir
16	Aktardırl	İletdirlar	Ulaşdırla	Bağışdırl	Sundırla	verdirlar	Ödedirla	Geçdirla	Uyardırla	Bildirdir
17	Aktardırl	İletdirler	Ulaşdırle	Bağışdırl	Sundırle	verdirler	Ödedirle	Geçdirle	Uyardırle	Bildirdir
18	Aktardu	İletdu	Ulaşdu	Bağışdu	Sundu	verdu	Ödedu	Geçdu	Uyardu	Bildirdu
19	Aktardü	İletdü	Ulaşdü	Bağışdü	Sundü	verdü	Ödedü	Geçdü	Uyardü	Bildirdü
20	Aktarduli	İletdular	Ulaşdula	Bağışdul	Sundular	verdular	Ödedula	Geçdular	Uyardula	Bildirdu
21	Aktardüle	İletdüler	Ulaşdüle	Bağışdül	Sundüler	verdüler	Ödedüle	Geçdüler	Uyardüle	Bildirdü
22	Aktarece	İletecekl	Ulaşecek	Bağışece	Sunecekl	verecekl	Ödeecek	Geçecekl	Uyarecek	Bildirece
23	Aktarece	İletecekl	Ulaşecek	Bağışece	Sunecekl	verecekl	Ödeecek	Geçecekl	Uyarecek	Bildirece
24	Aktarece	İletecekl	Ulaşecek	Bağışece	Sunecekl	verecekl	Ödeecek	Geçecekl	Uyarecek	Bildirece
25	Aktarece	İleteceks	Ulaşecek	Bağışece	Sunecekl	vereceks	Ödeecek	Geçecekl	Uyarecek	Bildirece
26	Aktarece	İletecekt	Ulaşecek	Bağışece	Sunecekt	verecekti	Ödeecek	Geçecekt	Uyarecek	Bildirece
27	Aktarece	İletecekt	Ulaşecek	Bağışece	Sunecekt	verecekti	Ödeecek	Geçecekt	Uyarecek	Bildirece
28	Aktareler	İleteler	Ulaşeler	Bağışeler	Suneler	vereler	Ödeeler	Geçeler	Uyareler	Bildirele
29	Aktareye	İleteye	Ulaşeye	Bağışeye	Suneye	vereye	Ödeeye	Geçeye	Uyareye	Bildireye
30	Aktarır	İletir	Ulaşır	Bağışır	Sunır	verir	Ödeir	Geçir	Uyarir	Bildirir
31	Aktarır	İletir	Ulaşır	Bağışır	Sunır	verir	Ödeir	Geçir	Uyarir	Bildirir
32	Aktarır m	İletir mi?	Ulaşır mı	Bağışır m	Sunır mı?	verir mi?	Ödeir mi	Geçir mi?	Uyarır mı	Bildirir n

