A RULE-BASED APPROACH USING THE ROUGH SET ON COVID-19 DATA

Rasim ÇEKİK [1*]

[1] Computer Engineering Department, Engineering Faculty. Şırnak University, 73000, Şırnak, Türkiye,
ORCID No : https://orcid.org/0000-0002-7820-413X

Keywords
COVID-19
Sentiment Analysis
Rough Set
Text Mining

Abstract

The COVID-19 pandemic has not only caused loss of life but also significantly affected people's emotional state. These emotional impacts have had serious consequences on societies and economies around the world. In order to repair these devastations in society, it is important to analyse these emotional effects in depth. In this study, the effects of the pandemic on human emotions are analysed using soft computing techniques. A rule-based approach is proposed for the analysis with the help of a rough set. The proposed method is based on two main components. The first one is the process of selecting the optimal subset (OFS) from the whole feature set with the help of k best known feature selection approaches. The second component involves the use of rough clustering methods to generate rules on the selected feature subset OFS. In the study, the first real data set called " Real World Concern Dataset", which is obtained from emotional responses to COVID-19, was used. The dataset consists of 5,000 items (2,500 short + 2,500 long). In the experimental studies, the proposed approach was tested with both labelled and unlabelled data, and it was observed that effective results were obtained with an accuracy rate of over 85%. It was also found that people were highly concerned about the future due to the pandemic.

COVID-19 VERİLERİ ÜZERİNDE KABA KÜME KULLANARAK KURAL TABANLI BİR YAKLAŞIMLA DUYGU ANALİZİ

Anahtar Kelimeler
COVID-19,
Kaba Kümeler,
Duygu Analizi,
Metin Madenciliği

Özet

COVID-19 salgını, sadece can kaybına neden olmakla kalmayıp aynı zamanda insanların duygusal durumlarını da önemli ölçüde etkilemiştir. Bu duygusal etkiler, dünya çapındaki toplumlar ve ekonomiler üzerinde ciddi sonuçlar doğurmuştur. Toplumda meydana gelen bu yıkımların onarılabilmesi için, bu duygusal etkilerin derinlemesine incelenmesi önemlidir. Bu çalışmada, salgının insan duyguları üzerindeki etkileri, yumuşak hesaplama teknikleri kullanılarak analiz edilmiştir. Analiz için kaba küme yardımıyla kural tabanlı bir yaklaşım önerilmiştir. Önerilen yöntem, iki temel bileşen üzerine kurulmuştur. Birincisi, k tane en iyi bilinen öznitelik seçme yaklaşımı yardımı ile tüm özellik kümesinden en uygun alt küme (OFS) seçme işlemdir. İkinci bileşen ise, seçilen özellik alt kümesi OFS üzerinde kurallar oluşturmak için kaba kümeleme yöntemlerinin kullanılmasını içermektedir. Çalışmada, COVID-19'a verilen duygusal tepkilerden elde edilen "Gerçek Dünya Endişe Veri Kümesi" adlı ilk elverişli gerçek veri kümesi kullanılmıştır. Veri kümesi 5.000 parçadan (2.500 kısa + 2.500 uzun) oluşmaktadır. Deneysel çalışmalarda, önerilen yaklaşımın hem etiketli hem de etiketsiz verilerle test edildiği ve %85'in üzerinde bir doğruluk oranıyla etkili sonuçlar elde edildiği gözlemlenmiştir. Ayrıca, insanların salgın nedeniyle geleceğe yönelik yüksek oranda endişe duydukları belirlenmiştir.

Araştırma Makalesi
Başvuru Tarihi : 15.01.2024
Kabul Tarihi : 16.05.2024

Research Article
Submission Date : 15.01.2024
Accepted Date : 16.05.2024

* Sorumlu yazar: rasimcekik@sirnak.edu.tr
https://doi.org/10.31796/ogummf.1420509

1. Introduction

The COVID-19 epidemic has affected people in almost every aspect such as economic, cultural, sociological, psychological, quality of life, as well as human health. The spread of its effects over such a wide area requires evaluating the COVID-19 epidemic from different

perspectives and from many aspects. Technological developments such as artificial intelligence and soft computing have provided ease of analysis and evaluation in many areas. Analyze and evaluate the emotions that people experience and feel about COVID-19 by using these possibilities of technology; in the future, it can provide a very important contribution to providing solutions to social problems such as increasing the economic welfare of the society, repairing its cultural wear, increasing the quality of life and questioning its psychological infrastructure. Therefore, successfully and effectively understanding and analyzing the emotions that people feel during the COVID-19 epidemic is a valuable and mandatory task in terms of public health. The primary objective of this research is to elucidate the ramifications of the COVID-19 pandemic on human emotions through the utilization of text mining methodologies and soft computing techniques. Through rigorous analysis and interpretation of textual data, the study endeavors to shed light on the nuanced manifestations of emotional states amidst the unprecedented challenges posed by the pandemic. In doing so, it seeks to contribute to the broader understanding of the psychological impact of the COVID-19 outbreak on individuals and societies at large.

The rapid development of the Internet and its very active use have enabled people to express their feelings easily. Moreover, companies can quickly access people's comments and feelings about products and adjust their sales policies accordingly. While this situation spreads over a very wide area, it has also created a very suitable environment for people to express their feelings about COVID-19. However, the huge amount of data that needs to be analyzed has made it impossible to make manual assessments. This problem is tried to be minimized by using high processing power of computers and various data processing approaches. Sentiment analysis, which emerged as a result of these needs, has been. A subject of interest in the fields of natural language processing and machine learning. Sentiment analysis aims to categorize the written and expressed judgments about a subject, according to the content, into categories such as the author's emotion, generally positive, negative, or neutral (Pak, 2015). However, depending on the content of the text written or expressed, emotions can vary such as pain, sadness, joy, anxiety, and fear. Many studies have been conducted on this subject so far, and most of these studies have focused on the document-level emotion classification problem, where the emotion expressed in the document is assumed to be about a single object (B. Liu, 2010). While some of the techniques used for document-level classification were unsupervised learning methods, mostly supervised learning methods were used. Supervised learning methods create a model using previously collected labeled data and classify new data according to this model (Aue & Gamon, 2005; Blitzer et al., 2007). Unsupervised methods (Turney, 2002) create a model without using any previously obtained labeled data and categorize the data according to the model. The absence of the desired amount of labeled data in each field represents an important problem in the field of sentiment analysis. In addition, the difficulty and cost of accessing labeled data is the main reason for this problem. Therefore, existing studies mostly focused on unlabeled data or tried to process unlabeled data using the few labeled data available.

