NATURAL LANGUAGE PROCESSING ANALYSIS OF COMMENTS ABOUT EDUCATION ON TWITTER DURING THE COVID-19

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF

HEALTHCARE SYSTEM ENGINEERING

OF ISTANBUL MEDIPOL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

HEALTHCARE SYSTEMS ENGINEERING

By

Lütviye Özge Polatlı

December, 2022

NATURAL LANGUAGE PROCESSING ANALYSIS OF COMMENTS ABOUT EDUCATION ON TWITTER DURING THE COVID-19

By Lütviye Özge Polatlı

21 December 2022

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

_____

Prof. Dr. Hakan Tozan (Advisor)

_____

Assist. Prof. Dr. Mustafa Yağımlı

_____

Assist. Prof. Dr. Kevser Banu Köse

Approved by the Graduate School of Engineering and Natural Sciences:

_____

Prof. Dr. Yasemin Yüksel Durmaz

Director of the Graduate School of Engineering and Natural Sciences

I hereby declare that all information in this document has been obtained and presented in accordance with the academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Signature :

Name, Surname: LÜTVİYE ÖZGE POLATLI

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my dear advisor Prof. Dr. Hakan TOZAN and Assist. Prof. Dr. Melis Almula Karadayı, who supported me all time, also who contributed greatly to the creation of the study, who supported and guided me throughout my academic studies. It's an honor to work with them.

I would like to thank my lovely family, who helped me get to where I am today, who are with me in every decision, who always support me. I am proud to be their daughter. Their endless love, endless patience, and endless support have helped me come to this day. I wouldn't have been such a strong woman without them.

I would like to thank my dear sister, Beril Isisağ, who I grew up with and who has always supported me as I got to where I am today. I am very lucky to have a sister like her in my life. I hope we celebrate many more successes together and always share our happiness. Finally, I would like to express my special thanks to my dear sister, Elif Delice, who supported me throughout the process. I hope we can achieve many more successes together.

Lütviye Özge Polatlı

December, 2022

CONTENTS

Page

ACKNOWLEDGEMENT ..... **iv**
CONTENTS…….. ..... **v**
LIST OF FIGURES ..... **vi**
LIST OF TABLES ..... **vii**
LIST OF SYMBOLS ..... **viii**
ABBREVIATIONS ..... **ix**
ÖZET ..... **x**
ABSTRACT ..... **xii**
1. INTRODUCTION **1**
1.1. COVID-19 In the World 2
1.2. Measures for COVID-19 3
1.3. COVID-19 Social Life Impacts 4
1.4. Education during the COVID-19 5
2. THEORETICAL PART **9**
2.1. Natural Language Processing 10
2.1.1. Natural language processing in health 12
2.1.2. Natural language processing in COVID-19 15
2.2. Topic Modeling 20
2.2.1. Topic modeling in health 23
2.2.2. Topic modeling in COVID-19 24
2.3. Sentiment Analysis 26
2.3.1. Sentiment analysis in health 27
2.3.2. Sentiment analysis in COVID-19 28
2.4. Topic Modelling and Sentiment Analysis 30
3. EXPERIMENTAL PART **33**
3.1. Octoparse 34
3.2. Transfer the Data to R 36
3.3. Pre-processing 36
3.4. Topic Modelling 39
3.4.1. Latent dirichlet allocation 43
3.4.1.1. How does latent dirichlet allocation work? 43
3.4.2. LDA application 47
3.5. Sentiment Analysis 48
3.6. Sentiment Analysis Application 49
4. RESULTS AND DISCUSSION **50**
4.1. Project Flow 51
4.2. Output of Topic Modelling 52
4.3. Output of Sentiment Analysis 53
5. CONCLUSIONS AND FUTURE WORK **57**
BIBLIOGRAPHY **59**
CURRICULUM VITAE **65**

LIST OF FIGURES

Figure 1.1: Mortality rates by current health problems 2
Figure 1.2: Daily new confirmed COVID-19 cases 3
Figure 1.3: Stay safe during COVID-19. 4
Figure 1.4: Share of countries implementing digital and broadcast remote learning policies, by education level 5
Figure 1.5: Percentage and number of students potentially reached and not reached by digital and broadcast remote learning policies 6
Figure 2.1: Literature review diagram. 10
Figure 3.1: Summary of data collection process and analysis. 33
Figure 3.2: Prepare Twitter URL. 35
Figure 3.3: Workflow created for Twitter. 35
Figure 3.4: LDA algorithm. 42
Figure 3.5: Distribution of tweets to the to word. 43
Figure 3.6: LDA parameters. 44
Figure 3.7: LDA flow. 44
Figure 3.8: Finding the frequencies of words. 45
Figure 3.9: Creating topic number. 46
Figure 3.10: Visualization and analysis stage. 46
Figure 3.11: Emotions in R. 47
Figure 4.1: Project flow diagram. 49
Figure 4.2: Output of LDA. 50
Figure 4.3: Output sentiment analysis for topic 1. 51
Figure 4.4: Output sentiment analysis for topic 2. 52
Figure 4.5: Output sentiment analysis for topic 3. 52
Figure 4.6: Output sentiment analysis for topic 4. 53

LIST OF TABLES

Table 1.1: SWOT Analysis. 7
Table 2.1: Natural language processing studies. 11
Table 2.2: Natural language processing studies in the healthcare field. 14
Table 2.3: Natural language processing in COVID-19 literature summary table. 18
Table 2.4: Topic modeling literature summary table. 21
Table 2.5: Topic modeling in COVID-19 literature summary table. 21
Table 2.6: Sentiment analysis literature summary table.. 21
Table 2.7: Sentiment analysis studies during the COVID-19. 21
Table 3.1: Twitter advanced search code. 34
Table 3.2: Read.CSV to Import Data in R 36
Table 3.3: Corpus in R. 36
Table 3.4: Remove punctuation in R. 36
Table 3.5: Strip whitespace in R. 37
Table 3.6: Tolower in R. 37
Table 3.7: Stopwords in R. 37
Table 3.8: Lemmatize in R. 37
Table 3.9: Removewords in R. 37
Table 3.10: RemoveNumbers in R. 38
Table 3.11: Before and after Data Output. 38
Table 3.12: Topic modeling method comparison. 39

LIST OF SYMBOLS

$\boldsymbol{k}$: Number of topics
$\boldsymbol{M}$: Number of texts
$\boldsymbol{z}$: Subject distribution of words
$\boldsymbol{\omega}$: Term
L : Text length
φ : Subject-word probability distribution
Θ : Text-topic probability distribution
$\boldsymbol{\alpha}$: Hyperparameter of Θ
$\boldsymbol{\beta}$: Hyperparameter of φ
$\boldsymbol{\beta_{ij}}$: The j -th word probability under the i

ABBREVIATIONS

WHO	**:** World Health Organization
NLP	**:** Natural Language Processing
BERT	**:** Bidirectional Encoder Representations from Transformers
MCC	**:** Matthews Correlation Coefficient
MAIL	**:** Malware analysis intermediate language
PPI	**:** Protein-Protein Interactions
NER	**:** Named-Entity Recognition
HPO	**:** Human Phenotype Ontology
CT	**:** Computed Tomography
ICD-O-M	**:** International Classification of Diseases for Oncology morphology classification standard
SARS	**:** Severe Acute Respiratory Syndrome
MERS	**:** Middle East Respiratory Syndrome
SAO	**:** Subject-Action- Object
LSA	**:** Latent Semantic Analysis
LSI	**:** Latent Semantic Indexing
HDP	**:** Hierarchical Dirichlet Process
LSVM	**:** Linear Support Vector Machine
AERTM	**:** Agriculture Named Entity Recognition using Topic Modelling Techniques
SS-LDA	**:** Sentence Segment LDA
DA	**:** Deterministic Annealing
pLSA	**:** Probabilistic Latent Semantic Analysis
GibbsLDA	**:** Gibbs Sampling with LDA
BTM	**:** Biterm Model
GSDMM	**:** Gibbs Sampling for Dirichlet Multinomial Mixture
MHANT	**:** Multi-Task Hierarchical Networks with Topic Attention
PHEIC	**:** Public Health Emergency of International Concern
PRSM	**:** Probabilistic Risk Stratification
SVM	**:** Support Vector Machine
LR	**:** Logistic Regression
EHR	**:** Electronic Health Records

COVID-19 SIRASINDA TWİTTER'DA EĞİTİM İLE İLGİLİ YORUMLARIN DOĞAL DİL İŞLEME ANALİZİ

ÖZET

Lütviye Özge Polatlı

Sağlık Sistemleri Mühendisliği, Yüksek Lisans

Tez Danışmanı: Prof. Dr. Hakan TOZAN

Eş Danışman: Dr. Öğr. Üyesi Melis Almula KARADAYI

Aralık, 2022

COVID-19'un ortaya çıkması, insanların günlük aktivitelerini yaşayamamasına, seyahat edememesine, çalışamamasına ve sosyal etkileşimlerin yaşanmamasına sebep olmuştur. Diğer birçok sektör gibi, eğitim sektöründe de dünya çapında öğrenciler ve eğitmenler ciddi sorunlar yaşamaktadır. Bu nedenle, COVID-19'un etkisini sınırlamak ve yaygınlaşmasını engellemek için eğitim kurumları okulları kapatmış ve akademik faaliyetlerini online platformlara taşımışlardır. Online eğitim süreci öğrenciler, eğitmenler ve veliler tarafından endişe ile karşılanmıştır. Bu süreçte, Doğal Dil işleme yöntemi kullanarak bu endişeleri sınıflandıran ve insanların yorum ve düşüncelerini çekinmeden paylaşabildikleri Twitter uygulaması alanında yapılan çalışmalar hızla artmıştır. Popüler sosyal medya programı olan Twitter'ın, dünya çapında 500 milyon kullanıcısı vardır. Dijital çağda internet, fikirlerin bilgilerin daha hızlı dolaşımını sağlamasıyla birlikte birçok doğru, yanlış, nefret söylemi olan fikir de hızla dolaşmaktadır. Twitter haberleri yaymak, dünya olayları hakkında fikir ve yorumları tartışmak için bir araç haline gelmiştir. COVID-19 salgını sırasında birçok yanlış bilgi, nefret söylemi toplumu korkutacak hileyeler ortaya çıkmıştır. Sokağa çıkma yasağı, maske takma zorunluluğu, evde çalışma, eğitimde aksaklıklar durumları, vb. insanların sosyal medyada duygu patlaması yaşamasına sebep olmuştur.

Çalışmanın amacı, COVID-19 sürecinde insanların uzaktan eğitim hakkında Twitter üzerinde yaptıkları yorumları analiz etmektir. Eğitim ile ilgili oluşturulan kelime bulutları Twitter üzerinden analiz edilmiştir. Octoparse program aracılığıyla 1 Ağustos 2020 ile 1 Ekim 2021 tarihleri arasında atılan tweetler analiz edilmiştir. Genel olarak öğrenciler evden çalışmanın rahat ve ulaşılabilir olmasından mutlu iken evde zaman yönetiminin olmaması, sosyal yaşantının az olması gibi problemlerle karşı karşıya kalmıştır. Ailelerin öğretmenlerin yüz yüze sınıflarda gerekli önemleri alıp öğrencileri sağlıklı bir ortam sunacaklarına inanmaktadır. Gelecek çalışmalarda çalışma ilköğretim lise ve üniversite öğrencileri için ayrı ayrı konu etiketleri oluşturup bakılabilir ve çalışma detaylandırılabilir

Sonuç olarak bu çalışma doğrultusunda öğrenciler evden çalışmanın rahat ve ulaşılabilir olmasından genel olarak mutlu olurken, evde zaman yönetimi ve sosyal ortam eksikliği gibi sorunlarla karşı karşıya kalmaktadırlar. Bu çalışma doğrultusunda doğal dil işleme alanında eğitimin çevrimiçi platforma aktarılması konusunda literatüre katkı sağlanması hedeflenmektedir. Tez kapsamında gerçekleştirilen uygulamanın sonuçları COVID-19

sürecinde eğitim politikası yapıcılarına ve ilgili araştırmacılara konu hakkında model sunarak ışık tutup gelecek çalışmalara rehberlik edebilir.

Anahtar sözcükler: COVID-19, eğitim, doğal dil işleme (NLP), konu modelleme, duygu analizi.

NATURAL LANGUAGE PROCESSING ANALYSIS OF COMMENTS ABOUT EDUCATION ON TWITTER DURING THE COVID-19

ABSTRACT

Lütviye Özge Polatlı

MSc in Healthcare Systems Engineering

Advisor: Prof. Dr. Hakan TOZAN

Co-Advisor: Asst. Prof. Dr. Melis Almula KARADAYI

December, 2022

The emergence of COVID-19 has caused people to be unable to live their daily activities, travel, work and social interactions. Like many other sectors, students, and educators around the world face serious problems in the education sector. Therefore, to limit the impact of COVID-19 and prevent its spread, educational institutions have closed schools and moved their academic activities to online platforms. The online education process has been met with concern by students, teachers, and parents. In this process, studies in the field of Twitter application, which classifies these concerns using Natural Language processing method and where people can share their comments and thoughts without hesitation, have increased rapidly. Twitter, the popular social media program, has 500 million users worldwide. In the digital age, as the internet allows ideas to circulate faster, many true, false and hate speech ideas are also circulating rapidly. Twitter has become a tool for disseminating news and discussing opinions and comments on world events. During the COVID-19 pandemic, many false information, hate speech, and tricks to scare the society have emerged. Curfew, obligation to wear a mask, working at home, disruptions in education, etc. It has caused people to experience an explosion of emotions on social media.

The purpose of the study is to analyze the comments people make on Twitter about distance education during the COVID-19 process. The word clouds created about education were analyzed with the support of Twitter data. Tweets sent between August 1, 2020, and October 1, 2021, through the Octoparse program were analyzed. In general, while students are happy that working from home is comfortable and accessible, they are faced with problems such as difficulty in time management at home and lack of social life. Families believe that teachers will take the necessary importance in face-to-face classes and provide students with a healthy environment. In future studies, the study can be created and looked at separately for primary school high school and university students and the study can be detailed.

As a result, in line with this study, while students are generally happy that working from home is comfortable and accessible, they face problems such as time management and lack of social environment at home. In line with this study, it is aimed to contribute to the literature on the transfer of education in the field of natural language processing to the online platform. The results of the application carried out within the scope of the thesis

can guide the future studies by presenting a model to the education policy makers and related researchers in the COVID-19 process.

Keywords: COVID-19, Education, Natural Language Processing (NLP), Topic Modeling, Sentiment Analysis.

CHAPTER 1

1. INTRODUCTION

As of July 10, 2020, while the COVID-19 disease was spreading, a state of emergency was declared by the World Health covert international health regulation emergency committee on January 30, 2020. The outbreak of COVID-19 has caused people to be unable to live their daily activities, travel, work, and social interactions. Like many other industries, the education system is facing serious problems for students and educators around the world. To limit the impact of COVID-19 and prevent its spread, educational institutions have closed schools and moved academic activities to their online platforms, and online education has created great concerns for both students, educators, and parents. In this process, the Natural Language Processing (NLP) method, to classify these concerns, studies in the field of Twitter application, where people can share their comments without hesitation, its value is increasing. Twitter, the popular social media program, has 500 million users worldwide. In the digital age, as the internet allows ideas to circulate faster, many are circulating true, false, and hate speech. It has become a tool for disseminating the news, discussing ideas and interpretations about world events. During the COVID-19 epidemic, much false information, hate speech, and tricks that will scare society have emerged. For this reason, studies such as the analysis of fake news, emotion classification, subject extraction have gained value. The curfew, the obligation to wear masks, working at home, and disruptions in education have caused people to experience an explosion of emotions on social media. The main purpose of this study is to analyze the comments made by people on Twitter about distance education during the COVID-19 process.

1.1. COVID-19 In the World

Coronavirus disease is a contagious disease. It has been observed that people with coronavirus disease have mild to moderate respiratory problems. However, some patients may become seriously ill and need medical attention. **Figure 1.1** shows these possibilities according to the resource Chinese Center for disease control and prevention. When the risk groups were examined, the death rates of cardiovascular and diabetes patients were found to be high according to the result of death rates according to existing health problems.

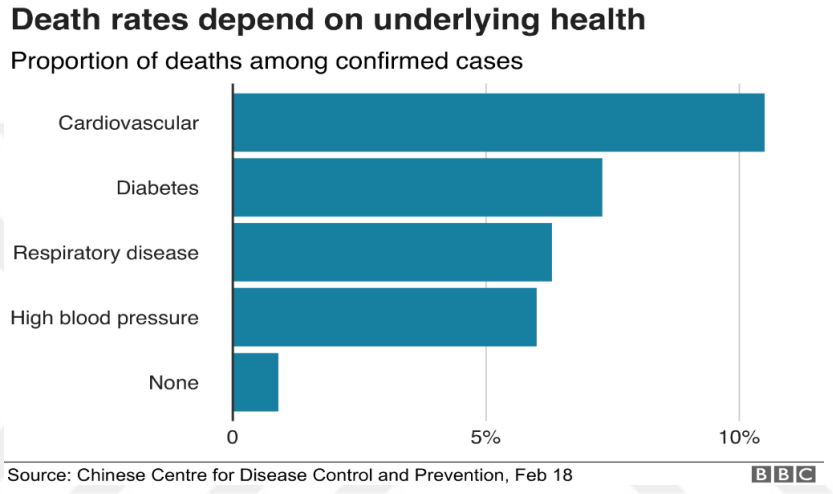

Figure 1.1: Mortality rates by current health problems [1].