Figure 4.25: Stem+suffix matchings

Ver,verdi,verdiler,verecekler,vereceklerdi,vereceklerdir,verecekse,verecekti,verecektir,vereler,verir,verirdi,verirdiyse,verirdiyeler,verirler,verirlerdi,verirlermis,verirlerse,verirmis,verirmis,verirse,vermeli,vermeliler,vermis,vermisdi,vermisdir,vermisler,vermislerdi,vermislerdir,vermismisse,vermismisseler,vermisse,vermisseler,vermisti,vermisti,versin,versinler
Ilet,Iletdi,Iletdiler,Iletecekler,Ileteceklerdi,Ileteceklerdir,Iletecekse,Iletecekti,Iletecektir,Iletir,Iletirdi,Iletirdiye,Iletirdiyeler,Iletirler,Iletirlerdi,Iletirlermis,Iletirlerse,Iletirmis,Iletirmis,Iletirse,Iletmeli,Iletmeliler,Iletmis,Iletmisdi,Iletmisdir,Iletmisler,Iletmislerdi,Iletmislerdir,Iletmismisse,Iletmismisseler,Iletmisse,Iletmisseler,Iletmisti,Iletsin,Iletsinler,Ietti,Iettiler,Ilettir
Birak,Birakacaklardi,Birakacaklardir,Birakacakti,Birakacaktir,Birakdi,Birakdilar,Birakdir,Birakir,Birakirdi,Birakirdiysa,Birakirdiysalar,Birakirlar,Birakirlardi,Birakirlarmis,Birakirlarsa,Birakirmis,Birakirsa,Birakmali,Birakmalilar,Birakmis,Birakmisdi,Birakmisdir,Birakmislar,Birakmislardi,Birakmismissa,Birakmismissalar,Birakmissa,Birakmissalar,Birakmisti,Biraksin,Biraksinlar,Birakti,Biraktilar,Biraktir
Bagis,Bagisa,Bagisdi,Bagisdir,Bagismis,Bagismisdi,Bagismisdir,Bagismissa,Bagismissalar,Bagismisti,Bagisti,Bagistir,Bagistirlar
Aktar,Aktaracaklardi,Aktaracaklardir,Aktaracakti,Aktaracaktir,Aktardi,Aktardilar,Aktarir,Aktarirdi,Aktarirlar,Aktarirlardi,Aktarirlarmis,Aktarirlarsa,Aktarmis,Aktarmisdi,Aktarmisdir,Aktarmislar,Aktarmislardi,Aktarmismissa,Aktarmismissalar,Aktarmissa,Aktarmissalar,Aktarmisti
Uzat,Uzatacaklardi,Uzatacaklardir,Uzatacakti,Uzatacaktir,Uzatdi,Uzatdilar,Uzatir,Uzatirdi,Uzatirdiysa,Uzatirdiysalar,Uzatirlar,Uzatirlardi,Uzatirlarmis,Uzatirlarsa,Uzatirmis,Uzatirsa,Uzatmali,Uzatmalilar,Uzatmis,Uzatmisdi,Uzatmisdi,Uzatmisdir,Uzatmislar,Uzatmislardi,Uzatmismissa,Uzatmismissalar,Uzatmissa,Uzatmissalar,Uzatmisti,Uzatsin,Uzatsinlar,Uzatti,Uzattilar,Uzattir
Götür,Götürdü,Götürdüler,Götürecekler,Götüreceklerdi,Götüreceklerdir,Götürecekse,Götürecekti,Götürecektir,Götürmeli,Götürmeliler,Götürmüs,Götürmüsdü,Götürmüsdür,Götürmüsler,Götürmüsmüsse,
Götürmüsmüsseler,Götürmüsse,Götürmüstü,Götürür,Götürürdü,Götürürdüyse,Götürürdüyseler,Götürürmüs
Sun,Sunacaklardi,Sunacaklardir,Sunacakti,Sunacaktir,Sundu,Sundular,Sunmali,Sunmalilar,Sunmus,Sunmusdu,Sunmusdur,Sunmuslar,Sunmusmussa,Sunmusmussalar,Sunmustu
Üret,Üretdi,Üretdiler,Üretecekler,Üreteceklerdi,Üreteceklerdir,Üretecekse,Üretecekti,Üretecektir,Üretir,Üretirdi,Üretirdiye,Üretirdiyeler,Üretirler,Üretirlerdi,Üretirlermis,Üretirlerse,Üretirmis,Üretirse,Üretmeli,Üretmeliler,Üretmis,Üretmisdi,Üretmisdir,Üretmisler,Üretmislerdi,Üretmislerdir,Üretmismisse,Üretmismisseler,Üretmisse,Üretmisseler,Üretmisti,Üretmiyecektir,Üretsin,Üretsinler,Üretti,Ürettiler,Ürettir
Ulas,Ulasacaklardi,Ulasacaklardir,Ulasacakti,Ulasacaktir,Ulasdi,Ulasdilar,Ulasdir,Ulasir,Ulasirdi,Ulasirdiysa,Ulasirdiysalar,Ulasirlar,Ulasirlardi,Ulasirlarmis,Ulasirlarsa,Ulasirmis,Ulasirsa,Ulasmali,Ulasmalilar,Ulasmis,Ulasmisdi,Ulasmisdir,Ulasmislar,Ulasmislardi,Ulasmissa,Ulasmissalar,Ulasmisti,Ulassin,Ulassinler,Ulasti,Ulastilar,Ulastir
Kazan,Kazanacaklardi,Kazanacaklardir,Kazanacakti,Kazanacaktir,Kazandi,Kazandilar,Kazandir,Kazanir,Kazanirdi,Kazanirlar,Kazanirlardi,Kazanirlarmis,Kazanirlarsa,Kazanirmis,Kazanirsa,Kazanmali,Kazanmalilar,Kazanmis,Kazanmisdi,Kazanmisdir,Kazanmislar,Kazanmislardi,Kazanmissa,Kazanmissalar,Kazanmisti
Öde,Ödedi,Ödediler,Ödemeli,Ödemeliler,Ödemeye,Ödemis,Ödemisdi,Ödemisdir,Ödemisler,Ödemislerdi,Ödemislerdir,Ödemismisse,Ödemismisseler,Ödemisse,Ödemisseler,Ödemisti,Öderler,Öderlerdi,Öderlermis,Öderlerse,Ödermis,Öderse,Ödesin,Ödesinler,Ödetir,Ödetirler,Ödeyecekler,Ödeyeceklerdi,Ödeyeceklerdir,Ödeyecekse,Ödeyecekti,Ödeyecektir
Hediye et,Hediye etdi,Hediye etdiler,Hediye etdir,Hediye etmis,Hediye etmisdi,Hediye etmisdir,Hediye etmisler,Hediye etmislerdi,Hediye etmislerdir,Hediye etmismisse,Hediye etmismisseler,Hediye etmisse,Hediye etmisseler,Hediye etmisti,
Harca,Harcadi,Harcadilar,Harcamali,Harcamalilar,Harcamis,Harcamisdi,Harcamisdir,Harcamislar,Harcamislardi,Harcamismissalar,Harcamissa,Harcamissalar,Harcamisti,Harcar,Harcardi,Harcarlar,Harcarlardi,Harcarlarmis,Harcarlarsa,Harcarmis,Harcarsa,Harcasin,Harcasinlar,Harcatir,Harcatirlar,Harcayacaklardi,Harcayacaklardir,Harcayacakti,Harcayacaktir