Sentiment analysis can be considered as a text mining issue since it is a process on written documents. Basic operations such as preprocessing, feature extraction and feature selection in text mining are also valid processes for sentiment analysis. High dimensionality is one of the most common problems encountered in the analysis of text documents. The high dimensionality problem expresses the high feature space obtained by taking each term in the documents as a feature. One of the most effective and fast solutions offered to this problem is to select the subset that best represents the entire feature space within the entire feature set. Feature selection approaches are used for these operations.

When there is not enough data or inconsistent, repetitive and incomplete data, tools are needed to provide confidential and meaningful inferences from the data. At this point, Rough Sets is a mathematical soft computation tool that does this task very successfully. This success of Rough Sets has led to its use in a very wide area. The amount of missing and inconsistent data in the fields of text mining and its sub-title sentiment analysis is quite high. This has made the use of Rough Sets effective in this field. Within the scope of this study, it was tried to obtain the optimal feature set with the help of the most known k feature selection approach on labeled data. Then, a rule-based model was put forward by applying Rough Sets on this cluster, and according to this model, labeled and unlabeled data were analyzed according to the content.

2. Related Works

Sentiment analysis is the process of collecting and analyzing people's ideas, thoughts, and perceptions on various topics, products, news, and services. People's opinions can help companies working for commercial purposes, individuals who want to know about a subject, collect information and make decisions based on these ideas. However, the sentiment analysis and evaluation process is fraught with many challenges. These difficulties make it difficult to accurately interpret emotions and determine the appropriate emotion pole. This prompted researcher to analyze

texts and offer various methods to identify feelings/emotions (Acheampong et al., 2020; Garcia-Garcia et al., 2017; Sarsam et al., 2021; Zhang et al., 2019). Although people frequently express their emotions through writing, it can be difficult to determine the emotion of the written text. A single text, for example, could contain several emotions. Furthermore, some words can have multiple meanings that correspond to various emotions, making it difficult to identify emotion-carrying words or phrases in a given text. Grammatical errors, typos, sarcasm, or abbreviations in texts compiled from online media platforms are among the challenges researchers face when detecting emotional words. Furthermore, natural languages contain numerous metaphors, which can make capturing the true meaning of the text more difficult than expected. Another difficulty is determining the emotion or feeling in a text with idioms or proverbs. "Break a leg," for example, means "good luck" in English. The meanings of the idioms, however, cannot be deduced from the meanings of the words that comprise it. As a result, numerous studies have been published in the literature that use a dictionary and/or a list of keywords to determine the true meaning of idioms and proverbs (Ibrahim et al., 2015; Klebanov et al., 2013; Williams et al., 2015)

Sentiment analysis can be performed at three different levels: sentence, document, and aspect. It aims to classify the emotion expressed in each sentence at the sentence level. In this kind of approach, the first step is to decide if the sentence is subjective or objective. If the sentence is subjective, it expresses positive or negative opinions at Sentence Level. Moreover, documents or paragraphs are divided into sentences in sentence-level or phrase-level sentiment analysis, and the poles of each sentence are determined (Arulmurugan et al., 2019; Meena & Prabhakar, 2007). Document-level sentiment analysis attempts to extract general sentiment from extensive texts that contain a lot of noise and needless local terms. Document-level sentiment analysis is also referred to as sentiment classification. The biggest challenge in this type of sentiment classification is the need for the full context of semantic information to extract the semantic relationship from the statements contained in the documents. It also requires a deeper understanding of the complex internal structure of emotions and dependent words (S. Liu et al., 2020). For example, speed and cost are two semantic expressions in a review document about a computer, such as "the processor is high, but this product is very expensive". While the meaning expression about speed is directly in the sentence, the expression about cost is mentioned implicitly. In this case, the cost expression is more difficult to derive and requires extra processing. (Devi Sri Nandhini & Pradeep, 2020) devised a method to extract implicit aspects from documents based on the frequency of coexistence of feature indicators and features, as well as the relationship between thought words and explicit aspects. Sentiment Analysis at the aspect-level (Schouten & Frasincar, 2015) is a subtask of Sentiment Analysis that focuses on determining the sentiment of a specific aspect or feature of a product, service, or entity. Unlike document-level, which analyzes the sentiment of a piece of text as a whole, aspect-level analyzes the sentiment of individual aspects of the entity, such as specific attributes or features. Aspect-level processing is carried out using natural language processing techniques such as named entity recognition, part-of-speech tagging, dependency parsing, and sentiment analysis algorithms. These techniques can assist in identifying the aspect or feature mentioned in the text, its sentiment polarity (positive, negative, or neutral), and the intensity of the expressed sentiment.

Measuring people's feelings and concerns regarding COVID-19 will be critical to understanding and dealing with the problem. For this reason, a lot of data about the epidemic has been collected. However, the dataset used in this study differs from others with some featuresCOVID-19 Real World Worry collection (RWWD) presents a high-quality, real-text collection of COVID-19-related emotions and concerns, as well as preliminary findings on emotional linguistic correlates, subject models, and prediction tests. RWWD is a ground truth dataset that uses a direct survey method to collect written descriptions of people as well as data on their sentiments and worries. As a result, the dataset no longer relies on third-party descriptions and can instead refer to self-reported feelings. Several research has been done to evaluate the epidemic's consequences. For example, (a) Coronavirus-related tweets have been collected since March 11, 2020, with an average of 4.4 million tweets per day (Banda et al., 2020). Tweets with keywords like "coronavirus" and "COVID-19" were collected using the Twitter streaming API. (b) Another Twitter collection of Coronavirus messages in many languages, including English, Spanish, and Indonesian, has been collected since January 22, 2020 (Chen et al., 2020).

3. Preliminaries and Background

In this section, basic information about the technology and approaches used in the background of the study is presented.

3.1. Text Mining and Sentiment Analysis

Text mining, also called text analytics, is an artificial intelligence technology that makes unstructured text in documents suitable for analysis or machine learning algorithms using natural language processing (ÇEKİK, 2022; ÇEKİK & Mahmut, 2023; Parlak & Uysal, 2020, 2023). Text mining, which is widely used in knowledge-

oriented organizations, is a review process that allows extracting new information from very large collections of texts or providing answers for specific research. Text mining, which derives information from written sources such as websites, books, e-mails, articles, and online news, organizes and structures data using advanced methodologies. This information can be used for descriptive, prescriptive, or predictive analytics by integrating it into databases, data warehouses, or business intelligence panels. Security, biomedicine, online media, sentiment analysis, commerce and marketing, digital humanities and computational sociology, scientific literature mining, and academic applications are only a few of the applications for text mining. Text mining tasks include text categorization, text clustering, concept or entity extraction, granular taxonomy modeling, sentiment analysis, entity connection model, and document summary.