According to the "Our world in data" website, where the COVID-19 data are compiled, the number of confirmed COVID-19 deaths worldwide reached 415 per million as of May 5. The country with the highest death rate by population factor has proven to be Hungary with a COVID-19 rate of 2 thousand 916 people per million. The country with the highest death rate by population factor has proven to be Hungary with a COVID-19 rate of 2 thousand 916 people per million. Czechia with 2 thousand 752, Bosnia and Herzegovina with 2 thousand 662, Bulgaria with 2,413, North Macedonia with 2 thousand 407, Slovakia with 2 thousand 177, Belgium with 2 thousand 105, Slovenia with 2 thousand 58, Italy with 2 thousand 17. It is noteworthy that Eastern European and Balkan countries are dominant among the countries with a high death rate according to population. Italy and Belgium from Western Europe were among the top 10 countries that included COVID-19 deaths, even if they were not tested. **Figure 1.2** shows Daily new confirmed COVID-19 cases. The proportion of countries with a positive test result is shown with the

blue line. The proportion of countries where the number of infections may be high is shown with a red line [2].

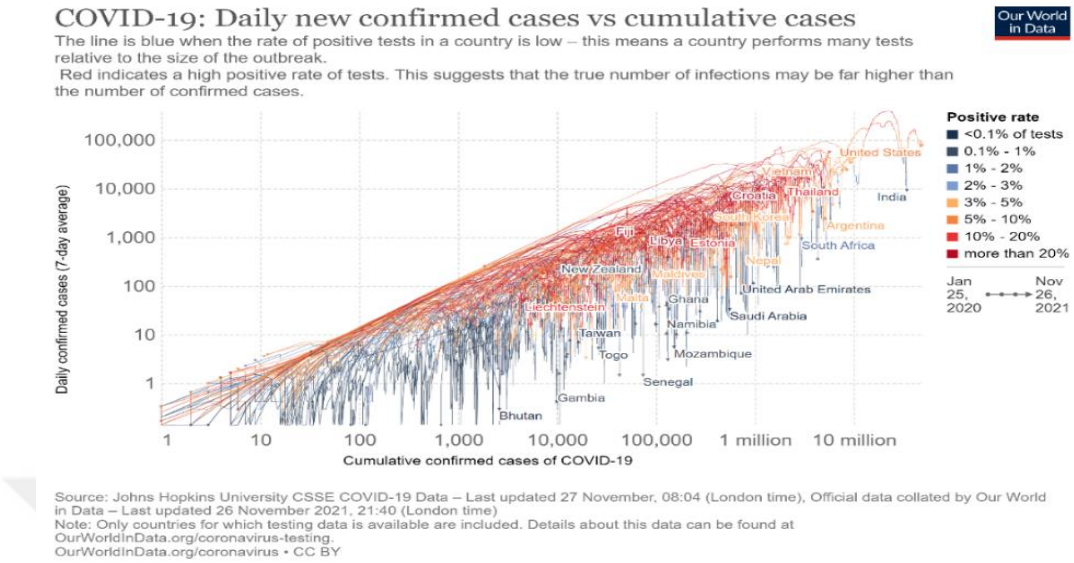

Figure 1.2: Daily new confirmed COVID-19 cases [3].

1.2. Measures for COVID-19

The World Health Organization has created a list of precautions for COVID-19. In this list, People should stay 1 meter away in social areas. WHO stated that masks should be worn at points where one cannot physically move away and wash hands. WHO has proven that the disease is more contagious indoors, in crowded places, with poor ventilation. If people infected with COVID-19 have touched it, people should avoid touching surfaces, especially in public settings or healthcare facilities. The disease is more contagious in crowded places, places with poor ventilation, and indoors. In cases such as coughing and sneezing, it should be covered with an elbow or tissue paper and used tissues should be immediately thrown into a closed trash bin. Together with, the World Health Organization emphasizes that people should be vaccinated [4]. The World Health Organization has prepared these issues as posters to inform the public (see **Figure 1.3**).

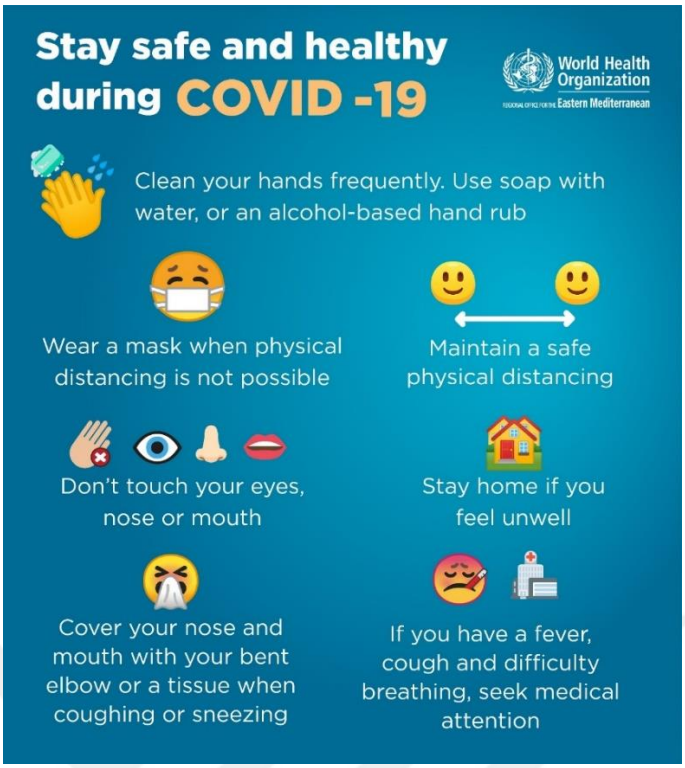

Figure 1.3: Stay safe during COVID-19 [5].

1.3. COVID-19 Social Life Impacts

COVID-19 process, great restrictions have come to people's social lives. Elderly people, people with disabilities, families, students, and employees have been psychologically affected in this process. Disabled people, on the other hand, faced problems such as interruption of services and support during the COVID-19 process. While families were worried about their children's education, students faced psychological problems caused by not being socialized. Its psychological and social effects on people have also caused deterioration. According to the researchers' research, students and health professionals emphasized that they were in a period when they could show symptoms. Social distancing and safety measures have affected the relationship between people and their perception of empathy towards others. The study included a sample of 1,143 parents with Italian and Hispanic children (range 3-18). Overall, parents have observed emotional and behavioral changes in their children who are quarantined at home during the COVID-19 process. Difficulty concentrating (76.6%), boredom (52%), nervousness (39%), restlessness (38.8%), irritability (38%), feeling of loneliness (31.3%), restlessness (30.4%) and anxiety (30.1%). From the comparison between the two groups (Hispanic and Italian parents), it was revealed that Italian parents reported more symptoms in their children

than Hispanic parents. Further data collected on a sample of university students during the epidemic in China showed that anxiety levels in young adults were mediated by certain protective factors, such as living in urban areas and family economic stability[6].

1.4. Education during the COVID-19

In the Scientific Advisory Board Study published by the Ministry of Health on March 9, 2021, under the title of Study Guide and Infection Control Measures in Health Institutions, the training given during the COVID-19 pandemic should be updated and continued. If possible, the training should be done online, recorded [7]. With the COVID-19 disease, it was decided to close the schools. More than 1 billion children continue their education online as schools are closed. However, poor families, students without internet access, and students without personal computers faced great difficulties in this process. This interruption of education worldwide and the disruption it has suffered during this period has risked undoing years of progress. With schools closing in 188 countries (as of April 2020), education has started to move to online platforms. Alternative ways of providing continuing education have been tried using technologies such as the Internet, TV and radio. (see **Figure 1.4).**

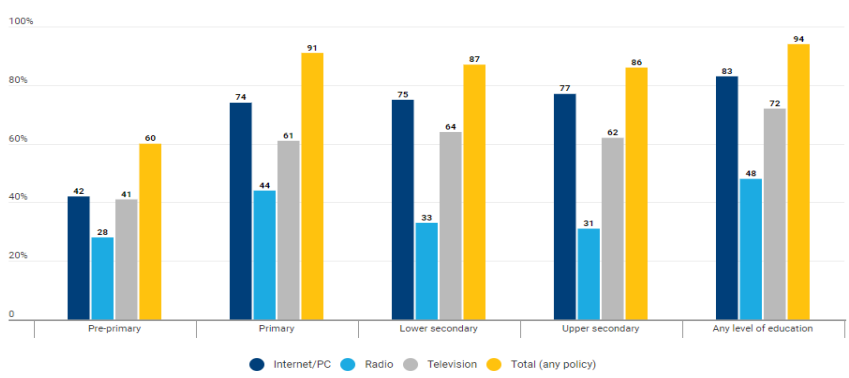

Figure 1.4: Share of countries implementing digital and broadcast remote learning policies, by education level [8].

Many studies have been carried out on the access of students in rural areas to education. It was observed that three of the four inaccessible students live in rural areas. Low-income families' lack of access to education has been studied. Among the out-of-reach students, male and female students were almost equally represented. Worldwide, 31% of schoolchildren do not have access to education (see **Figure 1.5)**

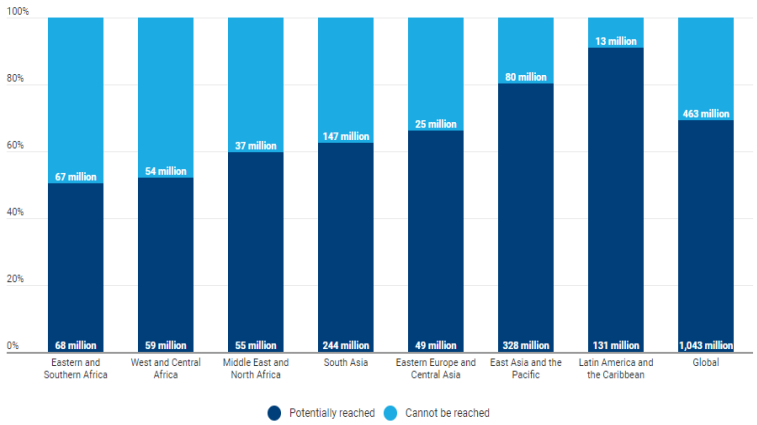

Figure 1.5: Percentages of students who could not reach education [8].

In the article, the researchers conducted a study in which they summarized the strengths, weaknesses, opportunities, and threats of the e-Learning concept. They interpreted distance learning and face-to-face education under the headings of strengths, weaknesses, opportunities and threats [9]. SWOT analysis results are presented in **Table 1.1.** There are many points that parents can worry about for their children. Problems such as the threats posed by distance education, internet problem, connection problem of the teacher, or late attendance of the student to the lesson may occur. The student may have problems communicating with the teacher or there may be situations where the teacher cannot reach the student. Online courses may not be safe for families. In addition, the weaknesses provided by distance education have led to problems such as the difficulty of individual work, the lack of the ability of the student to work in groups, the atrophy of the student's abilities, or the lack of recognition of his abilities. These problems have increased the concerns of families. There are also distance learning opportunities and strengths. Students have advantages such as time management is their own, education is fast and accessible. With education being extended, people have experienced a complexity of emotions about this topic. Educators, families, and students experience many emotional changes about this issue.

Table 1.1: SWOT Analysis.

Strengths	Weaknesses
• Easy access to the internet Ability to manage time • Ability to work independently • Flexibility in the learning program • Having more time for reduced physical activity	• Lack of discipline brought by individual work • Unable to create a daily schedule • Lack of access to course materials (computer cost, tablet cost, etc.) • Lack of ability to use technological tools. inability to adapt to technology (such as not knowing how to use a computer) Lack of access and communication to the teacher and other students • Loss of group workability
Opportunities	**Threats**
• Open and quickly accessible education • Making distance virtual education fun by gamification • Students have no fear of asking questions in the virtual environment • It is easy to balance non-educational work with time. • Students learn information about how to use online teaching software	• Parents may not want to pay for their paid online courses • Buying online education may not be safe. • Students may encounter problems such as lack of communication. The voice is not clear due to the teacher's internet problem • Classes not starting on time due to internet problem • Failure of students to attend classes due to internet problems

Section 2 presents the literature review, was conducted under three main headings: NLP, subject modeling, and sentiment analysis. It is aimed to find the most used database in

NLP studies. In addition, it is aimed to examine the place of the most used and best performing method in subject modeling in the literature.

Final section concludes the study, it is expected to examine the results of feeling the attitude of students working from home. The results of the application carried out within the scope of the thesis can guide the future studies by presenting a model to the education policy makers and related researchers in the COVID-19 process.

CHAPTER 2

2. THEORETICAL PART

NLP maximizes the human-computer relationship. It aims for computers to understand, process, and interpret this information in spoken language and to produce new information based on this information. Today, bank chatbots, cell phone assistants, Google translation, cell phone next word prediction, and other technologies rely on NLP. Recently popular text mining has been incorporated into NLP. The document creates a word cloud in the text and paragraph, creating the logical theme of the document. Topic modeling, determining the most talked about topics on the Twitter agenda, analysis of restaurant comments, the most talked about topic in the restaurant, such as cleanliness and taste, may come up. It is seen that Trendyol products use the word cloud technique in NLP for comment clustering. Topics consisting of the most spoken words about a product can be filtered. When examining the reviews of a phone product in Trendyol, it is possible to encounter word clouds such as charge, battery life, camera quality. By clicking on these word clouds created by filtering among the comments, all comments containing that word are received. In this way, when people buy a phone, they can read consumer comments about its charge. In this literature review, research on topic modeling was collected. Focus on research on the global COVID-19 pandemic. Aim to find the topics that people talk about the most about COVID-19. For the COVID-19 topics that people talk about the most, how they feel about these issues is significant.

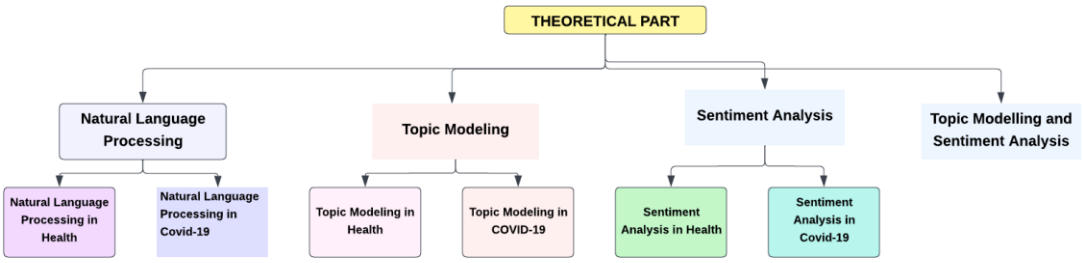

Figure 2.1: Literature review diagram.

Google Academic, Web of Science, Istanbul Medipol University Library-Electronic Information Resources databases were scanned with COVID-19 keywords. While examining the articles, natural language processing studies in the field of COVID-19 were examined. Current studies between 2020 and 2021 are summarized. When the studies in NLP are examined, it can be said that the most used database in studies is Twitter and the most researched subject in studies in the field of health is COVID-19. Studies show that LDA is the most used method among these methods and the method with the best performance is LDA. A diagram of the literature review was created (see **Figure 2.1)**.

2.1. Natural Language Processing

NLP technologies such as the chatbots we are talking about on the bank website today, the commands we send from our mobile phones to the assistant, the translations we make in google/Microsoft translate, the prediction of the next word from the mobile phone while composing. it's all-NLP. Thanks to text mining, many accumulated ideas on the Internet can be processed and analyzed. **Table 2.1**

Kasthuri et al. (2018) make a chatbot using the NLP method. Students can ask questions to the chatbot. After cleaning the data with NLP, questions that students can ask with the training set were created with IF-Else methods. The chat robot they created was able to grasp high-level questions without the need for human interaction, and they aimed to answer students' questions [10].

Chan et al. (2021) analyzed the relationship between entrepreneurship researchers' document writing quality of NLP techniques and new venture financing outcomes. Their main purpose is to predict crowdfunding results. It completed the study by using information such as campaign web pages, requested financing, duration, etc. To predict and understand funding outcomes within crowdfunding projects. They used 1359

crowdfunding campaign texts. risk definitions are in training to attract less funds and supporters [11].