Aktar,Aktaracaklardi,Aktaracaklardir,Aktaracakti,Aktaracaktir,Aktardi,Aktardilar,Aktarir,Aktarirdi,Aktarirlar,Aktarirlardi,Aktarirlarmis,Aktarirlarsa,Aktarmis,Aktarmisdi,Aktarmisdir,Aktarmislar,Aktarmislardi,Aktarmismissa,Aktarmismissalar,Aktarmissa,Aktarmissalar,Aktarmisti
Ilet,Iletdi,Iletdiler,Iletecekler,Ileteceklerdi,Ileteceklerdir,Iletecekse,Iletecekti,Iletecektir,Iletir,Iletirdi,Iletirdiye,Iletirdiyeler,Iletirler,Iletirlerdi,Iletirlermis,Iletirlerse,Iletirmis,Iletirmis,Iletirse,Iletmeli,Iletmeliler,Iletmis,Iletmisdi,Iletmisdir,Iletmisler,Iletmislerdi,Iletmislerdir,Iletmismisse,Iletmismisseler,Iletmisse,Iletmisseler,Iletmisti,Iletsin,Iletsinler,Ietti,Iettiler,Ilettir
Ulas,Ulasacaklardi,Ulasacaklardir,Ulasacakti,Ulasacaktir,Ulasdi,Ulasdilar,Ulasdir,Ulasir,Ulasirdi,Ulasirdiysa,Ulasirdiysalar,Ulasirlar,Ulasirlardi,Ulasirlarmis,Ulasirlarsa,Ulasirmis,Ulasirsa,Ulasmali,Ulasmalilar,Ulasmis,Ulasmisdi,Ulasmisdir,Ulasmislar,Ulasmislardi,Ulasmissa,Ulasmissalar,Ulasmisti,Ulassin,Ulassinler,Ulasti,Ulastilar,Ulastir
Bagis,Bagisa,Bagisdi,Bagisdir,Bagismis,Bagismisdi,Bagismisdir,Bagismissa,Bagismissalar,Bagismisti,Bagisti,Bagistir,Bagistirlar
Sun,Sunacaklardi,Sunacaklardir,Sunacakti,Sunacaktir,Sundu,Sundular,Sunmali,Sunmalilar,Sunmus,Sunmusdu,Sunmusdur,Sunmuslar,Sunmusmussa,Sunmusmussalar,Sunmustu
Ver,verdi,verdiler,verecekler,vereceklerdi,vereceklerdir,verecekse,verecekti,verecektir,vereler,verir,verirdi,verirdiye,verirdiyeler,verirler,verirlerdi,verirlermis,verirlerse,verirmis,verirmis,verirse,vermeli,vermeliler,vermis,vermisdi,vermisdir,vermisler,vermislerdi,vermislerdir,vermismisse,vermismisseler,vermisse,vermisseler,vermisti,vermisti,versin,versinler
Öde,Ödedi,Ödediler,Ödemeli,Ödemeliler,Ödemeye,Ödemis,Ödemisdi,Ödemisdir,Ödemisler,Ödemislerdi,Ödemislerdir,Ödemismisse,Ödemismisseler,Ödemisse,Ödemisseler,Ödemisti,Öderler,Öderlerdi,Öderlermis,Öderlerse,Ödermis,Öderse,Ödesin,Ödesinler,Ödetir,Ödetirler,Ödeyecekler,Ödeyeceklerdi,Ödeyeceklerdir,Ödeyecekse,Ödeyecekti,Ödeyecektir
Geç,geçdi,geçdiler,geçdir,geçdirler,geçecekler,geçeceklerdi,geçeceklerdir,geçecekse,geçecekti,geçecektir,geçir,geçirdi,geçirdiye,geçirdiyeler,geçirmis,geçirse,geçmeli,geçmeliler,geçmeye,geçmis,geçmisdi,geçmisdir,geçmisler,geçmislerdi,geçmislerdir,geçmismisse,geçmismisseler,geçmisse,geçmisseler,geçmisti,geçsin,geçsinler,geçti,geçtiler,geçtir
Uyar,Uyaracaklardi,Uyaracaklardir,Uyaracakti,Uyaracaktir,Uyardi,Uyardilar,Uyarir,Uyarirdi,Uyarirdiysa,Uyarirdiysalar,Uyarirlar,Uyarirlardi,Uyarirlarmis,Uyarirlarsa,Uyarirmis,Uyarirsa,Uyarmali,Uyarmalilar,Uyarmis,Uyarmisdi,Uyarmisdir,Uyarmislar,Uyarmislardi,Uyarmismissa,Uyarmismissalar,Uyarmissa,Uyarmissalar,Uyarmisti
Bildir,Bildirdiler,Bildirecekler,Bildireceklerdi,Bildireceklerdir,Bildirecekse,Bildirecekti,Bildirecektir,Bildirir,Bildirirdi,Bildirirdiye,Bildirirdiyeler,Bildirirler,Bildirirlerdi,Bildirirlermis,Bildirirlerse,Bildirirmis,Bildirirse,Bildirmeli,Bildirmeliler,Bildirmis,Bildirmisdi,Bildirmisdir,Bildirmisler,Bildirmislerdi,Bildirmislerdir,Bildirmismisse,Bildirmismisseler,Bildirmisse,Bildirmisseler,Bildirmisti,Bildirmiyeceklerdir,Bildirmiyecektir,Bildirsin,Bildirsinler,Bildirtir,Bildirtirler
Alinti,Alintidilar,Alintidir,Alintidirlar,Alintidirler,Alintimali,Alintimalilar,Alintiyacaklardi,Alintiyacaklardir,Alintiyacakti,Alintiyacaktir
Transfer,Transferdi,Transferdir,Transferdirler,Transfermis,Transfermisdi,Transfermisdir,Transfermisler,Transfermismisse,Transfermismisseler,Transfermisse,Transfermisseler,Transfermisti

Figure 4.26: Selected stem+suffix matchings

Since the intersection of the two stemming lists is not empty, the ‘bilgi’ words used in the two phrases are used with the same meaning. The quantity of the elements in the intersection set supports this idea.

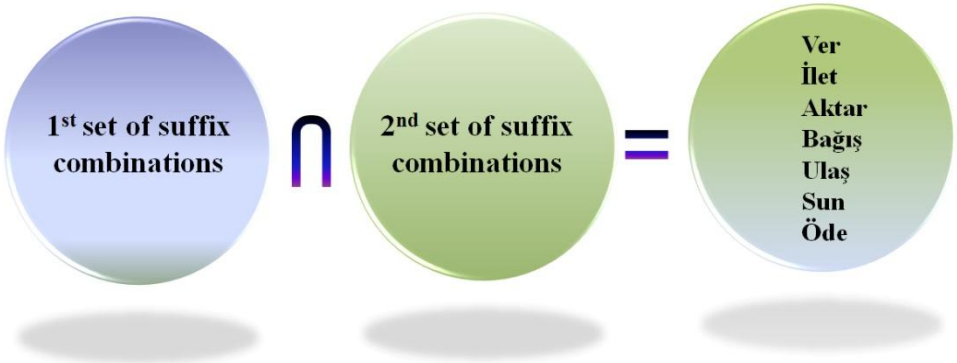


Figure 4.27: Intersection of the stemming lists

Therefore, the ‘bilgi’ words used in the two sentences are the same. In other words, they are not homonyms.