There are a number of processes that text mining tasks have in common:

- Preprocessing
    - tokenization
    - removal of stop words
    - letter transformation
    - stemming
- Feature Extraction
- Feature Selection

Sentiment analysis is basically a text analysis and aims to determine the class (positive, negative and neutral) that the given text wants to express emotionally. Since sentiment analysis can be used in different fields and for different purposes, it is possible to see it in the literature with various approaches. For example, while general sentiment analysis is used for tasks where it is sufficient to assign a single sentiment class to the entire text, target-based sentiment analysis may be needed in cases where there is more than one sentiment in the text. As a result of the frequent occurrence of these and similar situations, the most commonly used sentiment analysis types can be stated as follows:

- Subjectivity/Objectivity Analysis
- General Sentiment Analysis
- Aspect Based Sentiment Analysis
- Fine-Grained Sentiment Analysis
- Emotion Detection

The primary goal of feature selection approaches is to solve the high dimensionality problem by selecting the characteristics with the highest uniqueness from all available features. For this purpose, there are approaches with different working mechanisms such as filter, winding and embedded in the literature. However, in this study, these approaches were used when the filter approaches worked faster. Feature selection in such methods; It is accomplished by the use of functions based on statistical criteria such as feature distance, measure of dependency between features, or feature information. A score is calculated for each feature through these statistical functions, and The features with the highest ratings are selected to create the best feature subset (Cekik & Uysal, 2020; Chandrashekar & Sahin, 2014). In this study, the following filter feature approaches were used.

Information gain (IG): It is defined as the inverse of entropy, which expresses the disorder of a system. It is expressed statistically as follows:

$$IG(t) = -\sum_{i=1}^{M} P(C_i) log P(C_i) + P(t) \sum_{i=1}^{M} P(C_i|t) log P(C_i|t) + P(\bar{t}) \sum_{i=1}^{M} P(C_i|\bar{t}) log P(C_i|\bar{t}) \quad (1)$$

Chi-square (Chi2): It is a statistical approach that investigates whether the relationship between two relationships is dependent. It is expressed statistically as follows:

$$CHI2(t, C) = \sum_{t\in\{0,1\}} \sum_{C\in\{0,1\}} \frac{(N_{t,C} - E_{t,C})^2}{E_{t,C}} \quad (2)$$

$$CHI2(t) = \sum_{i=1}^{M} P(C_i) * CHI2(t, C) \quad (3)$$

Gini index(GI) is a mathematical method offered as an alternative to IG and does not use entropy. Mathematically:

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 P(C_i|t)^2 \quad (4)$$

Deviation from poisson distribution (DP): It is a widely used approach for selecting effective words in the field of information retrieval and is derived from the Poisson distribution.

$$DP(t,C) = \frac{(a-\hat{a})^2}{\hat{a}} + \frac{(b-\hat{b})^2}{\hat{b}} + \frac{(c-\hat{c})^2}{\hat{c}} + \frac{(d-\hat{d})^2}{\hat{d}}$$

$$\begin{aligned} \hat{a} &= n(C)\{1-\exp(-\mu)\} \\ \hat{b} &= n(C)\exp(-\mu) \\ \hat{c} &= n(\overline{C})\{1-\exp(-\mu)\} \\ \hat{d} &= n(\overline{C})\exp(-\mu) \\ \mu &= \frac{F}{N} \end{aligned} \quad (5)$$

Distinguishing feature selector (DFS): It is an effective feature selection approach that has been put forward recently and is formulated as follows (Uysal & Gunal, 2012):

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\overline{t}|C_i) + P(t|\overline{C_i}) + 1} \quad (6)$$

3.2. Rough Set Theory

Pawlak's Rough Set Theory (RST) (Pawlak, 1998) a mathematical method that successfully extracts hidden patterns or patterns in data. RST organizes partial, inconsistent, and ambiguous datasets such that they can be evaluated. Although it is typically employed as an auxiliary process, it is capable of performing operations like as classification, pattern extraction, rule creation, dimension reduction, feature selection, and feature extraction without the usage of any other method. Below are brief explanations of the fundamental ideas of rough sets.

Let$S = (U, A, C)$represents a decision table or information system, where $U = \{x_1, x_1, \dots, x_n\}$ is the universal set of objects, A is a conditional attribute set, and C is a decision attribute set. For any subset of $T \subseteq A$ conditional attributes, the T- indiscernibility relationship, whose designation is *IND(T)*, is defined as follows:

$$IND(T) = \{(x_i, x_j) \in U^2 | \forall a \in T, a(x_i) = a(x_j)\} \quad (7)$$

Where, the equivalence classes of the T- indiscernibility relationship are expressed as $[x]_T$. Lower and Upper set approaches express two basic concepts in rough set.

The representations of the T-lower and T-upper approaches of the set X on any subset of $X \subseteq U$ objects and the specified subset of $T \subseteq A$ attributes are $\underline{T}X$ and $\overline{T}X$, respectively. Also, their definitions are as follows:

$$\begin{aligned} \underline{T}X &= \{x \mid [x]_T \subseteq X\}, \\ \overline{T}X &= \{x \mid [x]_T \cap X \neq \emptyset\} \end{aligned} \quad (8)$$

A representative illustration showing the relationship of RST concepts is given in Figure 1.

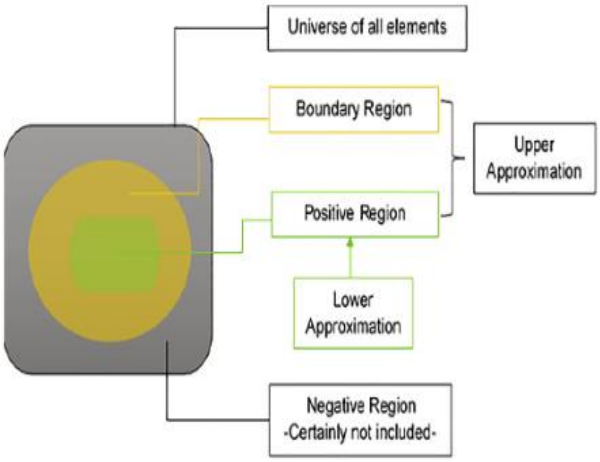


Figure 1. The regions and approximation sets in the RST.

4. Proposed Method

The proposed approach consists of two basic structures. One of them is the process of selecting a subset of the entire feature set with the help of the best-known k feature selection approach. The other basic structure is to obtain rules on the selected feature subset with the help of rough sets. Then, using these rules, sentiment analysis is performed on tagged and unlabeled COVID-19 data. The working principle of the proposed approach in Figure 2 is given below.