Alam. (2021) used the SIMP model they developed to find vicious patterns applications. SIMP provides control flow patterns to improve MAIL analysis. It makes the code accessible to NLP techniques to check for semantic similarities. The dataset consists of 2023 Android applications. Of these, 1023 are data collected for the Android, collected from two different sources. The proposed model is real malware and benign programs using different verification methods. When tested with benign Android apps, it achieved an MCC of 0.94 between the actual and predicted values. Based on the results of the outputs, proved to be predicted to be malicious or benign with a high success rate [12].

Perboli et al. (2021) as the current standard in aviation accidents, performed by trained personnel, there are currently no defined technical standards for automatic human factors identification. Using these standards, they classified the cause of aviation accidents with the NLP method. They tested on case studies of security events of varying severity, including system failures, minor incidents, and serious incidents involving loss of human life. They also estimated for at least 30% when applying the methodology to real documents controlled by experts compared to standard human factors identification methods [13].

Table 2.1: Natural language processing studies.

Author(s)	**Problem**	**Results**
Kasthuri et al. (2018)	Creating a chatbot.	They have created a chat robot where students can ask questions.
Alam (2021)	Providing control flow patterns to improve analysis	A new sample has been proven to be predicted as malicious or benign with a high success rate.
Perboli et al. (2021)	Finding the cause of aviation accidents using standards.	The cost and time were reduced by 30% and the tests were 86% accurate.
Chan et al. (2021)	Forecasting crowdfunding results	Risk definitions are in training to attract less funds and supporters.

NLP is the transfer of texts and sounds to computer language. The most striking areas in NLP were chatbot creation and predicting the next text.

2.1.1. Natural language processing in health

It is observed that research in the health sector has started to increase with the developing technology these days. In the health sector, studies on analyzing patient data using patient data are increasing. The field of NLP health is summarized in **Table 2.2.**

Doan et al. (2018), created an approach to extract causal relationships from tweets using NLP techniques. They focused on three health-related issues: “stress”, “insomnia” and “headache”. 501 results for stress out of 29705 (1.6%), 72/3827 (1.8%) for insomnia, and 94/11252 (0.8%) for headache. they found. They proved that the approximation reached an average accuracy between 74.59% and 92.27% [14].

Van Le et al. (2018), created a complementary risk estimation method in psychiatry inpatients. They used NLP to predict patients' endpoints of self-harm or victimization. The dataset in electronic health records consists of unidentified inpatient notes. They performed an automated analysis of case notes. They evaluated the presence or absence and frequency of words in the Electronic Health Records dataset. Seven machine learning algorithms created the risk assessment score. They used seven machine learning algorithms. (Bagging, J48, Grip, Logistic Model Trees, Logistic Regression, Linear Regression, and Support Vector Machine). Logistic model trees and Support Vector Machine gave the best results [15].

Lee et al (2020), used a dataset of patients with serious illnesses between 2008 and 2016. In this data set, 1435 patients are treated as inpatients. 1748 patients are treated as outpatients. They aimed to create a new study to measure patient care goals discussions using NLP and machine learning algorithms. They tested the performance with training and test sets, creating a quality criterion for maintenance. They repeated this study for the inpatient-only and outpatient-only subsets. As a result, 689 of the 3183 notes contain documentation of the care goals discussions. They found NLP mean sensitivity 82.3% and mean specificity 97.4%, the median positive likelihood ratio of NLP results was 32.2 and the median negative likelihood ratio was 0.18. They found better performance than inpatient-only samples compared to outpatient-only samples [16].

Kulshrestha et al. (2020), used retrospective trauma injury severity scores from the electronic health record. They have provided automatic injury scoring of clinical

documents with NLP. They proved that can distinguish cases of severe and non-severe post-traumatic chest injury with NLP. They examined documents consisting of patient records from the trauma center between 2014 and 2018. Severe chest injury was determined with a thorax shortened injury score greater than two and used as a reference standard for supervised learning Trauma documents consisting of 473,694 documents were examined. The output results found the trained dataset AUROC value to be 0.88 [17].

Forestiero et al. (2020), Doc2Vec was used. Doc2Vec creates vectors linked to documents, Word2vec creates vectors for words. They analyzed these two approaches using records from a clinic. User requests can reach the server containing content and service efficiently. Experimental results concluded that the effectiveness of the approach is proportional to how discovery processes become faster and the square root of the network size [18].

Mathews et al. (2021), analyzed the relationship between humans and Yersinia pestis. With the model they developed, they obtained a score of 0.92. They analyzed it in the framework of end-to-end deep learning with neural machine tools. They created a new model by combining the new preprocessing methods in the field of bioinformatics. It is demonstrated by the Human and Yersinia pestis PPI (Protein-Protein Interactions) that it provides comparable performance to other datasets using innovative methods with little preprocessing and hyperparameter tuning [19].

Solomon et al. (2021), develop and validate NLP algorithms to identify cases of aortic stenosis and associated parameters from semi-structured echocardiography reports. They described cases of aortic stenosis using a NLP algorithm and an echocardiography database. They used NLP, identifying a total of 927,884 eligible echocardiograms among 519,967 patients to obtain positive and negative predictive values. Echocardiography was classified as 104,090 (11.2%) patients with aortic stenosis [20].

Kormilitzin et al (2021) the Named-Entity Recognition (NER) modelWhile creating the model, 7 categories were examined. The model achieved an F1 score of 0.957 in all seven categories. They concluded that the model developed by examining the data from the Intensive Care Unit in the USA could be transferred to secondary mental health records in England [22].

Olthof et al. (2021) made comparisons using NLP to classify radiology reports for the presence of injury in orthopedic traumas. The NLP classification has been implemented and optimized by testing with Rule-based, Machine learning, and Bidirectional Encoder Representations from Transformers (BERT). The deep learning-based BERT model outperformed all other evaluated classification methods. They found that the deep learning-based BERT model outperformed all other evaluated classification methods [23].

Parikh et al. (2021) filtered the Human Phenotype Ontology (HPO) terms extracted from the NLP to resemble the manually extracted terms more closely and determined the filter parameters. Boston Children's Hospital used the dataset. Filtering resulted in NLP-based extraction of HPO terms in 92% of prospectively evaluated cases and was sufficient for manual subtraction. As a result, they developed methods to optimize NLP outputs for automatic diagnosis using electronic health records [24].

Table 2.2: Natural language processing studies in the healthcare field.

Author	**Problem**	**Result**
Forestiero et al. (2020)	To ensure efficient use of the targets of user requests and the server where the service is located.	Conclude that the effectiveness of the approach is how discovery processes become faster.
Mathews et al. (2021)	Analyzed the relationship between humans and Yersinia pestis (Bacteria)	It compares favorably with the study, which has a model that yields independent testing.
Solomon et al. (2021)	Identifying cases of aortic stenosis and associated parameters	Identified a total of 927,884 appropriate echocardiograms. Echocardiography was classified as 104,090 (11.2%) patients with aortic stenosis.
Olthof et al. (2021)	Classifying radiology reports	BERT model outperformed all other evaluated classification methods.

NLP is the transfer of text and sounds to the computer in a way that the human brain can perceive. NLP studies have been used efficiently for health research in recent years. When we examine the current studies in NLP health sector, studies in many fields such as classification of diseases, estimation of the next patient arrival, classification of reports, analysis of hospital standards, analysis of hospital service comments draw attention.

2.1.2. Natural language processing in COVID-19

World Health Organization declared COVID-19 a pandemic and a global health crisis. [25] Many studies have been started in the field of NLP COVID-19. In NLP, studies in the field of Twitter application, where people can share their comments without hesitation, are increasing. It became a vehicle for disseminating the news, discussing ideas and interpretations about world events. During the COVID-19 epidemic, much false information, hate speech, and tricks that will scare society emerged. In the digital age, with the internet allowing ideas to circulate faster, many ideas that are right, wrong, wrong, and hate speech are circulating. For this reason, studies such as fake news, emotion classification, topic extraction gained value. The curfew, the obligation to wear masks, working at home, and disruptions in education caused people to experience an explosion of emotions on social media. For this purpose, NLP studies, in which comments about COVID-19 were made in general within the scope of the compilation study, are summarized in **Table 2.3**.

Massaad et al. (2020) people's psychological problems emerged during the COVID-19 process. Aimed to analyze the volume, content, and geographic distribution of telehealth-related tweets during the COVID-19 pandemic. To access telehealth-related tweets from March 30, 2020, to April 6, 2020, Twitter polled public data and analyzed 41,328 tweets in total. They pulled tweets containing terms such as "telehealth" and the most used terms "covid", "health", "care", "services", "patients", and "pandemic" and analyzed them using NLP. In conclusion, mental health was the most prevalent health-related topic to emerge in this research, reflecting the high need for mental health services during the pandemic [26].

Ittamalla et al. (2020) using machine learning techniques, aimed to understand the feelings of the public towards contact tracing and to determine whether there was a change in the general public's thinking in the various crisis. Compared to the first month of the crises, they observed a significant increase in negative emotions related to digital

contact tracing and a decrease in neutral emotions in the following months. Study 2 revealed that the major problems that the public voiced with their negative emotions were a violation of privacy, fear of security, and lack of trust in government [27].

Chapman et al. (2020) developed NLP line to identify positively diagnosed COVID 19 patients. With the information received from the clinic, they created a system from the diagnoses found in the patient. The system was able to detect 6,360 positive cases. They estimated the performance of the system to be 82.4% precision. The outcome of this study for the United States Department of Veterans Affairs has accelerated the review of patient charts in the veteran's affairs surveillance system, where 36.1% of confirmed positive cases are identified using this system [28].

Ebadi et al. (2020) used PubMed and ArXiv as data sources in this study. They created a model to characterize COVID-19 research. They used publication similarity and sentiment data within the January-May 2020 time frame. As a result, they observed that the types of research available in PubMed and ArXiv differed significantly. They concluded that more focus should be placed on smart systems/tools to predict/diagnose COVID-19 [29].

Izquierdo et al. (2020) aimed to create a system that defines COVID-19 patients and predicts whether they should be hospitalized in the intensive care unit. They collected electronic health records from January 1 to March 29. Based on the combination of age, fever, and respiratory tract problems, they decided whether to be admitted to the intensive care unit or not. In conclusion, 6.1% (83/1353) of hospitalized patients should be admitted to the intensive care unit [30].

Banda et al. (2020), COVID-19 patients reported persistent symptoms three months after recovery. They aimed to find the most frequently mentioned symptoms after the disease. The follow-up of the patients in the clinical records was documented, but since this information was not public, they examined the comments made by the patients about the symptoms on Twitter. The most common symptom rates in the disease were fatigue and fatigue [31].

Hu et al. (2020 to analyze the studies in this field with the NLP method. They compared the bibliometric indices between COVID-19, SARS, and MERS up to 25 March 2020. They produced 1,480 documents for analysis. They pulled data including research publications from sources such as Scopus and PubMed. In total, they selected 13,945

research literature from 7 datasets for analysis. As a result, Weighted Citation Effect scores increased to 3.68, 6.63, and 11.35, respectively, for significant growth in research qualifications during these three outbreaks [32].

Liu et al. (2021), compare people's thoughts, concerns, behavior changes, discussion topics by applying NLP to data related to COVID-19. They analyzed the comments made about the COVID-19 outbreak of the Reddit social media platform. Between March and August 2020, they created word clouds of North Carolina's 18 subreddits and collected data on COVID-19. As a result, They observed that the public's attitudes towards wearing masks were positive [33].

Gupta et al. (2021) analyzes the comments made about COVID-19 on Twitter. They created an emotional care plan and a web-based platform to recognize the emotional state of people during the ongoing COVID-19 crisis. They divided them into categories. They found that the education system results in more trust (~29%) in terms of reputation. The health sector has found the result of attacks (~16% and fear (~18%) among the masses) because of fear of people [34].

Kasson et al. (2021), the negative health consequences associated with electronic cigarettes during the COVID-19 process created uncertainty and controversy on social media platforms. They aimed to identify at-risk electronic cigarette users. They extracted and analyzed a total of 794,620 comments on Twitter about electronic cigarettes with NLP methods. They concluded that the public was concerned about this issue [35].

Huerta et al. (2021), analyzed the tweets posted between the dates of the State of Emergency in Massachusetts, the closure of public schools, and the US State of Emergency, using NLP. They observed that with the increase in health discussions, the anxiety in people increased [36].

Klein et al. (2021) aimed to identify potential non-test-based COVID-19. By developing the NLP method, they developed an algorithm that can find the potential patient based on the tweets on Twitter. From January 23, 2020, they collected tweets in English on the Twitter app mentioning keywords related to COVID-19. They created a word cloud on twitter for COVID-19 and analyzed it with the BERT approach and found the Evaluator trustworthiness 0.77 based on binary annotations for 3644 (41%) of 8,976 tweets. [37].

Tang et al. (2021) used NLP to find out the level of public participation in the decisions made by public health institutions in Texas. Public health agencies started to work

towards the conclusion that they should or should not use Twitter to share information and to encourage the public. In the first 6 months of 2020, they examined all COVID-19-related tweets sent institutions and whether the tweet provided information. They categorize them as to whether they encourage action or build community. As a result, tweets that share information are more likely to be liked [38].

Table 2.3: Natural language processing in COVID-19 literature summary table.

Author	**Data**	**Problem**	**Result**
Oyebode et al. (2020)	Twitter, YouTube, Facebook, PushSquare.com, Archinect.com LiveScience.com	Analyzing people's impact on COVID-19 using social networks.	They have obtained negative comments on topics such as economic, socio-political and educational.
Izquierdo et al. (2020)	EHR	ICU admission analysis of COVID-19 patients	6.1% (83/1353) of hospitalized patients should be admitted to the intensive care unit
Hu et al. (2020)	Scopus, PubMed	Analyzing studies on SARS, MERS, and COVID-19	SARS, MERS, and COVID-19 n Weighted Citation Effect scores increased to 3.68, 6.63, and 11.35, respectively.
Ittamalla et al. (2021)	Reddit	To understand the public's feelings towards contact tracing	They observed a significant increase in negative emotions related to digital contact tracing and a decrease in neutral emotions in the following months.

Liu et al. (2021)	Reddit	Comparing the topics of people's thinking about COVID-19	The public's attitudes towards wearing masks were positive
Kasson et al. (2021)	Twitter	Identifying at-risk electronic cigarette users	The public was concerned about the risk electronic cigarette users
Banda et al. (2020).	Twitter	Identifying ten remaining symptoms of illness after recovery from COVID-19	The 10 most frequently mentioned symptoms are; weakness and fatigue, shortness of breath, palpitations, chest pain, insomnia/sleep disturbances, cough
Tang et al (2021)	Twitter	Using twitter to share health information	Twitter should be used to disseminate information.

When the NLP studies in the field of COVID-19 are examined, it is observed that there are many studies on fake news, emotion classification, subject extraction, ideas about curfew, thoughts about the obligation to wear masks, feelings of working from home, news of disruptions in education. In the digital age, with the internet allowing ideas to circulate faster, many articles about COVID-19 have emerged, with many ideas circulating right, wrong, wrong, and hate speech. Researchers, on the other hand, analyzed these articles and produced studies. As a result, when the studies in NLP are examined, it can be said that the most used database in the studies is Twitter and the most researched subject in the studies in the field of health is COVID-19.

2.2. Topic Modeling

Topic modeling, one of the NLP methods, is used to train machine learning models. Topic modeling is part of statistical text analysis. It is a method that analyzes text documents and aims to find the topics in the text. It analyzes the distribution of the words in the text and creates the topics based on the most repeated words. By looking at the comments on this survey and categorizing them under a topic such as the most common reasons for

their low ratings, their analysis can be followed quickly, and they can begin the improvement process on that issue. In this study, the topic modeling studies carried out in recent years are summarized in **Table 2.4.**

Pathan et al. (2019) made a comparative analysis of the most popular algorithms, LDA, LSA, and HDP algorithms within the scope of topic modeling. This study, which was examined with three models, also showed a good performance in three models. The consistency value of the LDA method had better results than LSA and HDP [39].

Ma et al. (2021) aimed to analyze the Derwent Innovation Index patent to present its potential contribution to R&D management. This study, which was analyzed in two ways, firstly found patent classification codes and term positions. Second, they integrated the SAO technique to explore semantic relationships between topics for these terms and to model the topic. They created the optimized model and made it applicable to extract subjects from DSSC patents. Complementing the shortcomings of SAO analysis, they defined semantic analysis and subject relationship [40].

Jisoo Ahn et al. (2021) analyzed tweets in categories such as agencies, media companies, non-profit organizations during the 2019 California earthquake. The fact that the categories were different made the tweets consist of different topics. Analysis of 9,391 tweets revealed that media companies' tweets on topics related to video materials and "others" on support and preparation generated more retweets and favorites than any other topic. LDA (Latent Dirichlet Allocation) method was used for analysis. These findings of similar and different models based on organization type and public participation may be helpful for social media managers in the future when communicating with the public during future disasters [41].