In Turkish even some words have the same stem, they have different meanings with different suffixes: **Göz**lemek/**Göz**lük = To wait/Glasses.

Original copy of the answer key:

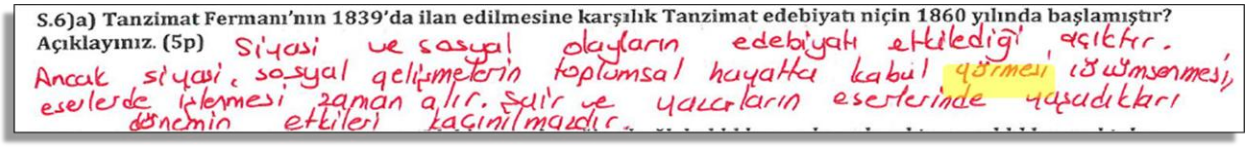

S.6)a) Tanzimat Fermanı'nın 1839'da ilan edilmesine karşılık Tanzimat edebiyatı niçin 1860 yılında başlamıştır? Açıklayınız. (5p) Siyasi ve sosyal olayların edebiyatı etkilediği açıktır. Ancak siyasi, sosyal gelişmelerin toplumsal hayatta kabul görmesi, özümsenmesi, eserlerde işlenmesi zaman alır. Şair ve yazarların eserlerinde yaşadıkları dönemin etkileri kaçınılmazdır.

Figure 4.28: Selected category from the answer key

Some words even have the same stem, the meaning of them are different with different suffixes. If we detect any situation like this, then we need to prepare separated stemming lists for the words for preventing the misunderstanding in the final grading stage.

Figure 4.28 is an original part of the answer key and figure 4.30 is an original part of the student answer paper. The terms ‘görmesi’ and ‘göreneklerinden’ from the texts have a similar problem described above. Both of them have the same stem: ‘gör’. We will identify if they have different meanings with different suffixes or not by using the following algorithm.

Original	• 6) Tanzimat Fermanı'nın 1839'da ilan edilmesine karşılık Tanzimat edebiyatı niçin 1860 yılında başlamıştır? Açıklayınız. (5p) • Siyasi ve sosyal olayların edebiyatı etkilediği açıktır. Ancak siyasi, sosyal gelişmelerin toplumsal hayatta kabul **görmesi**, özümsenmesi,eserlerde işlenmesi zaman alır. Şair ve yazarların eserlerinde yaşadıkları dönemin etkileri kaçınılmazdır.
Translation	• 6) Why Even though Tanzimat Edict was issued in 1839, Tanzimat literature started in 1860? Explain. (5p) • It is clear that political and social events affect the literature. However **to be accepted** and assimilated of these events in the social life and to be used in the literature is a matter of time. The era when the poets and authors lived has inevitable influence on the works of the poets and the authors.

Figure 4.29: Original and translated versions of the selected category

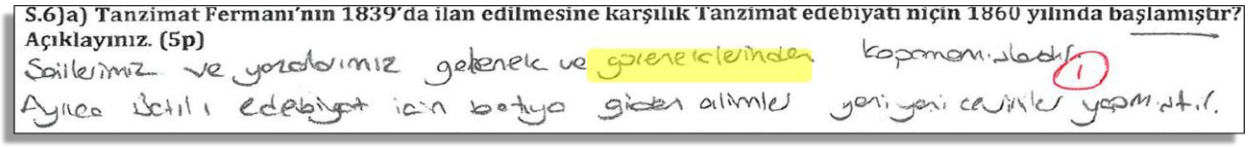


Figure 4.30: Student answer

All the synonym and related words of the terms are obtained from the TLA-BTD without making stem+suffix separation. Figure 4.32 is the list of the synonym and related words. Since the intersection of the two lists is empty, the terms have different meanings. Therefore even if the terms have the same stem, we need to prepare stemming lists for each one separately.