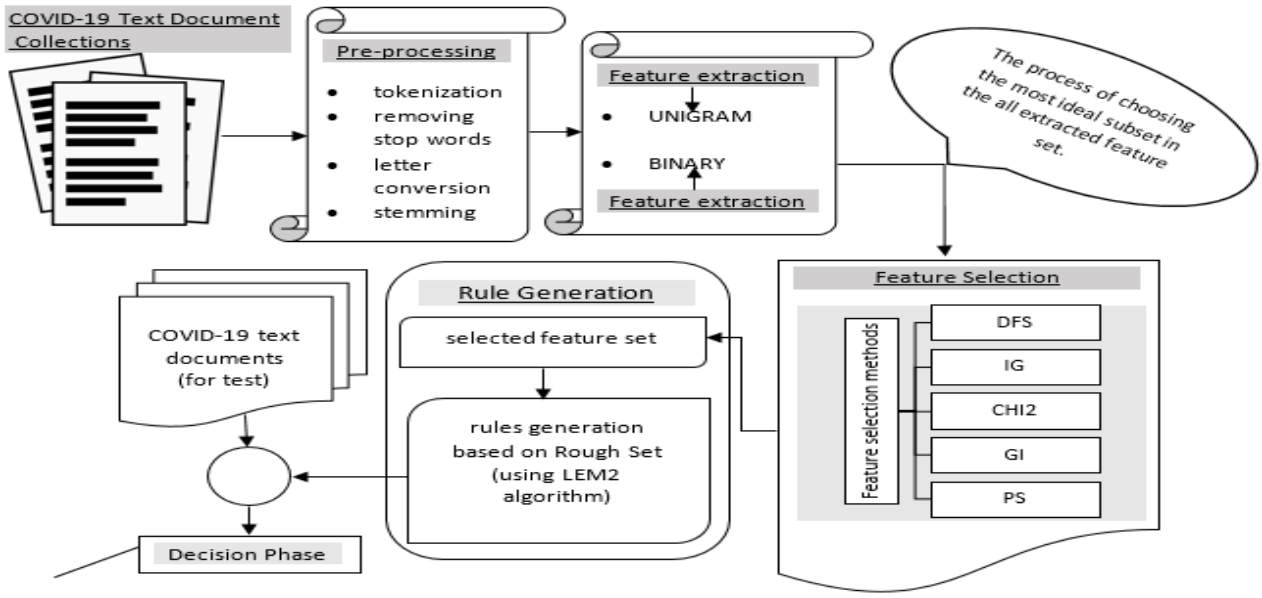


Figure 2. Working architecture of the proposed approach

In this study, the most well-known IG, CHI2, DFS, GI, and PS feature selection approaches in text mining were used. Each of these approaches separately selects top-n-size features. By taking the cluster combination of all selected top-n feature subsets, the Optimal Feature Set (OFS) is obtained to be presented to the rough set approach. The OFS equation is given below.

$$OFS = IG_{top-n} \cup CHI2_{top-n} \cup DFS_{top-n} \cup GI_{top-n} \cup PS_{top-n} \quad (9)$$

OFS refers to the columns in the decision information system/table for rough sets. Rules are obtained with the help of Rough Sets on the created information system. It utilized the ROSE 2 (Rough Set Data Explorer) system to derive the rules. ROSE 2 system uses LEM2 (Learning from Examples Module, version 2) (Stefanowski, 1998) algorithm to create rules. The LEM2 algorithm is a supervised learning technique that is used to construct rules from data and is based on the lower or upper approximation set of rough set theory (Grzymala-Busse, 2015). The algorithm operates at the (attribute, value) pair level. For each notion, a local coverage is calculated. The pseudocode of LEM2 (Grzymala-Busse, 1992; Stefanowski, 2001), is quoted:

LEM2

(**INPUT:** a set B,

OUTPUT: a single local covering T of set B);

BEGIN

$G := B$;

$T := \emptyset$;

WHILE $G \neq \emptyset$

BEGIN

$T := \emptyset$;

$T(G) := \{t | [t] \cap G \neq \emptyset\}$;

WHILE $T = \emptyset$ *or* $[T] \not\subseteq B$

BEGIN

select a pair $t \epsilon T(G)$ such that $|[t]| \cap G$ is max: if a tie occurs, select a pair $t \in T(G)$ with the smallest cardinality of $[t]$; if another tie occurs, select first pair; $T := T \cup \{t\}$;

$G := [t] \cap G$;

$T(G) := \{t | [t] \cap G \neq \emptyset\}$;

$T(G) := T(G) - T$;

END

FOR each $t \in T$ DO

IF $T - \{t\} \subseteq B$ THEN $T := T - \{t\}$;

$T := T \cup \{T\}$;

$G := B - U_{T \epsilon T}[T]$;

END

FOR each $T \in T$ DO

IF $U_{S \epsilon T - \{T\}}[S] = B$; THEN $T := T - \{T\}$;

END

A cross-section learning from the LEM2 results obtained within the scope of this study is shown on Figure 3. Equation 10, which is a statistical approach on the rules obtained by the LEM2 algorithm, is used to decide the sentiments of the text.

Where $Similarty_{Rule}$ shows the similarity ratio of the test data to the rule. $Rule_{rayting}$ indicates the number of rules with similarity. In other words, a rule occurs for each case in the training data during the rule generation stage with the rough set. In this case, the same rules can be generated. This also means that the situation was used frequently and is important information. In order to determine which class a data belongs to, the similarity rates of the rules on a class basis are analysed. The data is included in the class with a high similarity rate. Pseudocode of the proposed approach:

ALGORITHM 1: Proposed Method

```
PROCEDURE Proposed_Method(Feature Size N)
BEGIN
select_best_feature(N) // using DFS, IG, CHI2, GI, PS
```