Pérez et al. (2021) They used two types of software models were created. They used the LDA approach for code generation and interpretation. They developed a software model for the interpretation part of the LDA. They argued that Latent Semantic Indexing-based performs well. As a result, they showed that calibration approaches can be transferred from code to models [42].

Xie et al. (2021), Text analysis has two main limitations. First one limit is these methods often ignore contextual information of texts and have limited feature representation capability due to shallow feedforward network architecture, the second limit is the sparseness of representations in the subject semantic field is ignored. Addressing these

problems, this study also proposes a semantic reinforcement neural variable sparse topic model for the learning of explicable and sparse latent semantic representation. The proposed model has resulted in success [43].

Gangadharan et al. (2020), this study aims to classify crop diseases using fertilizer keywords. Agricultural dictionary AGROVOC and AERTM have presented a hybrid approach using the algorithm, but this hybrid study failed to detect Soil Types, Crop Diseases, and Fertilizers. Therefore, they proposed a LDA based topic modeling algorithm for these entities. The algorithm has been tested using 3000 sentences. This method can be used for answering questions in agricultural forums, and as a result of human evaluation of the method, 80% accuracy was achieved [44].

Goshima et al. (2018) proposed a new approach to improve Probabilistic latent semantic analysis (pLSA) performance, powered by a Deterministic annealing (DA) process, in which pLSA solutions are handled in the context of fuzzy co-clustering. The proposed DA treatment is performed by adjusting the intrinsic turbidity of the pLSA, and pLSA solutions are expected to improve initiation sensitivity. As a result, improve the quality of pLSA splitting, they fixed the final turbidity at 1, which is equivalent to pLSA. However, they suggested that the best common cluster division could be used at a different stage than the pLSA stage [45].

Cao et al. (2017), Mashup is a software development method. They used topic modeling to improve the quality of this software method. In which the LDA method was used, the Mashup method, which has high accuracy and good efficiency, was proposed. When mashup is compared to other methods, it has been shown that it provides improvement in some aspects such as precision. As a result of the Mashup method, Mashup helps improve the quality of service discovery and Mashup-based software development [46].

Table 2.4: Topic modeling literature summary table.

Author	**Method**	**Problem**	**Result**
Pathan et al. (2019)	LDA, LSA, HDP	Comparing the result of three different Topic modeling methods	The LDA method had better results.

Vasily Alekseev et al. (2021)	Multiple Model training	Suggesting a model validation method	Created a "Topic bank".
Ma et al. (2021)	SAO	Proposing a hybrid methodology analyzing the Derwent Innovation Index patent	Found shortcomings of SAO analysis and defined semantic analysis and topic relationship.
Jisoo Ahn et al. (2021)	LDA	Analyzing tweets posted during the 2019 California earthquake	Tweets about topics related to media companies' video materials and support and preparation from “others” generated more retweets and favorites than other topics.
Pérez et al. (2021)	LDA, LSI	Building two types of models in an industrial case study	It showed that LDA significantly outperformed the LSI method.

Topic modeling is one of the NLP methods used to train machine learning models. When topic modeling studies are examined, it is observed that a company that wants to identify improvement areas is frequently used in a survey that asks users to rate their services and comment, in Tweet analysis, in special reports, in finding the most talked about topic in restaurants, in finding the most talked about topic in hotels, and in hybrid studies. Topic modeling is one of the methods of analyzing text documents to find topics in the text as it is part of statistical text analysis. By analyzing the distribution of words in the text, themes are created on the most repeated words. For this reason, companies use this method to improve themselves in quality improvement. Studies are showing that LDA is the most used method among this method and that LDA is the method with the best performance.

2.2.1. Topic modeling in health

Although topic modeling is a very new method in the health sector, it is among the methods preferred by hospitals and pharmaceutical industries to improve the system for patients, for reports, and recently. Although the advantage of social media increases the quality of health services, companies that have started to consider the comments on social media have started to be successful. In conclusion, health sector is summarized.

Antonio et al. (2021) made a topic modeling study using the data received via Twitter and e-mail. They compared different datasets to explore health-related questions. TF-IDF and Doc2 Vec document vectorizations, LSI, LDA, GibbsLDA, Online LDA, Biterm Model (BTM) For Twitter Online LDA and Gibbs Sampling, they set up the Mixture Gibbs Sampling for Dirichlet Multinomial Dirichlet Multinomial Mixture (GSDMM) models. Twitter Online recommends LDA and GSDMM as best, while external indices are LSI and k-means and TF-means at best. He suggested the IDF [47].

Zhou et al. (2021) established a multi-task hierarchical network with topical attention (MHANT) model for health problem identification based on social media data. They aimed to detect health problems in social media and to predict whether they have a disease based on the writings of the authors. The experiments used on two datasets collected from public social media platforms. They concluded that the MHANT model outperformed the Bert approach by 0.73% and 0.4% in the dementia dataset and depression dataset [48].

Huang et al. (2015) proposed LDA and probabilistic risk stratification model (PRSM) models for a new approach to risk stratification by exploring large volumes of electronic health records in an unsupervised manner. The PRSM model recognizes the patient's clinical status as a probabilistic combination of latent sub-profiles. From their EHR, they created layers of risk for their patients. With the results obtained, they created a system that can be easily recognized as high, medium, and low risk, respectively. They validated the efficacy of the proposed approach in a clinical dataset containing 3463 coronary heart disease patient samples [49].

Liu et al. (2021), They analyzed the news about hospice care. The study consists of two time periods. They analyzed the remaining 2227 news stories after clearing data irrelevant to data cleaning processing. While the issue of care and health services was at the forefront for the first data set, patient treatment and health care development issues came to the fore in the second data set [50].

People who have moved away from traditional communication are now expressing their opinions on the internet and using social media tools to create a different communication model. The health sector has started to make improvements in their companies with the comments made by people on the internet. Therefore, studies in this field are very valuable. In conclusion, studies in the health sector are summarized.

2.2.2. Topic modeling in COVID-19

In the COVID-19 pandemic, machine learning algorithms can shed light on many problems. In machine learning, topic modeling algorithm, one of the most preferred algorithms, is used to cluster the topics that people talk about. Studies in this area are summarized in the **Table 2.6.**

Liu et al. (2021), Comments, discussions, and research about COVID-19 are increasing rapidly. Using the LDA model, they identified research topics related to Covid-19. They provided an overview of studies dealing with the COVID-19 crisis at different scales, including referencing/citing behaviour, topic diversity, and interrelationships. The results reveal the focus of scientific research, thus providing insight into how the academic community is contributing to the fight against the COVID-19 pandemic [51].

Cuaton et al. (2021) analyzed government responses regarding the management of the Covid-19 disease. Using the data from the Ministry of Health, they argued that the focus is more on nationwide analyzes of Covid-19 patients and reporting is done in this area. They have proven that the contact tracing of their infected patients is less and the government is less focused on this issue [52].

Yazdy et al. (2021), the government was in crisis when the WHO declared COVID-19 the International PHEIC on 31 January 2020. They used the topic modeling method to analyze these sentences. The results are the main concerns of half of the public; There have been awareness actions on (1) PCR testing, diagnosis, and screening, (2) shutdown of the education system, and (3) handwashing and face mask use [53].

Burel et al. (2020) investigated the relationship between the spread of misinformation and confirmation during the COVID-19. From December 2019 to January 4, 2021, they pulled the misinformation that surfaced on Twitter and the data that formed their fact-checks. They observed that confirmations about COVID-19 emerged rather quickly after misinformation spread [54].

Koh et al. (2021) investigated the loneliness experienced during the pandemic. They examined the main areas of this loneliness. They divided the effect of loneliness on society into 3 main headings. They focused on issues such as social distancing and mental health. As a result, they found that people made more comments about mental health [55].

Ordun et al (2020), They aimed to determine the subjects of their tweets during the pandemic. They created the topics discussed using the LDA model. They analyzed that news and rumors about Covid-19 spread very quickly [56].

Table 2.5: Topic modeling in COVID-19 literature summary table.

Author(s)	**Method**	**Problem**	**Result**
Liu et al. (2021)	LDA	Analyzing COVID-19 articles	To determine the research topics in the COVID-19 process.
Cuaton et al. (2021)	LDA	Analyzing the responses and statements of the Filipino government about the pandemic.	It revealed five hidden themes
Yazdy et al. (2021)	LDA	Analyzing COVID-19 articles	Found that people's concerns are about PCR, online education, and mask use.

The researchers' studies on the COVID-19 pandemic help to understand the feelings of the public during the pandemic process. The comments people make about COVID-19 on social media and news are increasing day by day. Studies show that people are the most talked about topics such as "social distance", "school closure", "wearing a mask", "curfew" and "vaccination in children". Studies in the field of health are summarized, the most researched health issue is COVID-19, and the most used method is observed as LDA.

2.3. Sentiment Analysis

Experienced people's feelings and opinions of the machine because of their experiences is called sentiment analysis. It can distinguish emotions as positive, negative, and neutral.

Thanks to the sentiment analysis, the satisfaction of the customers and their thoughts about the product can be determined. With the emotion dictionary, it can detect the general complaint about a brand or company and make improvements.

Carosia et al. (2021) to do sentiment analysis of Brazilian financial news. Using the outputs of sentiment analysis, they tested the relationship between emotions in the news and suggested an investment strategy. As a result, they found that Convolutional Neural Network is the most suitable for performing sentiment analysis in Portuguese, and they concluded that the dominant daily news sentiment in the stock market has a significant effect [57].

Daudert. (2021), made sentiment analysis of company reports. They found that there was anxiety over issues such as financial developments and domination. The proposed solution, on the other hand, has proven to improve performance by up to 15% and 234% over multiple baselines [58].

Li et al. (2021) analyzed the comments of online dating apps. They used machine learning algorithms for comments made on dating sites. They proved that machine learning methods outperform sentiment analysis [59].

Fang et al. (2015), aimed to do a sentiment analysis of online product reviews on Amazon.com. They proposed an emotion polarity categorization process with detailed explanations of each step. They made both sentence-level classification and analysis-level classification. Naive Bayesian, Random Forest, and Support Vector Machine are the selected classification models for categorization [60].

Table 2.6: Sentiment analysis literature summary table.

Author(s)	**Problem**	**Result**
Fang et al. (2015)	Sentiment analysis of online product reviews on Amazon.com	Proposed an emotion polarity categorization process with detailed explanations of each step
Carosia et al. (2021)	To do sentiment analysis of Brazilian financial news.	Concluded that the dominant daily news sentiment in the stock market has a significant impact.

Daudert. (2021)	Conducting detailed sentiment analysis of company and analyst reports	They have proven to increase performance by 15% and up to 234% over multiple baselines with the proposed solution.
Li et al. (2021)	To make sentiment analysis of Online Dating Services comments.	Proved that machine learning methods outperform sentiment analysis

When sentiment analysis studies are examined, there are many studies such as the classification of people's comments, the analysis of company reports, the reactions of the audience to the news.

2.3.1. Sentiment analysis in health

Advantage of social media increases the quality of health services, companies that have started to consider the comments on social media have started to be successful. Companies are trying to make new plans by listening to the voices on social media. Twitter, which is the platform where people express their feelings most comfortably, maintains its popularity in the sentiment analysis method. Studies in the field of emotion analysis health are summarized.

Palomino et al. (2015) analyzed people's feelings about nature. In this study about people's alienation from nature disorder, they analyzed people from Twitter using data on this issue. They offered suggestions for the good dissemination of people's views on public health [61].

Albornoz et al. (2018) They aimed to analyze patient opinions. They investigated health-related service quality by categorizing positive and negative comments. Opinions of the patients were collected, and they made a detailed analysis together with the emotion-based words. As a result, by performing sentiment analysis, it is possible to predict with very high accuracy (about 70 percent) the polarity of sentences written by the patient [62].

Zunic et al. (2020), Raw data were first collected from Drugs.com and they performed sentiment analysis to assess the health domain. They applied the emotion analysis method of positive and negative comments using drug reviews. results proved that graphical

convolution approach outperforms standard deep learning methods in dimension-based sentiment analysis task [63].

As a result of emotional studies in the field of health, it is very important to analyze issues such as comments about the pharmaceutical company, comments of patients about the hospital, comments about the health system. By separating these comments as good or bad, the performance of the system can be measured. Performance improvement can be made with the result of comments.

2.3.2. Sentiment analysis in COVID-19

In machine learning, the sentiment analysis algorithm, which is one of the most preferred algorithms, is used for clustering the topics that people talk about. In this study, sentiment analysis studies on COVID-19 are summarized in **Table 2.9.**

Fang et al. (2021) They conducted sentiment analysis of COVID-19 users using data from Weibo. They found that the pandemic significantly affected individual emotions, causing more passive emotions (for example, fear and sadness) [64].

Hussain et al. (2020) They wanted to analyze the public's thoughts and feelings about COVID-19 vaccines via social media. It has been observed that the public has concerns about the trial process of the vaccine, as well as fears about the economic situation. They obtained a result that they were confident about the efficacy of the vaccine [65].

Obembe et al. (2021) They analyzed the crisis experienced by COVID-19 tourists in communication using the model. They used tweets and news containing comments about tourists. This article investigated the main factors affecting public sentiment during the onset of the crisis. They identified the strategies that stakeholders should create for the crisis [66].

In the study of Basiri et al. (2021), tweets collected from eight countries about COVID-19 were analyzed. All news and events made to measure sensitivity in tweets are correlated. They concluded that the information about COVID-19 has increased, countries have unique emotion patterns, and negative emotional values are in active cases [67].

Yousefinaghani et al. (2021) They analyzed people's feelings and thoughts about COVID-19 vaccines. They observed that discussions on vaccine rejection and hesitation during the study were more than interest in vaccines. It has been determined that there are Twitter

bots or political activists, and that famous people and organizations write articles in favor of vaccines [68].

Rahman et al. (2021) to analyze people's feelings and thoughts about the economic crisis caused by COVID-19. They collected Using Twitter data, the analysis of people with low education level, people with low income, people with high rent levels was interested in economy [69].

Garcia et al. (2021) The effectiveness of sentiment analysis and topic models were compared using data from Portuguese and English languages. They observed that the emotions were negative in the outputs obtained [70].

Imran et al. (2020) They analyzed people's feelings about the measures taken for COVID-19. Dataset consists of tweets taken from the public Kaggle site. As a result, they found the relationship between the emotions of people in neighboring countries during the COVID-19 pandemic [71].

Table 2.7: Sentiment analysis studies during the COVID-19.

Author(s)	**Problem**	**Results**
Imran et al. (2020)	Analyzing positive and negative comments about COVID-19	The relationship between the emotions of people in neighboring countries
Hussain et al. (2020)	Analyzing people's emotions about the COVID-19 vaccines social media in the UK and United States	Public concerns about vaccine
Fang et al (2021)	Analyzing people's emotions during the COVID-19 process	People concluded that they felt fear and sadness.
Yousefinaghani et al. (2021)	To determine the public's feelings and thoughts on COVID-19 vaccines	There are Twitter bots or political activists, and famous people and organizations write articles in favor of vaccines.

Rahman et al. (2021)	To analyze people's feelings and thoughts about the economic crisis caused by COVID-19.	Analysis of people with low education level, people with low income, people with high rent levels was interested in economy

In conclusion COVID-19 studies, there are concerns about vaccines, fears about education, and psychological problems related to curfews. Twitter has been used as data in most of the studies in this field. As a result of the studies, people's feelings in the COVID-19 process are included.

2.4. Topic Modelling and Sentiment Analysis

As technology progresses, the use of social media has increased, and people have become comfortable reporting their opinions and thoughts. When topic modeling and sentiment analysis are used in the same study, it yields efficient results. Thanks to the sentiment analysis scores made on the subjects, the quality can be increased by studies such as improving the subjects.

Nguyen et al. (2015) created a model to predict stock price action using emotions in social media. They used a new feature in the prediction model that simultaneously captures the subjects and their emotions. In addition, a new topic model Topic Sentiment Latent Dirichlet Allocation (TSLDA) is proposed to achieve this feature. Compared with other sentiment analysis methods, the accuracy of the TSLDA method was also found to be 6.43% and 6.07% better than LDA and JST-based methods [72].

Chakraborty et al. (2020) They analyzed different data sets on two separate dates. The most retweeted tweets about COVID-19 (1 January 2019 - 23 March 2020) were analyzed. They observed that negative feelings were intense. The most retweeted tweets about COVID-19 (December 2019 - May 2020) were analyzed. They observed that positive feelings were intense [73].

Gregoriades et al. (2021) They aimed to increase the positive and negative issues with the tourist experience, service performance, and satisfaction The proposed method also considers the tourists' collective cultural and economic knowledge of their country through topic modeling and DT models. Based on the results, it is likely that high-star hotels will have complaints about the cost, while low-star hotels are likely to have complaints and criticisms about food and cleanliness [74].