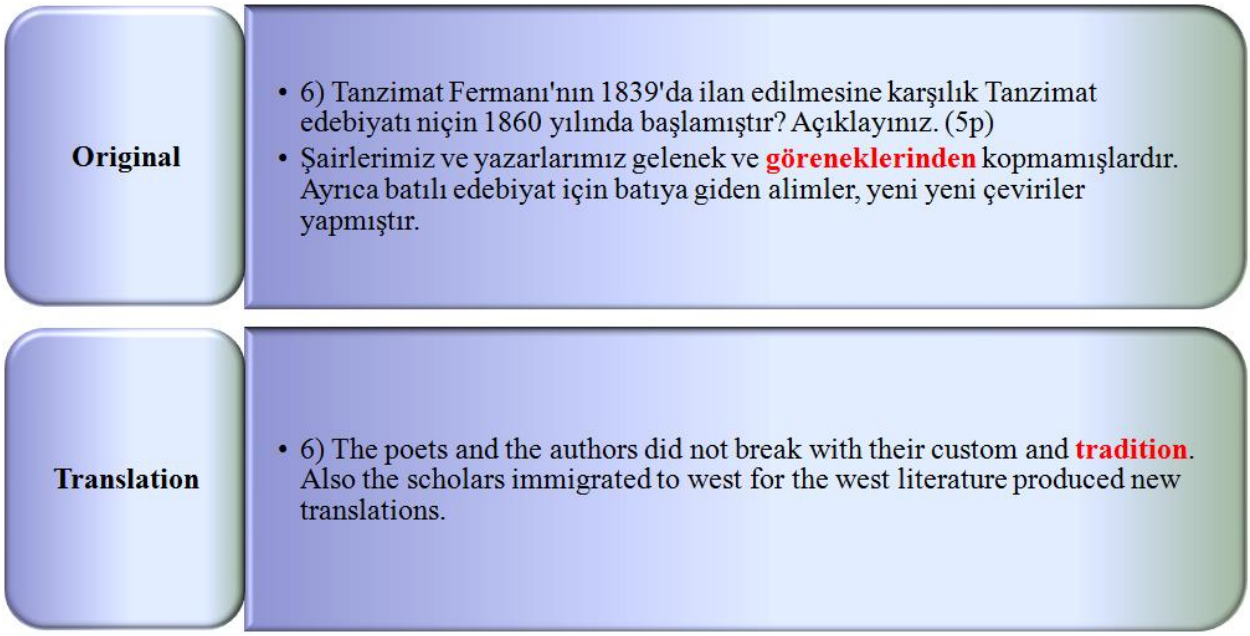


Figure 4.31: Original and translated versions of the student answer

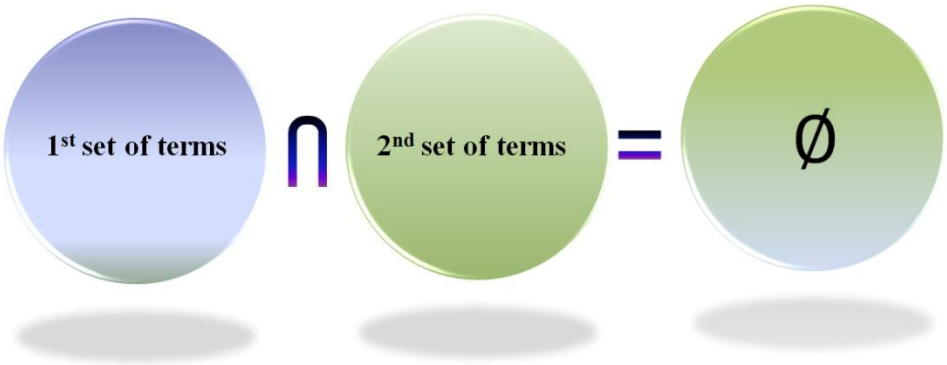


Figure 4.32: Intersection of the synonyms and related words lists

These words have different meanings even if they have the same stem, and so we need to prepare two different stem clusters for them.

4.5 Results

Student_ID	Grade_TM48	Grade_48	Error	Abs_Err	Ratio
1	**27**	**26**	-0.038	**0.038**	1.038
2	**29**	**30**	0.033	**0.033**	0.966
3	**32**	**37**	0.135	**0.135**	0.864

4	**25**	**24**	-0.041	**0.041**	1.041
5	**36**	**38**	0.052	**0.052**	0.947
6	**29**	**27**	-0.074	**0.074**	1.074
7	**28**	**33**	0.151	**0.151**	0.848
8	**27**	**29**	0.068	**0.068**	0.931
9	**31**	**31**	0.000	**0.000**	1.000
10	**25**	**32**	0.218	**0.218**	0.781
11	**41**	**37**	-0.108	**0.108**	1.108
12	**44**	**44**	0.000	**0.000**	1.000
13	**38**	**36**	-0.055	**0.055**	1.055
14	**33**	**27**	-0.222	**0.222**	1.222
15	**39**	**39**	0.000	**0.000**	1.000
16	**35**	**33**	-0.060	**0.060**	1.060
17	**30**	**42**	0.285	**0.285**	0.714
18	**42**	**44**	0.045	**0.045**	0.954
19	**35**	**37**	0.054	**0.054**	0.945
20	**31**	**36**	0.138	**0.138**	0.861
21	**29**	**36**	0.194	**0.194**	0.805
Mean	**32.667**	**34.190**	0.037	**0.094**	0.962

Table 4.1: Recalculated grades and comparison

The dictionaries created by using the developed algorithms are uploaded to the text miner tool for the each category separately, and the grade is calculated by the tool for the category. And then, all the grades are compiled together and for each student a general grade is calculated. There were some multiple choice questions in the original exam papers. Since we only focus on the written parts, we eliminated the multiple choice questions and then recalculated the students' grades just based on the written parts. The total grade of the written parts is 48. In the Table 4.1 column Grade_48 has the total grades of the students based on the original grades, Grade_TM_48 has the total grades of the students calculated by the text mining algorithm. Ratio = Grade_TM48/Grade_48, Error = 1 – Ratio, and Abs_Err = Absolute (Error). Some grades calculated by the algorithm are less than the original grades and some of them are more than the original grades. Therefore we have some positive and negative error scores. Abs_Err has the absolute error scores which is the most reliable measurement for assessing the success of the

algorithm. The mean Abs_Err is 0.09427. In other words, the mean of absolute accuracy of the algorithm is 1-0.09427=0.90573.