$$Similarty_{Rule} = Count(t_{rule} = t_{test}) * \log_{10} Rule_{rayting} \quad (10)$$

```
OFS ←calculated_OFS(N)
generation_rules(OFS) // using rough set (LEM2) algorithm
j←1
  WHILE j < size(test_data)
       FOR i ←1: size(rules)
         similarty(i) = Count(t_rule = t_test) log10 Rule_rayting
       END // for FOR
       decision(j) =max (similarty)
        ++j
   END // for WHILE
END // for BEGIN
```

```
#       LEM2
#       C:\Users\RSC\Desktop\Mahmut_Kaya\Criteria_main\Rose\Created_DataSet\F_50_ROSE_150.isf
#       objects = 1677
#       attributes = 150
#       decision = D
#       classes = {1, 2, 3, 4, 5, 6, 7, 8}
#       Wed Oct 12 20:24:03 2022
#       184 s

rule 1. (A27 = 1) & (A102 = 0) & (A104 = 0) & (A114 = 0) & (A121 = 1) & (A123 = 0) => (D = 1); [1, 1, 1.33%, 100.00%][1, 0, 0, 0, 0, 0, 0, 0]
[{29}, {}, {}, {}, {}, {}, {}, {}]
rule 2. (A7 = 0) & (A19 = 1) & (A27 = 0) & (A36 = 0) & (A98 = 0) & (A102 = 0) & (A108 = 0) & (A119 = 1) & (A123 = 0) & (A124 = 0) => (D = 1); [5, 5, 6.67%, 100.00%][5, 0, 0, 0, 0, 0, 0, 0]
[{5, 12, 18, 56, 63}, {}, {}, {}, {}, {}, {}, {}]
rule 3. (A12 = 0) & (A14 = 1) & (A19 = 0) & (A102 = 0) & (A104 = 0) & (A119 = 0) & (A125 = 0) & (A129 = 0) => (D = 1); [2, 2, 2.67%, 100.00%][2, 0, 0, 0, 0, 0, 0, 0]
[{10, 34}, {}, {}, {}, {}, {}, {}, {}]
rule 4. (A1 = 0) & (A100 = 0) & (A102 = 0) & (A104 = 1) & (A107 = 0) & (A110 = 0) & (A111 = 0) & (A113 = 0) & (A114 = 0) & (A122 = 0) & (A125 = 0) & (A126 = 0) & (A129 = 0) => (D = 1); [5, 5, 6.67%, 100.00%][5, 0, 0, 0, 0, 0, 0, 0]
[{17, 42, 68, 69, 73}, {}, {}, {}, {}, {}, {}, {}]
rule 5. (A47 = 0) & (A98 = 0) & (A103 = 1) & (A104 = 0) & (A107 = 0) & (A118 = 0) & (A122 = 0) & (A129 = 1) & (A147 = 0) => (D = 1); [2, 2, 2.67%, 100.00%][2, 0, 0, 0, 0, 0, 0, 0]
[{25, 59}, {}, {}, {}, {}, {}, {}, {}]
rule 6. (A3 = 0) & (A7 = 0) & (A36 = 0) & (A44 = 1) & (A47 = 0) & (A106 = 0) & (A112 = 0) & (A113 = 0) & (A114 = 0) & (A119 = 0) & (A122 = 0) & (A123 = 0) => (D = 1); [1, 1, 1.33%, 100.00%][1, 0, 0, 0, 0, 0, 0, 0]
[{75}, {}, {}, {}, {}, {}, {}, {}]
rule 7. (A1 = 0) & (A119 = 0) & (A122 = 0) & (A127 = 0) & (A144 = 1) => (D = 1); [2, 2, 2.67%, 100.00%][2, 0, 0, 0, 0, 0, 0, 0]
[{17, 47}, {}, {}, {}, {}, {}, {}, {}]
```

Figure 3. A cross-section of LEM2 results on OFS obtained according to Top-50

5. Experimental Works

5.1. Data Collection

Measuring Emotions in the COVID-19 Real World Worry Dataset (RWWD) was used as the dataset in this study. The dataset presents the first basic real dataset of emotional responses to COVID-19. The Dataset consists of 5000 texts (2,500 short + 2,500 long texts) asking participants to express their feelings and express them in text. This dataset reports initial findings for the RWWD, which examines people's emotions and concerns at a time when the impact of the COVID-19 pandemic is affecting the lives of all people living in the UK. Data were collected by Bennett K. (Kleinberg et al., 2020) on 6 and 7 April 2020, when the UK was under quarantine and deaths were rising. RWWD is a baseline fact dataset that obtains written emotion and anxiety data as well as people's other emotions using a direct survey method. As such, the dataset is not based on third-party disclosures, but may

Table 1. Results For Top-50, 100, 200, 300, 400, 500, 600, 700, 800, and 1000

Feature Size	50/150	100/283	200/484	300/612	400/878
75	0.5467	0.5733	0.4933	0.5733	0.5867
150	0.7200	0.7600	0.7133	0.7267	0.7667
250	0.7920	0.8320	0.7800	0.8040	0.8240
400	0.8375	0.8775	0.8225	0.8350	0.8600
600	0.7183	0.7350	0.7050	0.7067	0.7200
700	0.6214	0.6343	0.6057	0.6114	0.6300
Feature Size	500/1047	600/1184	700/1289	**800/1353**	1000/1562
75	0.5067	0.4267	0.4533	0.5600	0.5333
150	0.7267	0.5800	0.7133	0.7600	0.7600
250	0.8080	0.6560	0.8040	0.8360	0.8400
400	0.8275	0.7050	0.8625	**0.8875**	0.8800
600	0.6933	0.6100	0.7267	0.7467	0.7450
700	0.6071	0.5386	0.6300	0.6471	0.6486

refer directly to self-reported emotions. Kleinberg et al. (2020) offered two versions of RWWD, each consisting of 2,500 English sentences portraying participants' genuine concerns about the Corona situation in the UK: Long RWWD with open-ended duration and wishing inscriptions encouraging participants to express their emotions as they feel. In the brief RWWD, the same people were asked to convey their sentiments in Tweet-sized language. Furthermore, this short RWWD was chosen to make it easier to utilize for Twitter data research.

The emotions expressed in the dataset are Worry, Anger, Anxiety, Desire, Disgust, Fear, Happiness, Relaxation and Sadness, respectively. These emotions also determine the labels in the data set. That is, labeled data are classified according to these emotions. It was tried to find out which of these emotion types the unlabeled data belonged to. In the study, both labeled and unlabeled data were classified according to the model created by using the labeled data, and the classification success was shown in the tables.

5.2. Accuracy Analysis

The evaluation criterion employed at this stage is the measure of accuracy.

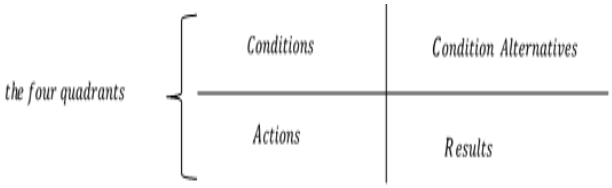
This section consists of a two-stage evaluation process. In the first stage, the performance of the proposed approach is tested on the labeled data and the basis of the model to be used in the next stage is established. In the first step, the OFS cluster is obtained using the best-known IG, CHI2, DFS, GI and PS feature selection approaches. This set constitutes the feature space of the information system presented to the rough sets. Rules are created with the help of rough sets over the obtained information system. These rules are one of the cornerstones of the working principle of the proposed approach and significantly affect the performance of the approach. Therefore, it is an important criterion to determine the feature space in which the approach gives the best results. To determine this criterion, the model is tested in different feature spaces. This process is been chosen at random from the top-n values in the study. The feature subsets chosen are 50, 100, 200, 300, 400, 500, 600, 700, and 800, in that order. In addition, the OFS numbers obtained for each selected feature size are 150, 283, 484, 612, 878, 1047, 1184, 1289, 1353 and 1562, respectively. The rule-based working model associated with each feature dimension is then tested for 75, 150, 250, 400, 600, and 700 feature dimensions. The dimensions for this test are also chosen at random. Table 1 shows the accuracy results obtained.