Melton et al. (2021), explored the discussion about the COVID-19 vaccine on social media, they conducted a sentiment analysis and LDA topic modeling analysis on textual data collected from 13 Reddit communities focused on the COVID-19 vaccine from December 1, 2020, to December 1, 2020. They distinguished the concerns, fears, and all the emotions they expressed about the vaccine as positive and negative. As a result of topic modeling, it was revealed that society was worried about the side effects of the vaccine rather than the fake news. As a result of LDA topic modeling, keywords showing vaccine hesitancy during the COVID-19 vaccine period were determined. It has emerged that people are concerned about vaccine side effects [75].

Wright et al. (2021) investigated public opinion about the central UK government during COVID-19, they identified the hottest topics by extracting themes from more than 4,000 free-text survey responses collected between 14 October and 26 November 2020. As a result, government corruption, nepotism, rules created in the process, inconsistency in messages, lack of clear planning, lack of openness, and transparency were the ten most discussed topics. As a result of the sentiment analysis of the subjects, their perspectives on the subjects were negative [76].

Satu et al. (2021), design an intelligent clustering-based classification and topic extraction model called TClustVID that analyzes public tweets about COVID-19. They collected COVID-19 Twitter datasets from the EEE Dataport repository and used several data preprocessing methods to clean up the raw data. They discovered that TClustVID outperformed traditional methodologies determined by clustering criteria, removing important topics from clusters, categorizing them according to positive, neutral, and negative emotions, and using the proposed model, they identified the most frequent issues [77].

In the study of Kaila et al. (2020), tweets with news content about #coronavirus spreading on Twitter were analyzed using sentiment analysis and LDA topic modeling. Coronavirus epidemic was minimal misinformation and accurate and reliable information. The sentiment analysis result confirmed the prevalence of positive emotions such as trust as well as negative emotions such as fear [78].

İt was observed that productive results were obtained by using topic modeling and sentiment analysis together. After the subject clouds are determined, sentiment analysis of each subject can improve the determined subject and increase the quality. When the

studies in the health sector in topic modeling and sentiment analysis are examined, it is shown that most studies are made about COVID-19. People are worried about vaccination and education in children during the COVID-19 process. It can be said that one of the most used applications in these studies is Twitter.

CHAPTER 3

3. EXPERIMENTAL PART

With the measures taken during the COVID-19 process, great restrictions have come in social life. Elderly people, people with disabilities, families, students, and employees have been psychologically affected in this process. The main purpose of the study is to examine the perspectives of the attitudes and beliefs of students, families, and teachers with the transfer of education to the online platform and to understand the reasons behind these beliefs. By using the R software program, it is aimed to analyze the tweets expressing their opinions about the online education on Twitter. **Figure 3.1** contains a summary image of data collection and analysis.

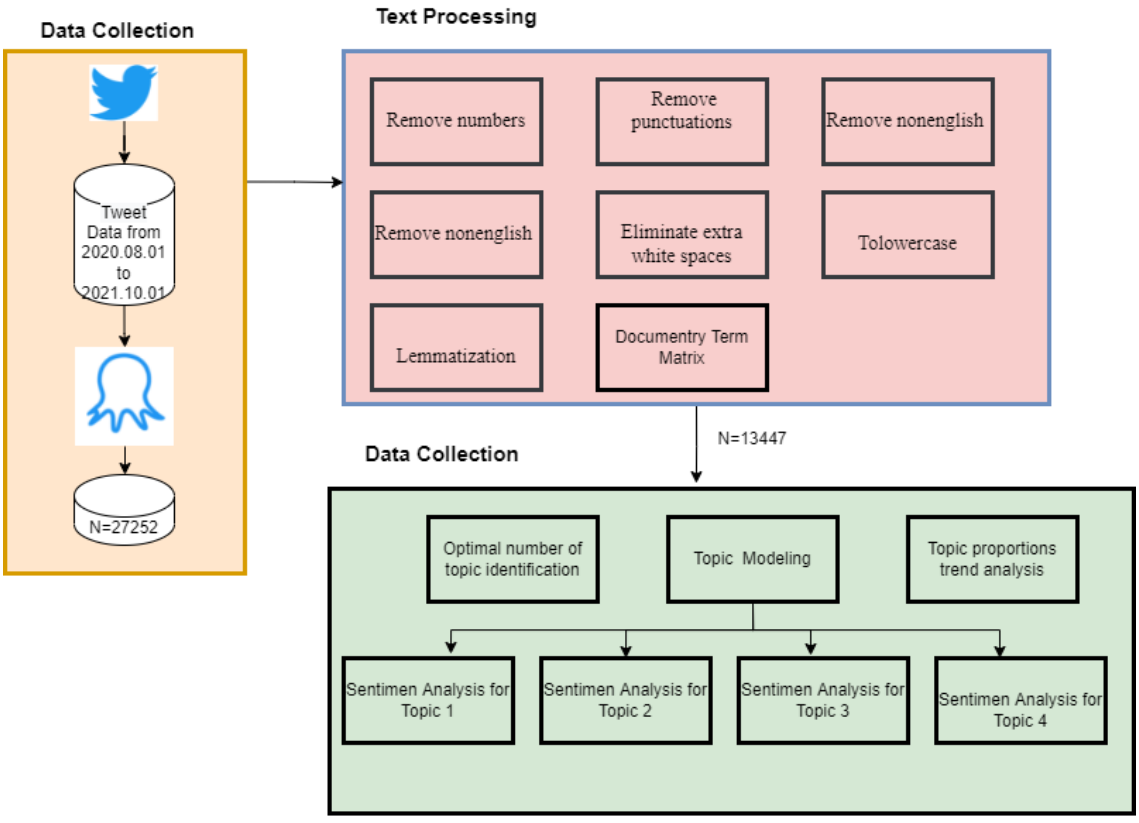

Figure 3.1: Summary of data collection process and analysis.

3.1. Octoparse

Octoparse is an easy-to-use tool that helps users extract content from the website. It works on dynamic web pages, including scraping pages with pagination. Moreover, the cloud service can receive and store large amounts of data. With the Octoparse application, data on the website can be collected by creating a web crawler without the need for coding. To access Twitter data, data can be accessed without a Twitter developer. It is possible to extract data using the URL of the hash or the pages of the people you want to reach on Twitter. The data to be accessed via the URL is selected, for example, data such as tweets on Twitter, the number of retweets received, username, the date can be selected and exported to CSV, HTML, or SQL.

Step 1: Enter the URL and create a pagination

In the first stage, a cloud of words requested from Twitter was created to create an octopus URL. A code was created in the advanced search section of Twitter. The creation of Twitter between August 1, 2020, and October 1, 2021, was requested and the start dates of the education period were considered. In addition, the processes involved in the COVID-19 process were discussed. It was aimed to receive tweets that commented on education during the COVID-19 process related to children. The code shown in **Table 3.1** was used for the Advanced Search part. Twitter URL was created using this code. A part of the generated URL is shown in Figure 3.1. In addition, the code used in the Advanced search is shown.

Table 3.1: Twitter advanced search code.

("remote learning" OR "distance learning" OR “virtual learning” OR “distance education” OR “freshman class” OR “online learning” OR “eLearning” OR “online school” OR “education” OR “school” OR “student” OR “freshmen class” OR “lecture”) (child OR boy OR child OR freshman class OR parents OR school OR children OR kid OR kids OR daughter OR boys OR son) (covid OR corona OR covid19 OR covid-19 OR pandemic OR coronavirus) lang: en until:2021-10-01 since:2020-08-01 -filter: links -filter: replies

Pagination is used to allow it to scroll down the page repeatedly. Pagination was obtained by clicking on the blank part of the page and clicking on the Allow rendering button. The purpose of this process is to create a page loop and scroll through the pages successfully.

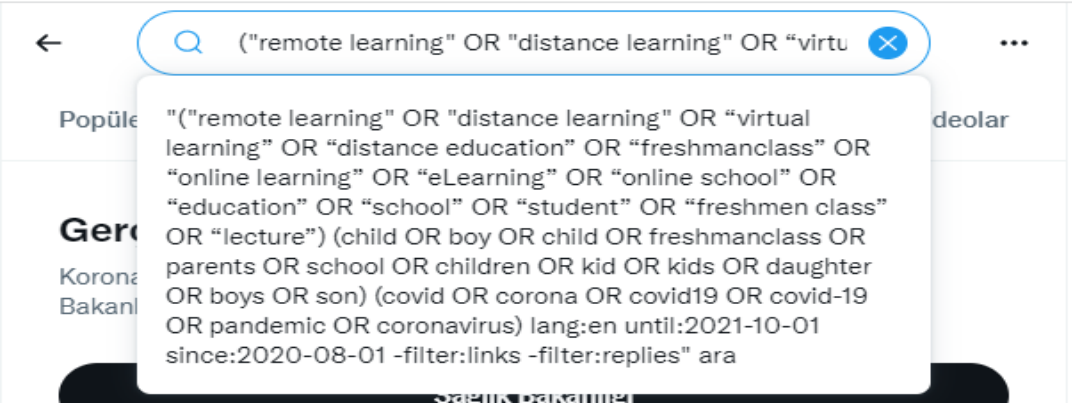

Figure 3.2: Prepare Twitter URL.

Step 2 Build a Loop Item

The data to be retrieved from Twitter is created. For example, the desired information such as a tweet, tweet likes, the username of the person who tweeted, date of tweeting is selected. A loop item is to request the desired titles by creating this loop on each page. In this study, the tweet, the username, the number of likes received for the tweet were taken. In **Figure 3.2**, there is a workflow for pagination and loop item. A loop was created for each page, and the tweet, username, and likes for the tweet were returned in the loop item. With pagination, this is repeated for each page.

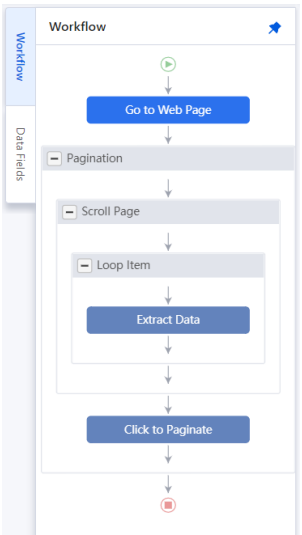

Figure 3.3: Workflow created for Twitter.

3.2. Transfer the Data to R

Text for NLP can be loaded into R from different source formats. It was requested to read the CSV file through the R program. For this reason, the read.csv command is used. As CSV, the name of the file is named education. It is shown in **Table 3.2.**

Table 3.2: Read.CSV to Import Data in R

```
education <- read.csv(file.choose(), header = T)
```

3.3. Pre-processing

Text preprocessing has steps. Text preprocessing generally refers to cleaning data or making data available for analysis. The text to be Data Mined can be loaded into R from different source formats. Extract from text files (.txt), pdf (.pdf), CSV files (.csv), etc. It may sound, but whatever the source format, the tm package to be used will be converted into a "corpus". A corpus is defined as “a collection of written texts, especially all the works of a particular author or writings on a particular subject”. The tm package uses the Corpus () function to create a collection. The conversion is performed, for example, using the tm_map() function to replace special characters in the text.

1- Corpus: A corpus is a large, structured set of machine-readable texts produced in a natural communication environment. A corpus is a collection of text or speech that has been put together according to a certain set of predetermined criteria. It is shown in **Table 3.3.**

Table 3.3: Corpus in R.

```
mydata_corpus <- Corpus (VectorSource(education$Tweet))
```

2- removePunctuation: It deletes unnecessary spaces and punctuation marks from the created corpus. It is shown in **Table 3.4.**

Table 3.4: Remove punctuation in R.

```
mydata <- tm_map(mydata , removePunctuation)
```

3- stripWhitespace(): It deletes the extra spaces in the sentences in the corpus. It is shown in **Table 3.5.**

Table 3.5: Strip whitespace in R.

```
mydata <- tm_map(mydata , stripWhitespace)
```

4- To lower(): It examines the upper and lower case letters of the sentences in the corpus and organizes the whole sentence in lower case. It shows in **Table 3.6.**

Table 3.6: Tolower in R.

```
mydata <- tm_map(mydata , content_transformer(tolower))
```

5- Stopwords(): They are English words that do not add much meaning to a sentence. Deletes the sentence without compromising the meaning. For example, it catches and ignores words like the, he, have. It shows in **Table 3.7.**

Table 3.7: Stopwords in R.

```
mydata <- tm_map(mydata , removeWords, stopwords("english"))
```

6- Lemmatisation: It provides the root of the word. It shows in **Table 3.8.**

Table 3.8: Lemmatize in R.

```
mydata<-tm_map(mydata, textstem::lemmatize_strings)
```

7- RemoveWords(): Words that are not intended to appear as topics in topic modeling have been cleaned. It shows in **Table 3.9.**

Table 3.9: Removewords in R.

```
mydata <- tm_map(mydata, removeWords,c("remote learning","distance learning", "virtual learning", "distance education", "freshmanclass","online learning","eLearning","online
school","education","school","student","freshmenclass","lecture","child","boy","freshmanclass","patents","shool","children","kid","kids","daughter","boys","son","covid","covid19","covid-
19","pandemic","coronavirus","corona""get","say","just","know","back","now","take","still","one","today","will","like","now","even","may","can","year","week","good","need","people","want","due","make","day","send","start","see","high","think","positive"
```

8- RemoveNumbers() in R: It shows in **Table 3.10.**

Table 3.10: RemoveNumbers in R.

```
mydata <- tm_map(mydata, removeNumbers)
```

Corpus was used to clean the data. It is aimed to remove punctuation marks with removePunctuation. It is aimed to clean the excess spaces with stripWhitespace. With Tower, it was aimed to complete the upper- and lower-case letters with the same expression. All words are aimed to be lowercase formants. The stopword, it is aimed to remove the ineffective words. In RemoveNumbers, unnecessary numbers in the tweet have been deleted. And in Removewords, the deleted words are the words searched on Twitter. These words are not requested to be included as a subject in topic modeling. It is expected that the subjects included in the education subject will come. Before and after the data prepared for the application are summarized in **Table 3.11.**

Table 3.11: Before and after Data Output.

Before	After	Function
The fact that education is online and can cause students to become angry!!....	The fact that education is online and can cause students to become angry	RemovePunctuation()
The fact that education is online and can cause students to become angry	The fact that education is online and can cause students to become angry	stripWhitespace
The fact that education is online and can cause students to become angry	the fact that education is online and can cause students to become angry	To lower
The fact that education is online and can cause students to become angry	fact that education online can cause students become angry	Remove Stop Words

The fact that education is online and can cause students to become angry due to covid-19	the fact that education is online and can cause students to become angry covid	RemoveNumbers

3.4. Topic Modelling

Topic modeling is a type of statistical model in machine learning and NLP that explores abstract topics within collected documents. Topic modeling, in a sense, is text mining that investigates the semantic links hidden within the texts. The results of topic modeling algorithms can be used to visualize, explore, summarize, and theorize the topic focused on. Apart from detecting trainer structures in data such as topic models, genetic information, images, and networks, it is also used for bioinformatics, NLP, and chatbot studies

First, LSA can be used for size reduction when the usage areas of LSA are examined. It creates a matrix of terms for document clustering. It is a linear algebra method. Therefore, it does not perform well on non-linear datasets. The vector size can be greatly reduced without losing any context or information. It does not work well with small documents. Another disadvantage is that it cannot catch many meanings (synonyms) of the word. Its application is easy and fast. It has consistent results. As a result, it reduces the computational power and the time taken to perform the computation. Second, pLSA models word-document formations. Its results have clear probabilistic interpretation but cannot provide a document-level probability. It is a complex model. It is not a defined manufacturer model for new documents. It can filter on requests to get information. Third, LDA is advantageous over LSA and PLSA due to its size reduction feature. It finds topics in the document collection and automatically classifies them based on how "relevant" they are in the document. LDA is the Bayesian version of pLSA. When the disadvantages of LDA are examined, it cannot capture the correlation between subjects, it is static, so there is no change in subjects over time, and finally, it cannot model sentence structure. Fourth, the HDP hybrid model can be used in a non-parametric natural generalization where the number of subjects is unlimited and can be learned from the data. Therefore, it cannot work well on non-parametric data. Fifth, CTM is used in the fields of Text mining, multimedia, image retrieval, biomedical, role discovery, emotion subject, anti-phishing. Its advantages are performing well on complex issues. It is dynamic. Topics change over

time. Its disadvantages are that it does not fit in its multinomial analysis. Many numbers also have computation. There are too many general words on the subject. Sixth, in the Dynamic Topic Model, each topic defines a polynomial distribution over a set of terms. They are generative models that can be used to analyze the evolution of (unobserved) topics of the document collection over time. It can capture both power and content development at the same time. However, it is not used in complex temporal patterns. Finally, the Pachinko Allocation Model analyzes the text by looking at the correlation between topics. Improvises by modeling the correlation between subjects. Therefore, the efficiency of the results decreases.

Table 3.12: Topic modeling method comparison.