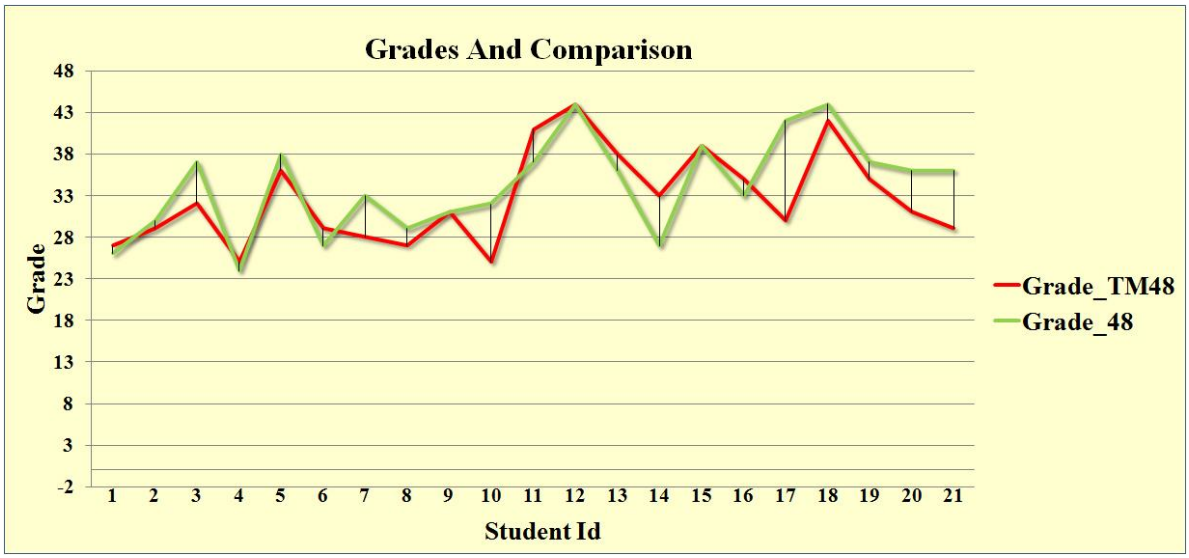


Figure 4.33: Comparison of the results

As given in Figure 4.33, the original and algorithm calculated grades have similar distribution. This shows the success of the algorithm. Also, the Figure 4.34 has the distribution of the absolute errors. The maximum error rate is 0.3, and also there are some 0 error rates which mean that the algorithm calculated exactly the same grades with the original grades.

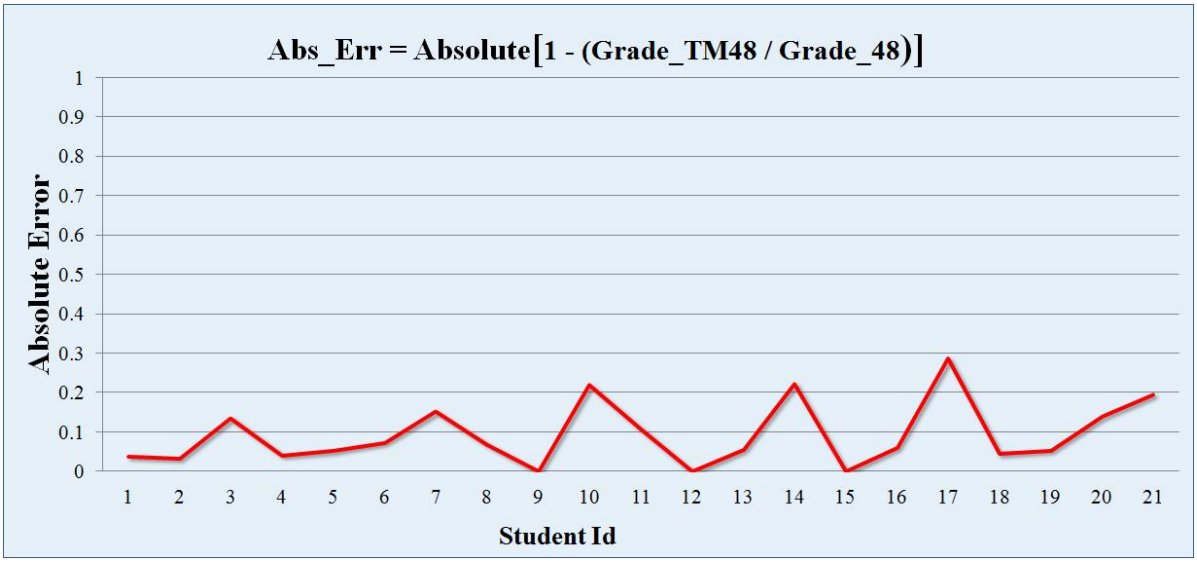


Figure 4.34: Absolute error rates

4.6 Future Work: Open-ended questions

In the exam papers in the data set, there is no open-ended question but it is another difficulty that researchers may have during the creation of the algorithm. Since the answer of such questions is completely based on the personal opinions, there are infinitely many possibilities for the correct answer. And so, the dictionary of such a question would include almost infinitely many words/phrases. Also, to create dictionaries for such questions is very time-consuming and needs so much energy and a teamwork. Therefore, the accuracy rates of the algorithms based on these kind questions would be not as good as the algorithms based on the close-ended questions. The only hint based on that a dictionary can be created is the key words in the question. Following is an open-ended question example.

Example 4.3: Mutluluk nedir?/What is happiness?

Mutluluk/Happiness is the only significant word in the question but millions of books can be written about it.

For such a word the dictionary would be as following:

Mutluluk
Kıvanç
Saadet
Bahtiyarlık
Huzur
Kut
Ongunluk
Neşe
Gülümsemek
Tebessüm

These words are the synonymous or related words of ‘mutluluk’. A dictionary based on ‘mutluluk’ would consists of these words and their all the suffix combinations.

“Mutluluk, ağlayabilmektir bir dostun derdiyle”/”Happiness is to be able to cry for a friend’s misery”

The person who wrote this sentence explains the happiness with these words:

Ağlamak/To cry
Dost/Friend
Dert/Misery

However, none of them is matching to a word in the dictionary prepared for the question. Also any suffix combinations of the words do not match any word in the dictionary prepared for the question. So the regular algorithms do not work so good for the open-ended questions/answers.

Bibliography

[1] Hogenboom A, Heerschop B, Frasincar F, Kaymak U, de Jong F. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision support systems. 2014;62:43-53.

[2] Lau RYK, Li C, Liao SSY. Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. Decision Support Systems. 2014;65:80-94.

[3] Wang H, Wang W. Product weakness finder: an opinion-aware system through sentiment analysis. Industrial Management & Data Systems. 2014;114:1301-20.

[4] Yee Liau B, Pei Tan P. Gaining customer knowledge in low cost airlines through text mining. Industrial Management & Data Systems. 2014;114:1344-59.

[5] Yu Y, Duan W, Cao Q. The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision Support Systems. 2013;55:919-26.

[6] Hatzivassiloglou V, McKeown KR. Predicting the semantic orientation of adjectives.Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics: Association for Computational Linguistics; 1997. p. 174-81.

[7] Turney P, Littman ML. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. 2002.

[8] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis. Proceedings of the 14th ACM international conference on Information and knowledge management: ACM; 2005. p. 625-31.