In 50/150 notation, 50 denotes the top-50 value for feature selection approaches, and 150 denotes the OFS value obtained based on this top-50 value. Table-1 also includes all other top-n values. According to Table 1, the top-800 has the highest accuracy when the test size is 400. This data determines the top-n and associated OFS values that will be used to build the model structure in the next step. As a result, the feature size for the second stage is 800, and the OFS value is 1353.The second phase is divided into two parts: *comparing against existing rule-based working models and running on unlabeled data.*

5.3. Comparing against existing rule-based working models

Performance comparisons are made with the proposed approach, rule-based ZeroR, Decision Table, JRip and Part approaches.

- The ZeroR method calculates the mean of numerical or nominal test data. It follows the basic coverage algorithm rules. ZeroR essentially just attempts to detect the majority class distribution, which is assumed as a rule. ZeroR estimates the arithmetic mean for numeric class characteristics and the value of the mode class for nominal class attributes. Other than this, ZeroR does not generate any rules.
- Decision Tables are a technique for modeling complex logic, systematizing decision making, and testing all combinations that influence a decision. Decision Table Structure; Decision tables consist of 4 parts called “Conditions”, “Condition Alternatives”, “Actions” and “Results”. This particular structure is called “The Four Quadrants”.



- JRip is a basic and extensively used algorithm. Classes are inspected in increasing order, and a starting rule set for the class is produced using incremental reduced error JRip. (RIPPER). It then treats all instances of a certain decision in the training data as a class and searches for a set of rules that applies to all members of that class. The method is then repeated for the next class until all classes have been processed. They start with a default rule and aim to learn rules that predict default exceptions using a training dataset. Each learned rule is made up of a set of propositional invariants. Each literal indicates a single-value data division.
- The algorithm generates "decision lists" of rules that represent the intended set of rules. A new item is compared to each rule in the list in turn, and the class assigned to it is that of the first matching rule. PART produces a partial C4.5 decision tree at each iteration, with the "best" leaf being a rule. PART is an extremely efficient method in terms of both computational performance and results. The decision tree learning divide-and-conquer strategy is generally used with the rule-learning divide-and-conquer method in PART. The C4.5 method is used to build a decision tree, and the leaf with the maximum coverage is turned into a rule. This rule's set of examples is then discarded, and the procedure is repeated. As a result, an ordered collection of rules is generated, which is augmented with a default rule that applies in cases where none of the preceding rules apply.

The proposed approach's performance is compared to the existing rule-based classifier approaches given above based on the success of the 75, 150, 250, 400, 600, and 700 feature dimensions. For existing classification approaches, DFS, CHI2, IG and GI feature selection approaches are used. The performance of each classifier on the feature subset presented by the feature selection approaches is compared to the success of the proposed approach for the feature subset of the same size. The results are shown in Figures 4, 5, 6 and 7.

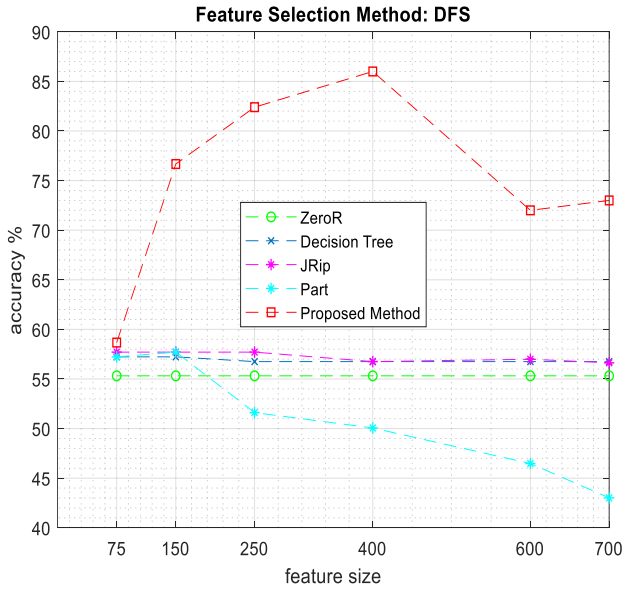


Figure 4. Classifier accuracy results for DFS

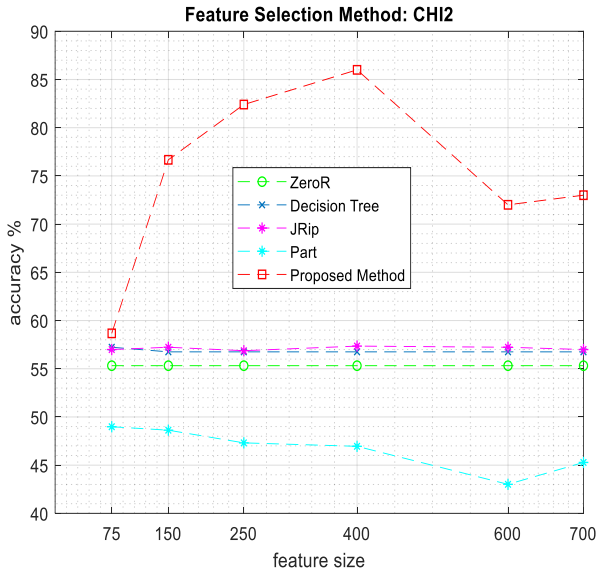


Figure 5. Classifier accuracy results for CHI2

Figure 4 and Figure 5 show the performance of the current classifiers for the DFS and CHI2 feature selection approaches, as well as the performance of the proposed approach for the same feature size regardless of feature selection approach. When Figure 4 and Figure 5 are examined, it is seen that the proposed approach gives much better results in all feature dimensions. The reason for this is the similarity calculation ability of the presented model, depending on the fact that the proposed approach successfully reveals the hidden and meaningful information in the data with the help of rough sets.