	Advantage	Disadvantage	Usage areas
LSA	❖ It has much better performance than the flat vector remote model. ❖ Efficient and easy to apply. ❖ It is faster than other topic modeling algorithms. ❖ It has consistent results.	❖ Does not perform well on non-linear datasets. ❖ Does not perform well on small datasets. ❖ Can't find the synonym of the word. ❖ Hard to interpret outputs	It is a linear algebra method. LSA can be used for size reduction, document clustering. It creates a matrix of terms.
pLSA	❖ Its results have a clear probabilistic interpretation. ❖ Information filtering can also be used.	❖ It cannot provide document-level possibilities. ❖ It is a complex model. It is not cluster oriented. ❖ It is not a well-defined generating pattern for new documents.	NLP can be used in machine learning from text. Models word-document formations.
LDA	❖ B t is advantageous compared to LSA and PLSA due to its size reduction feature. ❖ It can be incorporated into more complex methods.	❖ Slowly trained. ❖ The number of topics is fixed and must be known beforehand. ❖ Subjects do not change over time, they are static.	It finds topics in the document collection and automatically classifies them based on how "relevant" they are in the document. LDA is the Bayesian version of pLSA.

HDP	❖ Parametric sets give good results.	❖ Does not work well on non-parametric sets.	The HDP complicated model is used in a non-parametric data set where the number of subjects is unlimited.
CTM	❖ Performs well on complex issues. ❖ Subject ideal number subject n>=60 ❖ Topics change over time. It is dynamic.	❖ It is not polynomial. ❖ Has many calculations. ❖ There are not many general words on the subject.	Text mining, multimedia, image retrieval, biomedical, role exploration, sentiment.
Dynamic Topic Model	❖ Capable of capturing both power and content development at the same time. ❖ It is used in documents suitable for dynamic topic models. ❖ Can classify and group according to time zone.	❖ It is not used in complex temporal patterns.	For each word of each document, a topic is drawn from the mix, and then a term is extracted from the corresponding polynomial distribution for that topic.
Pachinko Allovation	❖ Subject ideal number subject n>=60 ❖ More flexible and more powerful.	❖ The startup process is random. ❖ Improvises by modeling the correlation between subjects. Therefore, the efficiency of the results decreases.	It analyzes the text by looking at the correlation between the subjects.

3.4.1. Latent dirichlet allocation

Topic Modeling is a machine learning method that determines the semantic structure of a document containing text. NLP is also cited as a research area. Topic modeling methods can organize and summarize high-content text documents. It can be successfully applied in many areas such as topic modeling, automatic document indexing, document classification, topic discovery. LDA is a probability-based topic modeling method. At the core of LDA, topics have a probability distribution over words, and text documents have a probability distribution over topics. Each subject has a distribution over the word string. LDA is an unsupervised learning algorithm, it does not need predefined words. After the

number of topics is determined, labels are assigned to the topics according to the classes. Working principle; For each document, it assigns a random topic to the words in the document. Using this information, the model generates various statistics. After the statistical information is obtained, the subject assignment of each word is performed for each document. For this, the existing vocabulary information should be updated as much as the number of iterations.

3.4.1.1. How does latent dirichlet allocation work?

LDA is a method that divides sentences into topics by matching them to topics. They are probability distributions on specific topics. The purpose of LDA is to find documents with similar topics and use them by sorting them according to similar word groups. It finds the pass rate in the documents according to the word frequency and divides them into topics. Lda flow representation is shown in **Figure 3.3.**

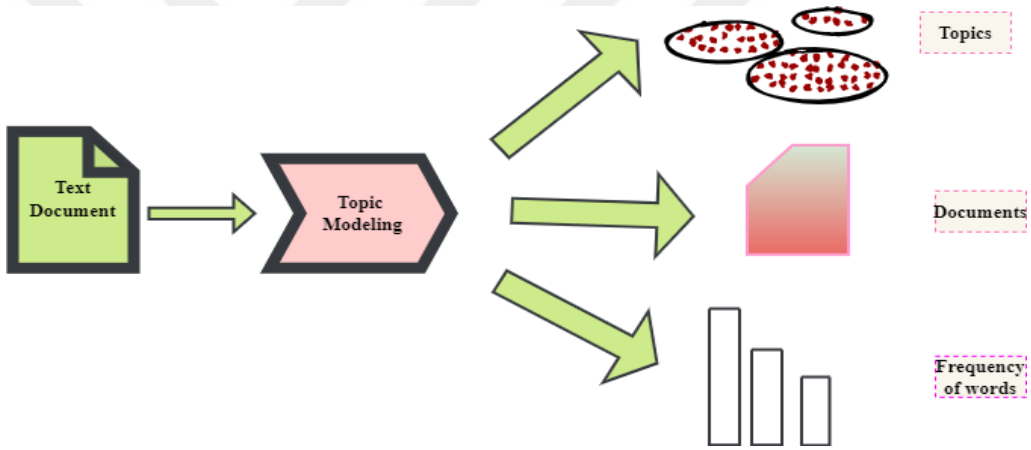

Figure 3.4: LDA algorithm.

For example, let's examine a few tweets.

Tweet 1: I'm sorry I didn't go to school; homeschooling is so bad.

Tweet 2: I miss my teachers so much.

Tweet 3: All my family's COVID-19 test result is positive.

Tweet 4: My COVID-19 test was positive, I have a severe illness.

Tweet 5: I won't be able to see my teacher because my test result is positive.

Tweet 1 and Tweet 2 both belong to topic 1

Tweet 3 and Tweet 4 both belong to topic 2

Sentence 5 describes 70% topic 1 and 30% topic 2

LDA assumes certain rules and regulations before producing a document. Just as there should be a word limit, a document must have a certain number of user-specified words. There should also be diversity in the content of the document. The distribution of tweets into words is shown in **Figure 3.4.**

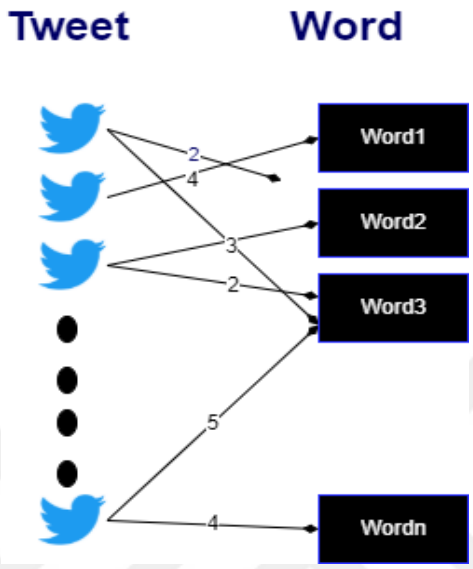

Figure 3.5: Distribution of tweets to the to word.

It refers to documents within the dataset (tweets). A fixed number of K topics are to be discovered.

The LDA algorithm loops through each document and randomly assigns each word in the document to one of the K topics. This random assignment already gives both the topic representation of all documents and the word distribution of all documents and the word distribution of all topics.

The LDA model has two parameters that control the distributions:

1. Alpha (α) means the distribution of the topic falling into the documents and.Beta (β) means is provides topic word distribution control.

To summarize: Its shows **Figure 3.5.**

- M: means total documents in the corpus
- N: refer the words (number of) in the document
- w: means Word
- z: is the latent topic assigned to a word

- theta (θ) means the topic distribution
- LDA model s parameters: Alpha (α) and Beta (β)

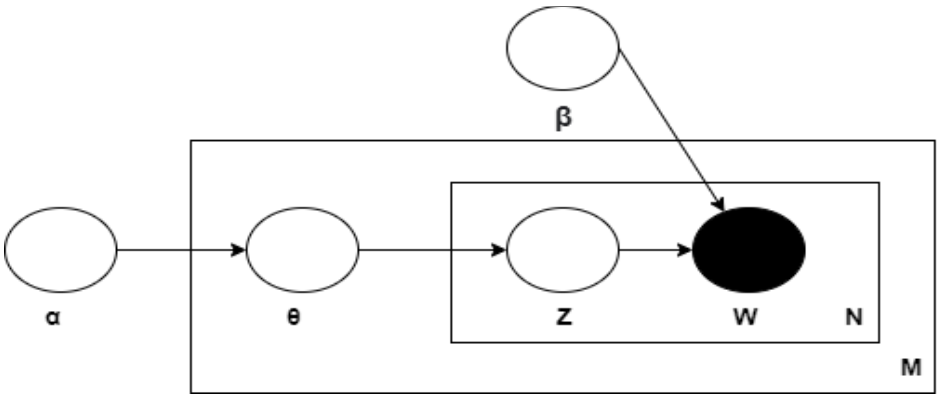

Figure 3.6: LDA parameters.

LDA does two tasks: it finds the topics from the corpus, and at the same time, assigns these topics to the document present within the same corpus. Its shows in **Figure 3.6** schematic diagram summarizes the process of LDA.

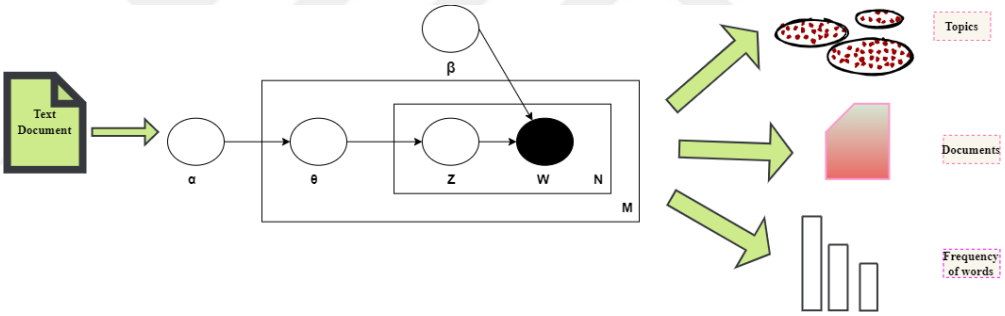

Figure 3.7: LDA flow.

LDA is a general probabilistic hierarchical Bayesian model that initiates topics from a collection of documents in three ways.

1. Each document in the collection is distributed among the subjects sampled for this document according to the Dirichlet distribution.

2. Each word in the document is associated with a unique topic based on this Dirichlet distribution.

3. Each topic is represented by a polynomial distribution over the words specified for the sampled topic [79].

The formula includes the probability that a document will discuss the subject and the probability that the words will be seen. The distribution of documents by topics, the

distribution of topics by terms, the probabilistic distribution of words for each topic among topics are calculated with the LDA formula (1).

Based on this informations, the following formula was appearing in the use of the LDA method;

$$P(W, Z, \Theta, \varphi; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i, \beta) \prod_{j=1}^{M} P(\Theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\Theta_j) P(W_{j,t}|\varphi_{Z_{j,t}}) \quad \textbf{(1)}$$

3.4.2. LDA application

In this application, the LDA algorithm has been chosen to allocate subjects into clusters. LDA technique is used with R program support. The process was started by reducing the vector size of the data set prepared within the scope of this study. As seen in **Figure 3.7**, it is aimed to find the frequency of the words. The columns are summed to narrow the matrix. The frequencies of the terms were recorded. All rows with a value of zero are deleted. In this way, the model is ready to create subject sets.

```
"Collapsing matrix by summing over columns"
dtm <- removeSparseTerms(dtm, sparse = 0.95)
frequency <- colSums(as.matrix(dtm))
"Length should be total number of terms"
length(frequency)
"Creating sort order (descending)"
ord <- order(frequency, decreasing = TRUE)
"Listing all terms in decreasing order of freq and write to disk"
frequency[ord]
write.csv(frequency[ord], "word_freq.csv")
"Removing the 0 rows"
raw.sum=apply(dtm,1,FUN=sum)
dtm=dtm[raw.sum!=0,]
```

Figure 3.8: Finding the frequencies of words.

It is expected that 4 subject clusters will be created and 5 words expressing these subject clusters will be formed. **Figure 3.8** shows how the number of subjects in topic modeling is determined. In the next step, a CSV file was created to see the tweets of each subject in the sentiment analysis. In this file, there are tweets belonging to the topics.

```
"Creating model with 4 topics"
k=4
seed=1234
lda_fit <- LDA(dtm, k=k, control=list(seed=seed))
lda_fit@alpha
topics(lda_fit, k)
terms(lda_fit, 4)

lda_fit.topics <- as.matrix(topics(lda_fit))
write.csv(lda_fit.topics,file=paste("docstotopics", k, "DocsToTopics.csv"))
tdm <- TermDocumentMatrix(tweet_corpus)
tdm <- removeSparseTerms(tdm, sparse = 0.95)
m <- as.matrix(tdm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

Figure 3.9: Creating topic number.

In **Figure 3.9**, the visualization of the topic modeling has been made. After this process was completed, the dataset of each subject was analyzed by the sentiment analysis method.

```
top_terms <-
  topics %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)