[9] Dag A. MMF, Oztekin A., Chen Y., Yucel A. A Hybrid Data Analytic Approach to Predict Heart Transplant Success. European Journal of Operational Research (under second revision). 2014.

[10] Dag A. TK, Megahed M. F., Oztekin A. A Probabilistic Data-Driven Methodology to Score Heart Transplant Survival. DEcision Support Systems (Under second revision). 2015.

[11] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques.Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10: Association for Computational Linguistics; 2002. p. 79-86.

[12] Genc O. DA. A Bayesian Network-Based Data Analytical Approach to Predict Velocity Distribution in Small Streams. Journal of Hydroinformatics (under second revision). 2015.

[13] Genc O. DA. A Data Mining Based Approach to Predict the Velocity Profiles in Small Streams. Water Resources Management (under second revision). 2014.

[14] Dag A. GO, Gonen B., Dinc S. Developing an Automated Tool to Predict Point Velocities in Small Streams via Artificial Neural Networks. Soft Computing (under revision). 2015.

[15] Da Silva NFF, Hruschka ER, Hruschka ER. Tweet sentiment analysis with classifier ensembles. Decision Support Systems. 2014; 66:170-9.

[16] Fersini E, Messina E, Pozzi FA. Sentiment analysis: Bayesian Ensemble Learning. Decision Support Systems. 2014; 68:26-38.

[17] Mullen T, Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources.Emnlp 2004. p. 412-8.

[18] Sarkar S, Mallick P, Banerjee A. A Real-Time Machine Learning Approach for Sentiment Analysis.Information Systems Design and Intelligent Applications: Springer; 2015. p. 705-17.

[19] https://wordnet.princeton.edu

[20] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information.Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1: Association for Computational Linguistics; 2003. p. 149-56.

[21] Kamps J, Marx M, Mokken RJ, De Rijke M. Using WordNet to Measure Semantic Orientations of Adjectives.Lrec: Citeseer; 2004. p. 1115-8.

[22] Ding X, Liu B. The utility of linguistic rules in opinion mining.Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval: ACM; 2007. p. 811-2.

[23] Kaji N, Kitsuregawa M. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents.EMNLP-CoNLL2007. p. 1075-83.

[24] Turney PD. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.Proceedings of the 40th annual meeting on association for computational linguistics: Association for Computational Linguistics; 2002. p. 417-24.

[25] Li G, Liu F. A clustering-based approach on sentiment analysis.Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on: IEEE; 2010. p. 331-7.

[26] Chaovalit P, Zhou L. Movie review mining: A comparison between supervised and unsupervised classification approaches.System Sciences, 2005 HICSS'05 Proceedings of the 38th Annual Hawaii International Conference on: IEEE; 2005. p. 112c-c.

[27] Boiy E, Moens M-F. A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval. 2009; 12:526-58.

[28] Prabowo R, Thelwall M. Sentiment analysis: A combined approach. Journal of Informetrics. 2009; 3:143-57.

[29] Ye Q, Zhang Z, Law R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications. 2009; 36:6527-35.

[30] Wang H, Lu Y, Zhai C. Latent aspect rating analysis on review text data: a rating regression approach.Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining: ACM; 2010. p. 783-92.

[31] Bjørkelund E, Burnett TH, Nørvåg K. A study of opinion mining and visualization of hotel reviews.Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services: ACM; 2012. p. 229-38.

[32] http://www.tripadvisor.com/PressCenter-c6-About_Us.html

[33] http://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html

[34] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection.Ijcai1995. p. 1137-45.

[35] Olson DL, Delen D. Advanced data mining techniques: Springer Science & Business Media; 2008.

[36] Han J, Kamber M, Pei J. Data mining: concepts and techniques: concepts and techniques: Elsevier; 2011.

[37] Bespalov D, Bai B, Qi Y, Shokoufandeh A. Sentiment classification based on supervised latent n-gram analysis.Proceedings of the 20th ACM international conference on Information and knowledge management: ACM; 2011. p. 375-82.

[38] Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews.Aaai2006. p. 1265-70.

[39] Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: A comparative study. Decision Support Systems. 2011; 50:602-13.

[40] Breiman L. Random forests. Machine learning. 2001; 45:5-32.

[41] Oztekin A, Kong ZJ, Delen D. Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations. Decision Support Systems. 2011; 51:155-66.

[42] Kuzey C, Uyar A, Delen D. The impact of multinationality on firm value: A comparative analysis of machine learning techniques. Decision Support Systems. 2014; 59:127-42.

[43] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC press; 1984.

[44] Quinlan JR. Induction of decision trees. Machine learning. 1986; 1:81-106.

[45] Quinlan JR. C4. 5: programs for machine learning: Elsevier; 2014.

[46] Quinlan JR. Bagging, boosting, and C4. 5.AAAI/IAAI, Vol 11996. p. 725-30.

[47] Roe BP, Yanga H, Zhub JI. Boosted decision trees, a powerful event classifier. Signal. 2005; 30:50.

[48] Misiunas N, Oztekin A, Chen Y, Chandra K. DEANN: A Healthcare Analytic Methodology of Data Envelopment Analysis and Artificial Neural Networks for the Prediction of Organ Recipient Functional Status. Omega. 2015.

[49] http://documentation.statsoft.com/STATISTICAHelp.aspx?path=TextMiner (Singular Value Decompositionin STATISTICA TextMining and Document Retrieval)

[50] http://cs-www.cs.yale.edu/homes/mmahoney/pubs/kdd07.pdf

[51] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

[52] Stanton, Angela D'Auria, Irvine Clarke III, and Wilbur W. Stanton. "Guidelines for Assessing and Establishing Effective Questionnaires in a Multicultural Context." Proceedings of the 1996 Multicultural Marketing Conference. Springer International Publishing, 2015.

[53] Woodruff, Robert B. "Customer value: the next source for competitive advantage." Journal of the academy of marketing science 25.2 (1997): 139-153.