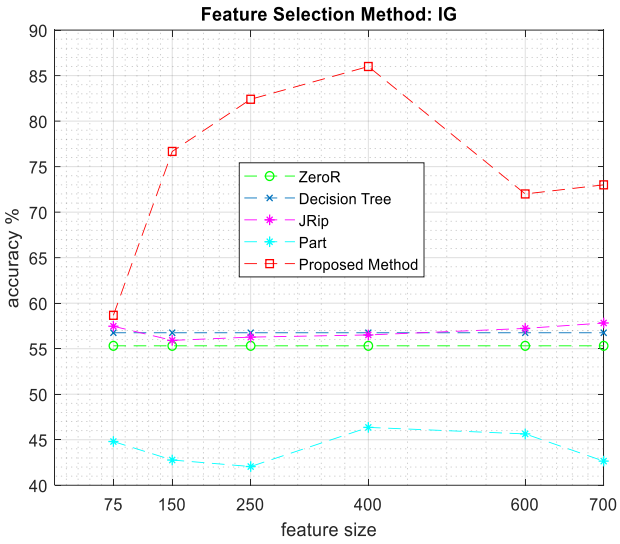


Figure 6. Classifier accuracy results for IG

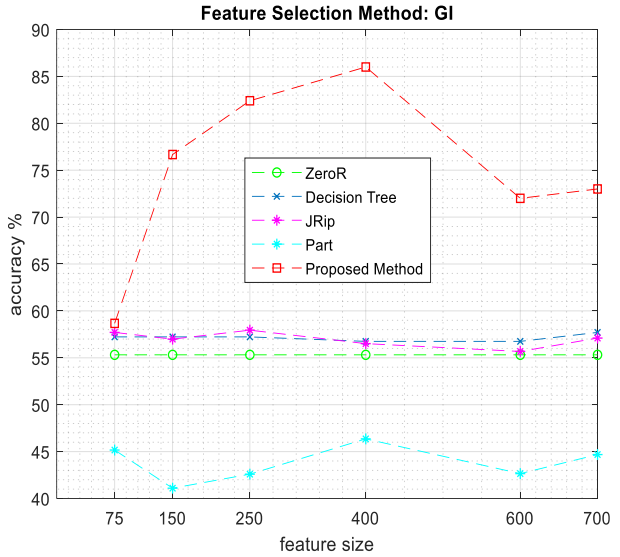


Figure 6. Classifier accuracy results for GI

Similarly, Figure 5 presents the results of the classifiers for the GI and IG feature selection approaches. Again, Figure 5 shows the performance of the proposed approach for the same feature size, regardless of the feature selection approach. When Figure 5 is examined, it is clear that the proposed approach yields the best results. The proposed approach's working mechanism outperforms other existing classifiers in terms of extracting meaningful content from the Covid 19 dataset.

The second stage consists of evaluation on unlabeled data. The label of unlabeled data will be estimated with the rules generated with rough sets on top-n feature sizes that give the best results (top-800).

5.4. Running on unlabeled data

In the previous stage, it was learned that the best results were obtained in 400 test feature sizes (See Table 1). At this stage, firstly, experimental results for top-100, 1000 and 5000 were obtained by using 400 unlabeled data in line with this information. The results are given in Table 2.

Table 2. Results For Top-50, 100, 200, 300, 400, 500, 600, 700, 800, and 1000 by using 400 unlabeled text documents

Emotion (Class)	Top -50		Top -100		Top -200		Top -300		Top -400	
	Number	Rate	Number	Rate	Number	Rate	Number	Rate	Number	Rate
Anger	18	0.0450	21	0.0525	7	0.0175	18	0.0450	18	0.0450
Anxiety	**355**	**0.8875**	**364**	**0.9100**	**353**	**0.8825**	**353**	**0.8825**	**360**	**0.9000**
Desire	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Disgust	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Fear	12	0.0300	1	0.0025	23	0.0575	11	0.0275	0	0.0000
Happiness	1	0.0025	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Relaxation	4	0.0100	1	0.0025	1	0.0025	5	0.0125	16	0.0400
Sadness	10	0.0250	13	0.0325	16	0.0400	13	0.0325	6	0.0150
Emotion (Class)	Top -500		Top -600		Top -700		Top -800		Top -1000	
	Number	Rate	Number	Rate	Number	Rate	Number	Rate	Number	Rate
Anger	5	0.0125	5	0.0125	2	0.0050	3	0.0075	3	0.0075
Anxiety	**362**	**0.9050**	**311**	**0.7775**	**376**	**0.9400**	**369**	**0.9225**	**383**	**0.9575**
Desire	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Disgust	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Fear	2	0.0050	22	0.0550	7	0.0175	5	0.0125	2	0.0050
Happiness	1	0.0025	0	0.0000	1	0.0025	1	0.0025	0	0.0000
Relaxation	18	0.0450	23	0.0575	6	0.0150	12	0.0300	5	0.0125
Sadness	12	0.0300	39	0.0975	8	0.0200	10	0.0250	7	0.0175

Table 2 contains eight different emotional expressions. These emotional expressions also denote social class. In the table, “Number” refers to the number of emotions and “Rate” refers to their ratio. Table 2 shows the results of the proposed model for various top-size properties. The results clearly show that 'Anxiety' is the dominant emotion. It is also useful to compare these results to the labeled data distribution in the dataset. The distribution of labeled data in terms of emotion expressions is shown in Table 3.

Table 3. Text data descriptive statistics and emotion ratings.

Emotions	Ratio (%)
Anger	4.33%
Anxiety	**55.36%**
Desire	1.09%
Disgust	0.69%
Fear	9.22%
Happiness	1.38%
Relaxation	14.36%
Sadness	13.36%

The intense emotion in the Covid 19 dataset is 'Anxiety' as shown in Table 3. Furthermore, when other emotions are considered, it is very close to the emotion classification of the proposed model on unlabeled data. This demonstrates that the proposed method successfully performs sentiment analysis. To gain a better understanding of the situation, examine the emotion classification based on the top-800, which previously provided the best results. Table 4 shows the Top-800 results using unlabeled text document data from 75, 150, 250, 400, 600 and 700.

The DN on Table 4 indicates the number of documents with the relevant emotion, and the UDD-n indicates the n number of unlabeled documents. In addition, when Table 4 is taken into account, it is seen that the emotion distribution is better for UDD-400.

As a result, it was observed in the experiments that the proposed approach gave more effective results on the RWWD dataset than the other existing rule-based models. In addition, it has been determined that the model gives more successful results in Top-800 feature size on unlabeled data. After all, it can be said that people mostly feel "Anxiety" about COVID-19.