top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 18)) +
  labs(title = "LDA ", caption= "Top Terms") +
  ylab("") +
  xlab("") +
  coord_flip()
```

Figure 3.10: Visualization and analysis stage.

3.5. Sentiment Analysis

Sentiment analysis means analyzing and finding the emotion or intention behind a text or a document or any piece of writing. Sentiment analysis analyzes the ideas and feelings of texts, speeches, authors. This analysis, which adopts the text analysis technique, assigns a sentiment for each document. Sentiment Analysis or Idea Mining usage areas;

Marketing and customer service teams use social media platforms quite frequently to help their customers identify their feelings/ideas.

3.6. Sentiment Analysis Application

It can classify emotions as positive, negative, neutral. Sentiment analysis used in this study is a NLP technique used to measure an idea or emotion expressed in a series of tweets. The emotions used are shown in **Figure 3.10.**

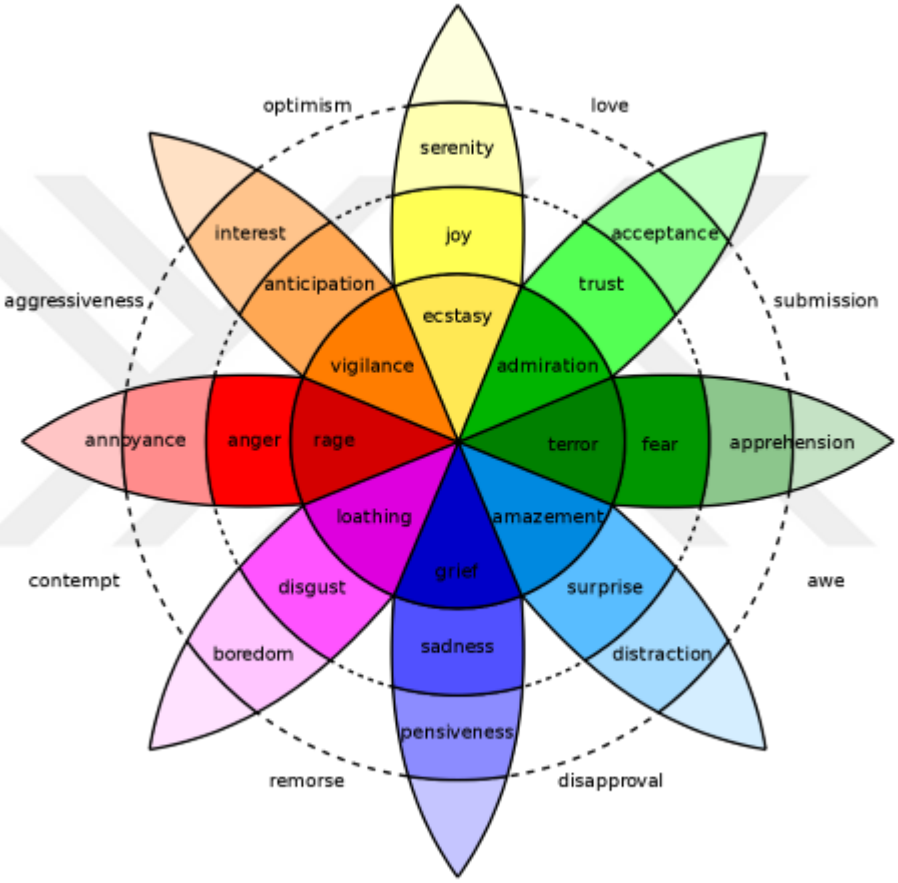

Figure 3.11: Emotions in R.

CHAPTER 4

4. RESULTS AND DISCUSSION

During the COVID-19 process, with education becoming remotely accessible via the internet, families, students, and teachers have entered a period of emotional turmoil. Among the advantages, the fact that the education is virtual causes the students not to hesitate when asking questions, in this case, the students like online education, but this may cause the student's fears to increase in the society. Although flexible time is an advantage for students who can manage time, students with poor time management suffer serious losses in this process. In the university, it is preferred that some of the short-term courses without practice are online. Decisions have been taken, such as holding practical courses at school. Considering the content of the course, it may be a reason for the course to be held at school, even if there is no application. It can be divided into courses that should be taken at school and courses that should not be, in addition to this, freshman students who have just started their education must be at school, and this is a process where they can gain self-confidence. In addition, with the emergence of face-to-face education in the COVID-19 process, it has been observed that families have positive thoughts about the importance that teachers take. Families believe that teachers will take the necessary importance in the classroom if education is face-to-face. This study can be divided into university students, high school students, and primary school students in detail. This study is separated by using the sentiment analysis library. A library related to the subject can be created in the R program within further studies.

4.1. Project Flow

Taken from tweets about education between August 1, 2020, and September 1, 2021. The transfer of education to online platforms during the COVID-19 period, it is aimed to make sense of the emotional complexity experienced by people. LDA, a topic modeling technique, was used. Word clouds were created in the Twitter Search section. Among

these created words are words such as distance education, COVID-19 , first-class, family, pandemic. These words were drawn with the Octoparse program. A total of 14 months of data were received as 27252 tweets. These data were taken between 1 August 2020 and 1 September 2021. The reason for choosing these dates is the dates when education and training started. This study, which includes a total of 14 months of data, it is aimed to analyze how students are affected during the COVID-19 process. These tweets were cleaned using corpus in the R program. After cleaning the data, 13 337 data remained. The cleaned data was used in LDA, which is the most used method in topic modeling. Sentiment analysis was performed for each subject cluster in four subject clusters and the results were interpreted. The flow chart of the project is shown in **Figure 4.1**.

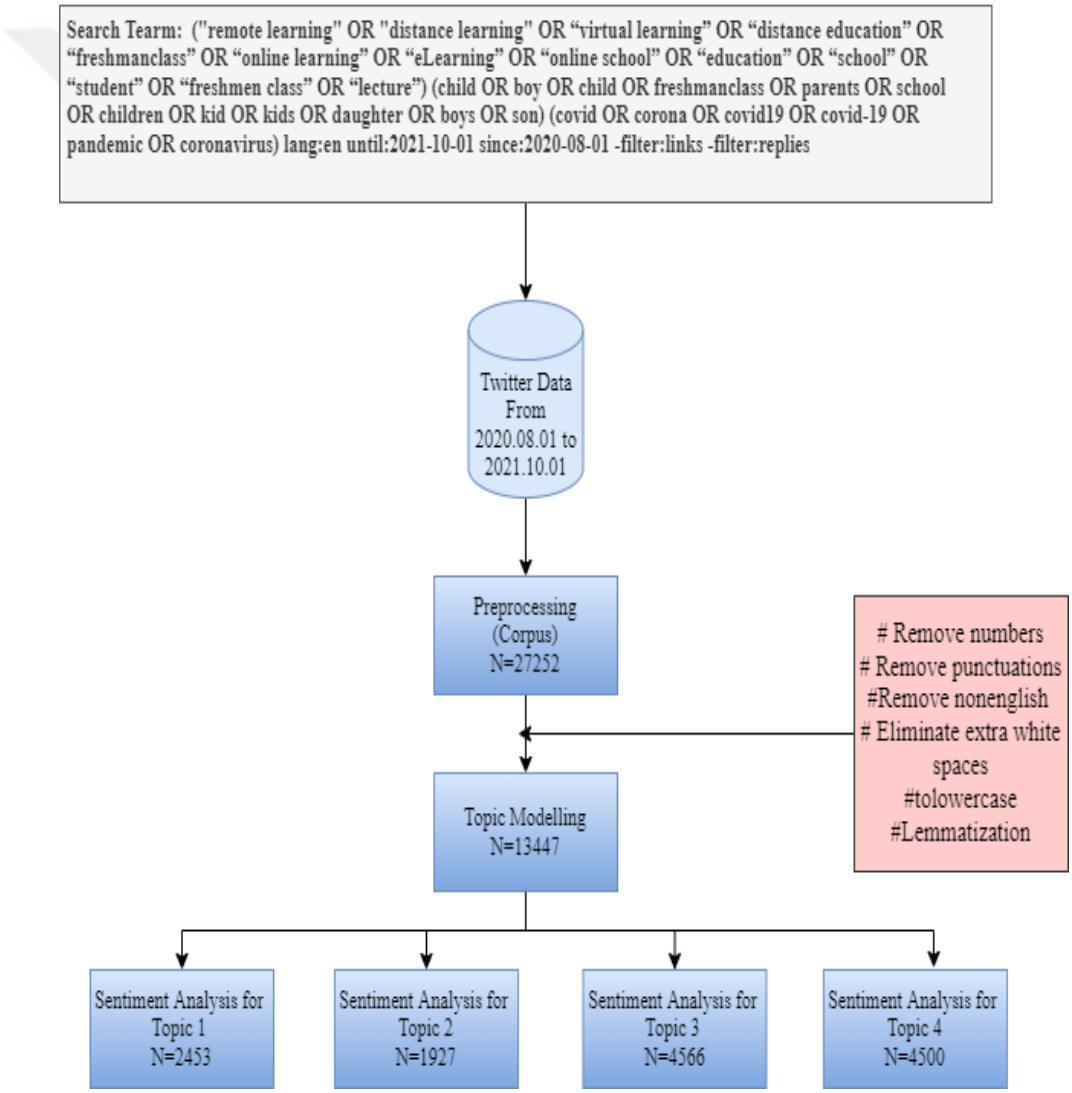

Figure 4.1: Project flow diagram.

4.2. Output of Topic Modelling

LDA output is divided into 4 clusters and labeled.

1. Topic: It is a set of topics that includes general feelings about working from home during the COVID-19 process.

2. Topic: It is a set of topics that includes the general feelings of families about the importance that teachers take with the emergence of face-to-face education in the COVID-19 process.

3. Topic: It is a set of topics that includes the general feelings of families and students about COVID-19 tests.

4. Topic: It is a set of topics that include their general feelings about time management because of working from home.

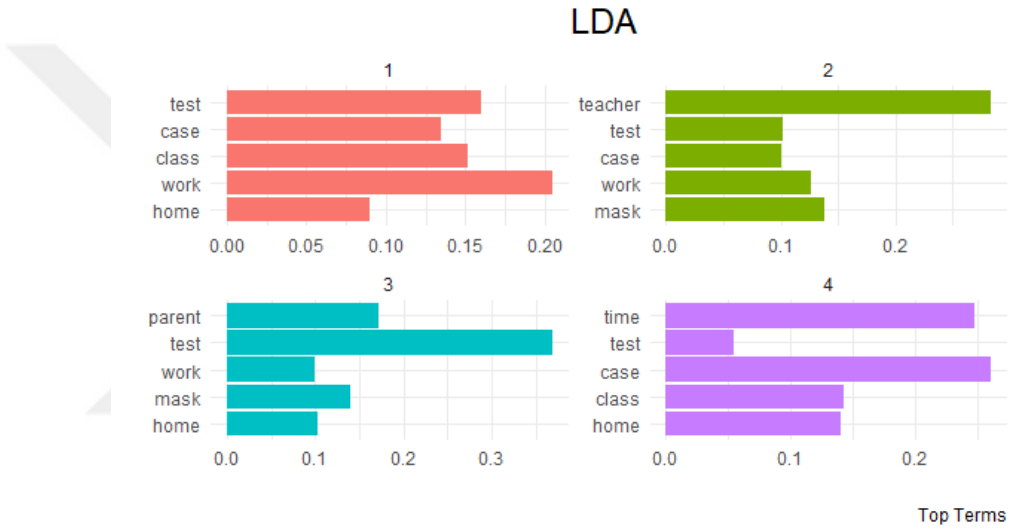

Figure 4.2: Output of LDA.

4.3. Output of Sentiment Analysis

1. Topic: It is a set of topics that includes general feelings about working from home during the COVID-19 process. When the sentiment analysis output is examined, the general thoughts of people about education being online and working from home were both positive and negative in half. This may be because people have quick access to education. In addition, being able to reach education without wasting time and effort on the way to school, without getting tired. In this process, many universities that do not have practical courses have supported online education. Many companies have even calculated that it can be profitable for the company to work from home before people come to work. However, when the graph is examined, the graphs of people such as fear, anxiety, and sadness, along with negative thoughts, were found to be excessive. In this

process, students may be confused with problems such as having internet problems, being a social worker, and expensive technology tools. These problems may have caused fear and anxiety in them.

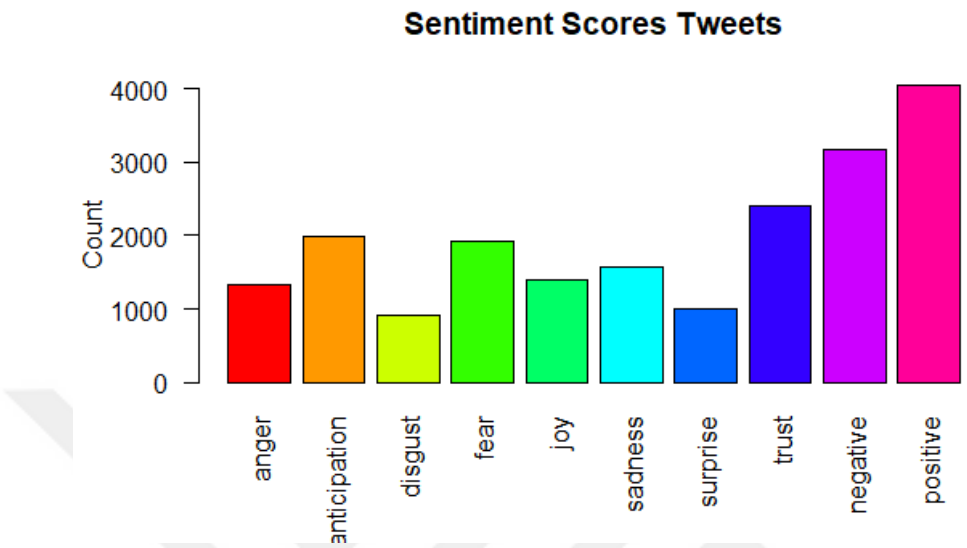

Figure 4.3: Output sentiment analysis for topic 1.

2.Topic: To includes the general feelings of families about the importance that teachers take with the emergence of face-to-face education in the COVID-19 process. In the emotional analysis output, the attitudes of the families towards the importance that the teachers took were positive with the emergence of face-to-face education in the COVID-19 process. In this process, families may believe that teachers will take the necessary importance in face-to-face education and that teachers will pay attention to social distance and classroom cleanliness issues in this process, as the education is face-to-face. In this process, negative thoughts of families who are worried about their children are included in the graph.

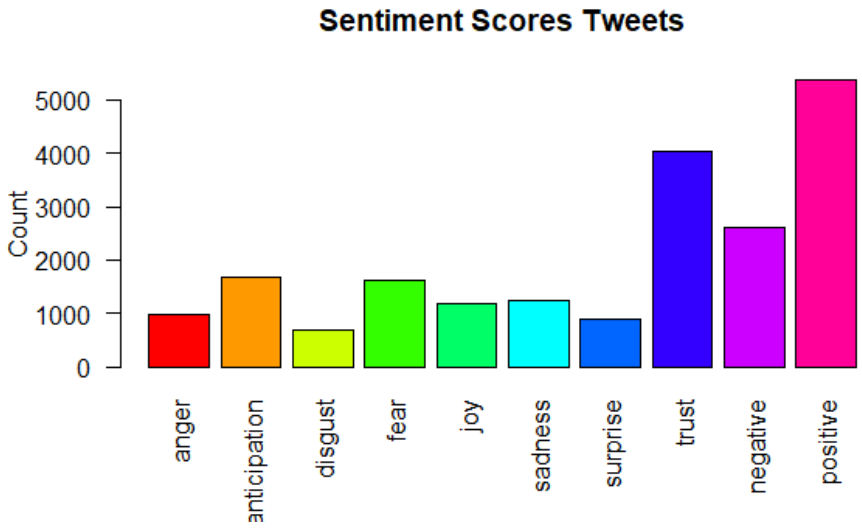

Figure 4.4: Output sentiment analysis for topic 2.

3. Topic: To includes the general feelings of families and students about COVID-19 tests. In the sentiment analysis output, it was revealed that the Families thought positively and trusted the students' COVID-19 tests. People believe and trust that the results of the PCR test are correct. People have fears and negative thoughts about doing the test through the nose and mouth. Although they have a few fears about the construction of the test, people generally believe the COVID-19 test and COVID-19 test result.

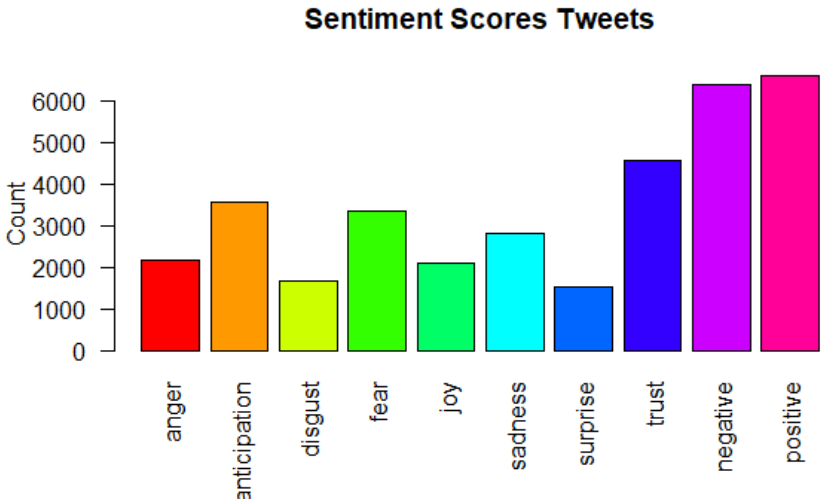

Figure 4.5: Output sentiment analysis for topic 3.

4. Topic: To include their general feelings about time management because of working from home. Working independently from home and being flexible in programs during the

COVID-19 process have mainly provided positive, negative, and trusting feelings for people. Situations such as making presentations from afar and doing homework from afar may have increased people's self-confidence. In addition, students who had difficulty in managing time in this process demanded that this process end in a short time. The students who could not manage their time had more feelings of anticipation and fear.

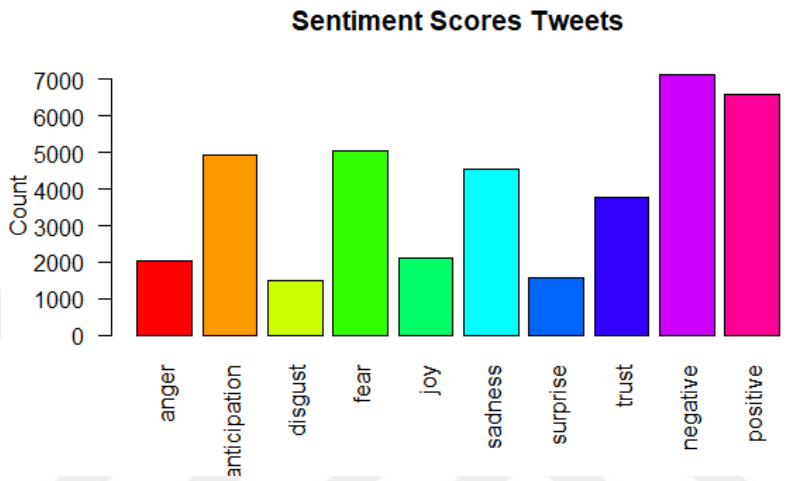

Figure 4.6: Output sentiment analysis for topic 4.

CHAPTER 5

5. CONCLUSIONS AND FUTURE WORK

Since July 2020, COVID19 has been spreading rapidly to every corner of the world. With the outbreak of the COVID19 epidemic, great changes have occurred in people's daily routines. One of the biggest changes has started with the transfer of school processes of students in education age to the online platform. With the transfer of education to online platforms, people have experienced the advantages and disadvantages of being online and have experienced many mixed feelings about this situation. People express their opinions by announcing these feelings and thoughts on the Twitter platform. People write here the right news, the wrong news, the intensity of their feelings and thoughts. In the digital age, where the internet allows ideas to flow faster, more ideas are circulating about right, wrong, wrong and hate speech. During the COVID-19 epidemic, many fake news, hate speech and social scare tactics emerged. In the study, target labels that can be made about education and training during the COVID-19 process were created. These tags were created using the Octoparse program and the comments of the people were taken on Twitter. There are word clouds such as education, freshman, lecture in the tags. It is aimed to find the target topics that students, families, and teachers talk about the most about working from home. In this direction, the data were analyzed using the LDA technique collected from the Topic model and the maximum 4 topics discussed in this field were determined. While creating topic tags, the word cloud searched on Twitter was cleaned during the data cleaning phase. To see the topics more clearly, these words have been cleaned during the cleaning phase so that they are not caught in the most repeated topic frequency. At the same time, during the cleaning phase, the words on Twitter were reduced to the root, non-English numeric numbers and unnecessary spaces were also deleted. Conditions that would affect sentiment analysis were cleared during the data cleaning phase.

1.Topic: It is a set of topics that includes general feelings about working from home during the COVID-19 process.

2.Topic: It is a set of topics that includes the general feelings of families about the importance that teachers take with the emergence of face-to-face education in the COVID-19 process.

3.Topic: It is a set of topics that includes the general feelings of families and students about COVID-19 tests.

4. Topic: It is a set of topics that include their general feelings about time management because of working from home.

Sentiment analysis was performed for each subject. After the creation of hashtags, are people happy about that topic? Is he afraid? Do you have negative thoughts? Sentiment analysis was carried out to get answers to questions such as: In line with these results, while students are generally happy that working from home is comfortable and accessible, they are faced with problems such as lack of time management at home and lack of social life. Families believe that teachers will take the necessary importance in face-to-face classes and provide students with a healthy environment. In line with this study, it is aimed to contribute to the literature on the transfer of education to the online platform in the field of NLP. As a result of the application, it is aimed to shed light on the education policy in the COVID-19 process by presenting a model on the subject and to guide future studies. The next study can be created and looked at separately for primary school high school and university students, and the study can be detailed. At the same time, this study was separated using the sentiment analysis library. A library related to the subject can be created in further studies.

BIBLIOGRAPHY

[1] National Health Commission of the People's Republic of China.“Chinese Center for Disease Control and Prevention.” https://www.chinacdc.cn/en/ (accessed Nov. 02, 2021).

[2] Johns Hopkins Coronavirus. “COVID-19 Map-Johns Hopkins Coronavirus Resource Center.” https://coronavirus.jhu.edu/map.html (accessed Nov. 24, 2021).

[3] Our World in Data. Anadolu Ajansı “ people per million died in the world due to covid-19”. “https://www.aa.com.tr/tr/info/infografik/23093 (accessed Nov. 02, 2021).

[4] “Coronavirus disease (COVID-19): How is it transmitted?” https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted (accessed Nov. 02, 2021).

[5] “WHO EMRO | EMRO home page | Landing | Front page.” http://www.emro.who.int/index.html (accessed Nov. 27, 2021).

[6] W. Cao et al., “The psychological impact of the COVID-19 epidemic on college students in China,” Psychiatry Res., vol. 287, no. March, p. 112934, 2020, doi: 10.1016/j.psychres.2020.112934.

[7] B. Danışma and K. Çalışması, “COVID-19 Pandemisinde Sağlık Kurumlarında Çalışma Rehberi ve Enfeksiyon Kontrol Önlemleri.”