[54] Anstead, Nick, and Ben O'Loughlin. "Social media analysis and public opinion: the 2010 UK General Election." Journal of Computer-Mediated Communication 20.2 (2015): 204-220.

[55] Chalothom, Tawunrat, and Jeremy Ellman. "Simple Approaches of Sentiment Analysis via Ensemble Learning." Information Science and Applications. Springer Berlin Heidelberg, 2015. 631-639.

[56] Desai, Jayraj M., and Swapnil R. Andhariya. "Sentiment analysis approach to adapt a shallow parsing based sentiment lexicon." Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on. IEEE, 2015.

[57] Denecke, Kerstin, and Yihan Deng. "Sentiment analysis in medical settings: New opportunities and challenges." Artificial intelligence in medicine (2015).

[58] Agarwal, B., Mittal, N., Bansal, P., & Garg, S. Sentiment Analysis Using Common-Sense and Context Information. Computational intelligence and neuroscience, 2015.

[59] Bignami, Francesca. "European versus American liberty: a comparative privacy analysis of anti-terrorism data-mining." Boston College Law Review 48 (2007): 609.

[60] Guzik, Keith. "Discrimination by Design: Data Mining in the United States's 'War on Terrorism'." Surveillance & Society 7.1 (2009): 3-20.

[61] Wilkins, Hugh. "Using importance-performance analysis to appreciate satisfaction in hotels." Journal of Hospitality Marketing & Management 19, no. 8 (2010): 866-888.

[62] Ekiz, E., Khoo-Lattimore, C., & Memarzadeh, F. (2012). Air the anger: Investigating online complaints on luxury hotels. Journal of Hospitality and Tourism Technology, 3(2), 96–106.

[63] Schuckert, Markus, Xianwei Liu, and Rob Law. "Hospitality and tourism online reviews: Recent trends and future directions." Journal of Travel & Tourism Marketing 32.5 (2015): 608-621.

[64] Li, Xinxin, and Lorin M. Hitt. "Self-selection and information role of online product reviews." Information Systems Research 19.4 (2008): 456-474.

[65] Mauri, Aurelio G., and Roberta Minazzi. "Web reviews influence on expectations and purchasing intentions of hotel potential customers."International Journal of Hospitality Management 34 (2013): 99-107.

[66] Moutinho, Luiz. "Consumer behaviour in tourism." European journal of marketing 21.10 (1987): 5-44.

[67] Kozak, Metin. "Repeaters' behavior at two distinct destinations." Annals of tourism research 28.3 (2001): 784-807.

[68] Yoon, Yooshik, and Muzaffer Uysal. "An examination of the effects of motivation and satisfaction on destination loyalty: a structural model." Tourism management 26.1 (2005): 45-56.

[69] Pileliene, Lina, and Viktorija Grigaliunaite. "Lithuanian tourist satisfaction index model." (2014).

[70] Ying, Li. "An Analysis of Tourists' Satisfaction and Influencing Factors in Tourist Destinations——Taking Xi'an Domestic Tourism Market as An ExampleJ." Tourism Tribune 4 (2008): 43-48.

[71] World Travel and Tourism Council http://www.wttc.org/

[72] Richter, Linda K. The politics of tourism in Asia. University of Hawaii Press, 1989.

[73] Nikhil, R., et al. "A Survey on Text Mining and Sentiment Analysis for Unstructured Web Data." Journal of Emerging Technologies and Innovative Research. Vol. 2. No. 4 (April 2015). JETIR, 2015.

[74] Turney, Peter D., and Michael L. Littman. "Measuring praise and criticism: Inference of semantic orientation from association." ACM Transactions on Information Systems (TOIS) 21.4 (2003): 315-346.

[75] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning. New York: springer, 2013.

[76] Ferrucci, David A. "Introduction to "this is watson"." IBM Journal of Research and Development 56.3.4 (2012): 1-1.

[77] Dolatsara, Hamidreza Ahady. "Development of Safety Performance Functions for Non-Motorized Traffic Safety." (2014).

[78] Elberrichi, Zakaria, Abdellatif Rahmoun, and Mohamed Amine Bentaallah. "Using WordNet for Text Categorization." Int. Arab J. Inf. Technol. 5.1 (2008): 16-24.

[79] D.L. Olson, D. Delen, Advanced data mining techniques, Springer Publishing Company, Incorporated2008.

[80] https://en.wikipedia.org/wiki/Turkish_grammar

[81] B. Cromwell, Almost All Turkish Suffixes © by Jan 2016 http://cromwell-intl.com/turkish/turkish-suffixes.html.

[82] http://www.turkishlanguage.co.uk/vh1.htm

[83] Y. Göknel, Turkish Grammar, Updated Academic Edition, Vivatinell Bilim-Kültür Yayınları (2013)

[84] http://www.turkishclass.com/turkish_lesson_63

[85] http://text-processing.com/demo/stem/

[86] N. S. Moghaddami, H. S. Yazdi, H. Poostchi, Correlation based splitting criterionin multi branch decision tree, Ferdowsi University of Mashhad, Mashhad, Iran, Central European Journal of Computer Science Received 01 Feb 2011; accepted 07 Jun 2011

[87] K. Agron, Creating Composite Measures, Lecture 6: Computing Variables John Jay College

[88] Andrew P. Armacost, Cynthia Barnhart, Keith A. Ware, (2002) Composite Variable Formulations for Express Shipment Service Network Design. Transportation Science 36(1):1-20. http://dx.doi.org/10.1287/trsc.36.1.1.571

[89] H. Walkey Frank, Composite variable analysis: A simple and transparent alternative to factor analysis Department of Psychology, Victoria University of Wellington, P.O. Box 600, Wellington, New Zealand

[90] Handbook On Constructing Compoiste Indicators: Methodology And User Guide – Isbn 978-92-64-04345-9 - © Oecd 2008

Appendices