Table 4. Results For Top-50, 100, 200, 300, 400, 500, 600, 700, 800, and 1000 by using 400 unlabeled text documents

Emotion (Class)	UDD -75		UDD -150		UDD -250		**UDD -400**		UDD -600		UDD -700	
	DN	Rate	DN	Rate	DN	Rate	**DN**	**Rate**	DN	Rate	DN	Rate
Anger	4	0.0533	5	0.0333	5	0.0200	**3**	**0.0075**	8	0.0133	8	0.0114
Anxiety	68	0.9067	140	0.9333	235	0.9400	**369**	**0.9225**	571	0.9517	661	0.9443
Desire	0	0.0000	0	0.0000	0	0.0000	**0**	**0.0000**	0	0.0000	0	0.0000
Disgust	0	0.0000	0	0.0000	0	0.0000	**0**	**0.0000**	0	0.0000	0	0.0000
Fear	0	0.0000	0	0.0000	1	0.0040	**5**	**0.0125**	2	0.0033	3	0.0043
Happiness	1	0.0133	1	0.0067	1	0.0040	**1**	**0.0025**	1	0.0017	1	0.0014
Relaxation	0	0.0000	1	0.0067	3	0.0120	**12**	**0.0300**	5	0.0083	10	0.0143
Sadness	2	0.0267	3	0.0200	5	0.0200	**10**	**0.0250**	13	0.0217	17	0.0243

6. Conclusion

The COVID-19 pandemic crashed into people's lives like a nightmare in 2020. People have suffered greatly as a result of this epidemic. They lost their lives, health, social activities, freedom, and economic, mental, and emotional well-being. These effects appear to last a long time. To mitigate the effects, accurately and effectively analyze people's feelings about the epidemic and develop solutions accordingly. This study presents effective implications on the RWWD dataset, which examines people's emotions and concerns at a time when the impact of the COVID-19 pandemic is affecting the lives of all people living in the UK. In the study, features that provide important information were selected with feature selection approaches that are effective in classifying text data, and a rule-based model with a rough set that provides effective information from insufficient data was presented. The model successfully revealed private information in text data. In the study, while manually labeled data was used in model creation, unlabeled data was provided in the background to determine the label with the proposed model. Experimental results show that the proposed model outperforms existing classical approaches in classification. The results also revealed that people are very worried and fearful about the pandemic. Uncertainty, fear and anxiety caused by the pandemic can lead to a wide range of negative consequences such as anxiety and fear in society, psychological problems, insecurity and increased violence, disruption in family structure, weakening of social ties, economic difficulties, interruptions in education. Therefore, it is important to develop urgent and comprehensive policies to address and minimise the psychological and social impacts of the pandemic. Building solidarity and support mechanisms at the individual and societal level is vital in mitigating the negative effects of the pandemic and making our society more resilient.

Contribution of Researchers

In this study, the author worked alone in all processes of the study, such as literature research, conducting experimental studies, and analyzing the results.

Conflict of Interest

No conflict of interest was declared by the authors.

References

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7), e12189.

Arulmurugan, R., Sabarmathi, K. R., & Anandakumar, H. (2019). RETRACTED ARTICLE: Classification of sentence level sentiment analysis using cloud machine learning techniques. Cluster Computing, 22(Suppl 1), 1199–1209.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. Proceedings of Recent Advances in Natural Language Processing (RANLP), 1(3.1), 1–2.

Banda, J., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., & Chowell, G. (2020). A Twitter dataset of 150+ million tweets related to COVID-19 for open research, April 5.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 440–447.

ÇEKİK, R. (2022). Metin Siniflandirma İçin Filtre Öznitelik Seçim Yaklaşimlari. Mühendislik Alanında Uluslararası Araştırmalar II, 87.

ÇEKİK, R., & Mahmut, K. (2023). A New Feature Selection Metric Based on Rough Sets and Information Gain in Text Classification. Gazi University Journal of Science Part A: Engineering and Innovation, 10(4), 472–486.

Cekik, R., & Uysal, A. K. (2020). A novel filter feature selection method using rough set for short text data. Expert Systems with Applications, 160, 113691.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.

Chen, E., Lerman, K., & Ferrara, E. (2020). Covid-19: The first public coronavirus twitter dataset.

Devi Sri Nandhini, M., & Pradeep, G. (2020). A hybrid co-occurrence and ranking-based approach for detection of implicit aspects in aspect-based sentiment analysis. SN Computer Science, 1, 1–9.

Garcia-Garcia, J. M., Penichet, V. M. R., & Lozano, M. D. (2017). Emotion detection: a technology review. Proceedings of the XVIII International Conference on Human Computer Interaction, 1–8.

Grzymala-Busse, J. W. (1992). LERS-a system for learning from examples based on rough sets. Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory, 3–18.

Grzymala-Busse, J. W. (2015). Rule induction from rough approximations. Springer Handbook of Computational Intelligence, 371–385.

Ibrahim, H. S., Abdou, S. M., & Gheith, M. (2015). Idioms-proverbs lexicon for modern standard Arabic and colloquial sentiment analysis. ArXiv Preprint ArXiv:1506.01906.

Klebanov, B. B., Burstein, J., & Madnani, N. (2013). Sentiment profiles of multiword expressions in test-taker essays: The case of noun-noun compounds. ACM Transactions on Speech and Language Processing (TSLP), 10(3), 1–15.

Kleinberg, B., Van Der Vegt, I., & Mozes, M. (2020). Measuring emotions in the covid-19 real world worry dataset. ArXiv Preprint ArXiv:2004.04225.

Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of Natural Language Processing, 2(2010), 627–666.

Liu, S., Lee, K., & Lee, I. (2020). Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. Knowledge-Based Systems, 197, 105918.

Meena, A., & Prabhakar, T. V. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. Advances in Information Retrieval: 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007. Proceedings 29, 573–580.

Pak, M. Y. (2015). Metinlerde duygu analizi ve sınıflandırma için yeni yöntemler. Anadolu University (Turkey).

Parlak, B., & Uysal, A. K. (2020). On classification of abstracts obtained from medical journals. Journal of Information Science, 46(5), 648–663.

Parlak, B., & Uysal, A. K. (2023). A novel filter feature selection method for text classification: Extensive Feature Selector. Journal of Information Science, 49(1), 59–78.

Pawlak, Z. (1998). Rough set theory and its applications to data analysis. Cybernetics & Systems, 29(7), 661–688.

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., Alnumay, W., & Smith, A. P. (2021). A lexicon-based approach to detecting suicide-related messages on Twitter. Biomedical Signal Processing and Control, 65, 102355.

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(3), 813–830.

Stefanowski, J. (1998). On rough set based approaches to induction of decision rules. Rough Sets in Knowledge Discovery, 1(1), 500–529.

Stefanowski, J. (2001). Algorithms of decision rule induction in data mining. Poznan University of Technology Press, Poznan, Poland.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. ArXiv Preprint Cs/0212032.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36, 226–235.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., & Spasić, I. (2015). The role of idioms in sentiment analysis. Expert Systems with Applications, 42(21), 7375–7385.

Zhang, S., Zhang, X., Chan, J., & Rosso, P. (2019). Irony detection via sentiment-based transfer learning. Information Processing & Management, 56(5), 1633–1644.