[8] Unicef. “COVID-19 and School Closures: Are children able to continue learning - UNICEF DATA.” https://data.unicef.org/resources/remote-learning-reachability-factsheet/ (accessed Nov. 28, 2021).

[9] C. Lee and C. M. Lee, “Descriptive SWOT Analysis about Online Learning,” no. April, pp. 0–10, 2021.

[10] E. Kasthuri and S. Balaji, “Natural language processing and deep learning chatbot using long short term memory algorithm,” Mater. Today Proc., pp. 4–7, 2021, doi: 10.1016/j.matpr.2021.04.154.

[11] C. S. R. Chan, C. Pethe, and S. Skiena, “Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowdfunding outcomes,” J. Bus. Ventur. Insights, vol. 16, no. June, p. e00276, 2021, doi: 10.1016/j.jbvi.2021.e00276.

[12] S. Alam, “Applying Natural Language Processing for detecting malicious patterns in Android applications,” Forensic Sci. Int. Digit. Investig., vol. 39, p. 301270, 2021, doi: 10.1016/j.fsidi.2021.301270.

[13] G. Perboli, M. Gajetti, S. Fedorov, and S. Lo Giudice, “Natural Language Processing for the identification of Human factors in aviation accidents causes: An application to the SHEL methodology,” Expert Syst. Appl., vol. 186, no. July, p. 115694, 2021, doi: 10.1016/j.eswa.2021.115694.

[14] S. Doan, E. W. Yang, S. Tilak, and M. Torii, “Using Natural Language Processing to Extract Health-Related Causality from Twitter Messages,” Proc. -

2018 IEEE Int. Conf. Healthc. Informatics Work. ICHI-W 2018, pp. 84–85, 2018, doi: 10.1109/ICHI-W.2018.00031.

[15] D. Van Le, J. Montgomery, K. C. Kirkby, and J. Scanlan, "Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting," J. Biomed. Inform., vol. 86, no. February, pp. 49–58, 2018, doi: 10.1016/j.jbi.2018.08.007.

[16] R. Y. Lee et al., "Identifying Goals of Care Conversations in the Electronic Health Record Using Natural Language Processing and Machine Learning," J. Pain Symptom Manage., vol. 61, no. 1, pp. 136-142.e2, 2021, doi: 10.1016/j.jpainsymman.2020.08.024.

[17] S. Kulshrestha et al., "Prediction of severe chest injury using natural language processing from the electronic health record," Injury, vol. 52, no. 2, pp. 205–212, 2021, doi: 10.1016/j.injury.2020.10.094.

[18] A. Forestiero and G. Papuzzo, "Natural language processing approach for distributed health data management," Proc. - 2020 28th Euromicro Int. Conf. Parallel, Distrib. Network-Based Process. PDP 2020, pp. 360–363, 2020, doi: 10.1109/PDP50117.2020.00061.

[19] N. Mathews, T. Tran, B. Rekabdar, and C. Ekenna, "Jou rna lP pro of," Informatics Med. Unlocked, p. 100738, 2021, doi: 10.1016/j.imu.2021.100738.

[20] M. D. Solomon, G. Tabada, A. Allen, S. H. Sung, and A. S. Go, "Large-scale identification of aortic stenosis and its severity using natural language processing on electronic health records," Cardiovasc. Digit. Heal. J., vol. 2, no. 3, pp. 156–163, 2021, doi: 10.1016/j.cvdhj.2021.03.003.

[21] A. Kormilitzin, N. Vaci, Q. Liu, and A. Nevado-Holgado, "Med7: A transferable clinical natural language processing model for electronic health records," Artif. Intell. Med., vol. 118, no. August 2020, p. 102086, 2021, doi: 10.1016/j.artmed.2021.102086.

[22] A. W. Olthof et al., "Machine learning based natural language processing of radiology reports in orthopaedic trauma," Comput. Methods Programs Biomed., vol. 208, p. 106304, 2021, doi: 10.1016/j.cmpb.2021.106304.

[23] J. R. Parikh et al., "A data-driven architecture using natural language processing to improve phenotyping efficiency and accelerate genetic diagnoses of rare disorders," Hum. Genet. Genomics Adv., vol. 2, no. 3, p. 100035, 2021, doi: 10.1016/j.xhgg.2021.100035.

[24] "WHO 2020, "Coronavirus disease 2019 (COVID-19) Situation Report – 55" https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200315-sitrep-55-covid-19.pdf?sfvrsn=33daa5cb_8.

[25] E. Massaad and P. Cherfan, "Social Media Data Analytics on Telehealth During the COVID-19 Pandemic," Cureus, vol. 12, no. 4, pp. 1–7, 2020, doi: 10.7759/cureus.7838.

[26] P. S.V and R. Ittamalla, "General public's attitude toward governments implementing digital contact tracing to curb COVID-19 – a study based on

natural language processing," Int. J. Pervasive Comput. Commun., 2020, doi: 10.1108/IJPCC-09-2020-0121.

[27] A. Chapman, K. Peterson, A. Turano, T. Box, K. Wallace, and M. Jones, "A Natural Language Processing System for National COVID-19 Surveillance in the US Department of Veterans Affairs," Proc. 1st Work. NLP COVID-19 ACL 2020, 2020.

[28] A. Ebadi, P. Xi, S. Tremblay, B. Spencer, R. Pall, and A. Wong, "Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing," Scientometrics, vol. 126, no. 1, pp. 725–739, 2021, doi: 10.1007/s11192-020-03744-7.

[29] J. L. Izquierdo, J. Ancochea, and J. B. Soriano, "Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients with COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing," J. Med. Internet Res., vol. 22, no. 10, pp. 1–13, 2020, doi: 10.2196/21801.

[30] J. M. Banda, "Long-term patient-reported symptoms of COVID-19 : an analysis of social media data," 2020.

[31] Y. Hu et al., "From SARS to COVID-19: A Bibliometric study on Emerging Infectious Diseases with Natural Language Processing technologies," 2020, doi: 10.21203/rs.3.rs-25354/v1.

[32] Y. Liu, C. Whitfield, T. Zhang, A. Hauser, T. Reynolds, and M. Anwar, "Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning," Heal. Inf. Sci. Syst., vol. 9, no. 1, pp. 1–16, 2021, doi: 10.1007/s13755-021-00158-4.

[33] V. Gupta et al., "An Emotion Care Model using Multimodal Textual Analysis on COVID-19," Chaos, Solitons and Fractals, vol. 144, p. 110708, 2021, doi: 10.1016/j.chaos.2021.110708.

[34] E. Kasson, A. K. Singh, M. Huang, D. Wu, and P. Cavazos-Rehg, "Using a mixed methods approach to identify public perception of vaping risks and overall health outcomes on Twitter during the 2019 EVALI outbreak," Int. J. Med. Inform., vol. 155, no. March, p. 104574, 2021, doi: 10.1016/j.ijmedinf.2021.104574.

[35] D. Thorpe Huerta, J. B. Hawkins, J. S. Brownstein, and Y. Hswen, "Exploring discussions of health and risk and public sentiment in Massachusetts during COVID-19 pandemic mandate implementation: A Twitter analysis," SSM - Popul. Heal., vol. 15, no. June, 2021, doi: 10.1016/j.ssmph.2021.100851.

[36] A. Z. Klein, A. Magge, K. O'Connor, J. I. F. Amaro, D. Weissenbacher, and G. G. Hernandez, "Toward Using Twitter for Tracking COVID-19:A Natural Language Processing Pipeline and Exploratory Data Set," J. Med. Internet Res., vol. 23, no. 1, 2021, doi: 10.2196/25314.

[37] L. Tang et al., "Texas public agencies' tweets and public engagement during the COVID-19 pandemic: Natural language processing approach," JMIR Public Heal. Surveill., vol. 7, no. 4, 2021, doi: 10.2196/26720.

[38] A. F. Pathan and C. Prakash, “Unsupervised Aspect Extraction Algorithm for Opinion Mining using Topic Modeling,” Glob. Transitions Proc., pp. 0–9, 2021, doi: 10.1016/j.gltp.2021.08.005.

[39] T. Ma, X. Zhou, J. Liu, Z. Lou, Z. Hua, and R. Wang, “Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies,” Technol. Forecast. Soc. Change, vol. 173, no. June 2020, p. 121159, 2021, doi: 10.1016/j.techfore.2021.121159.

[40] J. Ahn, H. Son, and A. D. Chung, “Understanding public engagement on twitter using topic modeling: The 2019 Ridgecrest earthquake case,” Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, p. 100033, 2021, doi: 10.1016/j.jjimei.2021.100033.

[41] F. Pérez, R. Lapeña, A. C. Marcén, and C. Cetina, “Topic modeling for feature location in software models: Studying both code generation and interpreted models,” Inf. Softw. Technol., vol. 140, no. November 2020, p. 106676, 2021, doi: 10.1016/j.infsof.2021.106676.

[42] Q. Xie, P. Tiwari, D. Gupta, J. Huang, and M. Peng, “Neural variational sparse topic model for sparse explainable text representation,” Inf. Process. Manag., vol. 58, no. 5, p. 102614, 2021, doi: 10.1016/j.ipm.2021.102614.

[43] V. Gangadharan and D. Gupta, “Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques,” Procedia Comput. Sci., vol. 171, no. 2019, pp. 1337–1345, 2020, doi: 10.1016/j.procs.2020.04.143.

[44] T. Goshima, K. Honda, S. Ubukata, and A. Notsu, “Deterministic annealing process for pLSA-induced fuzzy co-clustering and cluster splitting characteristics,” Int. J. Approx. Reason., vol. 95, pp. 185–193, 2018, doi: 10.1016/j.ijar.2018.02.005.

[45] B. Cao, X. Frank Liu, J. Liu, and M. Tang, “Domain-aware Mashup service clustering based on LDA topic model from multiple data sources,” Inf. Softw. Technol., vol. 90, pp. 40–54, 2017, doi: 10.1016/j.infsof.2017.05.001.

[46] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, “Evaluation of clustering and topic modeling methods over health-related tweets and emails,” Artif. Intell. Med., vol. 117, no. May 2020, p. 102096, 2021, doi: 10.1016/j.artmed.2021.102096.

[47] D. Zhou, J. Yuan, and J. Si, “Health issue identification in social media based on multi-task hierarchical neural networks with topic attention,” Artif. Intell. Med., vol. 118, no. May, p. 102119, 2021, doi: 10.1016/j.artmed.2021.102119.

[48] Z. Huang, W. Dong, and H. Duan, “A probabilistic topic model for clinical risk stratification from electronic health records,” J. Biomed. Inform., vol. 58, pp. 28–36, 2015, doi: 10.1016/j.jbi.2015.09.005.

[49] Q. Liu et al., “Health Communication About Hospice Care in Chinese Media: Digital Topic Modeling Study,” JMIR Public Heal. Surveill., vol. 7, no. 10, p. e29375, 2021, doi: 10.2196/29375.

[50] J. Liu et al., “Tracing the Pace of COVID-19 Research: Topic Modeling and Evolution,” Big Data Res., vol. 25, p. 100236, 2021, doi: 10.1016/j.bdr.2021.100236.

[51] G. P. Cuaton, L. J. B. Caluza, and J. F. V. Neo, “A topic modeling analysis on the early phase of COVID-19 response in the Philippines,” Int. J. Disaster Risk Reduct., vol. 61, no. May 2020, p. 102367, 2021, doi: 10.1016/j.ijdrr.2021.102367.

[52] F. Kaveh-Yazdy and S. Zarifzadeh, “Track Iran’s national COVID-19 response committee’s major concerns using two-stage unsupervised topic modeling,” Int. J. Med. Inform., vol. 145, no. September 2020, p. 104309, 2021, doi: 10.1016/j.ijmedinf.2020.104309.

[53] G. Burel, T. Farrell, and H. Alani, “Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter,” Inf. Process. Manag., vol. 58, no. 6, p. 102732, 2021, doi: 10.1016/j.ipm.2021.102732.

[54] J. X. Koh and T. M. Liew, “How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds,” J. Psychiatr. Res., no. September, 2020, doi: 10.1016/j.jpsychires.2020.11.015.

[55] C. Ordun, S. Purushotham, and E. Raff, “Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs,” no. March, 2020, [Online]. Available: http://arxiv.org/abs/2005.03082.

[56] A. E. de Oliveira Carosia, G. P. Coelho, and A. E. A. da Silva, “Investment strategies applied to the Brazilian stock market: A methodology based on Sentiment Analysis with deep learning,” Expert Syst. Appl., vol. 184, no. September 2020, p. 115470, 2021, doi: 10.1016/j.eswa.2021.115470.

[57] T. Daudert, “Exploiting textual and relationship information for fine-grained financial sentiment analysis,” Knowledge-Based Syst., vol. 230, p. 107389, 2021, doi: 10.1016/j.knosys.2021.107389.

[58] H. Li, Q. Chen, Z. Zhong, R. Gong, and G. Han, “E-word of mouth sentiment analysis for user behavior studies,” Inf. Process. Manag., vol. 59, no. 1, p. 102784, 2022, doi: 10.1016/j.ipm.2021.102784.

[59] X. Fang and J. Zhan, “Sentiment analysis using product review data,” Journal of Big Data, vol. 2, no. 1. 2015, doi: 10.1186/s40537-015-0015-2.

[60] M. Palomino, T. Taylor, A. Göker, J. Isaacs, and S. Warber, “The online dissemination of nature–health concepts: Lessons from sentiment analysis of social media relating to ‘nature-deficit disorder,’” Int. J. Environ. Res. Public Health, vol. 13, no. 1, 2016, doi: 10.3390/ijerph13010142.

[61] J. Carrillo-de-Albornoz, J. R. Vidal, and L. Plaza, “Feature engineering for sentiment analysis in e-health forums,” PLoS One, vol. 13, no. 11, pp. 1–25, 2018, doi: 10.1371/journal.pone.0207996.

[62] A. Žunić, P. Corcoran, and I. Spasić, “Aspect-based sentiment analysis with graph convolution over syntactic dependencies,” Artif. Intell. Med., vol. 119, no. December 2020, 2021, doi: 10.1016/j.artmed.2021.102138.

[63] C. Liu et al., “Improving sentiment analysis accuracy with emoji embedding,” J. Saf. Sci. Resil., 2021, doi: 10.1016/j.jnlssr.2021.10.003.

[64] A. Hussain et al., “Artificial intelligence-enabled analysis of public attitudes on facebook and twitter toward COVID-19 vaccines in the United Kingdom and the

United States: Observational study," J. Med. Internet Res., vol. 23, no. 4, pp. 1–10, 2021, doi: 10.2196/26627.

[65] D. Obembe, O. Kolade, F. Obembe, A. Owoseni, and O. Mafimisebi, "Covid-19 and the tourism industry: An early stage sentiment analysis of the impact of social media and stakeholder communication," Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, p. 100040, 2021, doi: 10.1016/j.jjimei.2021.100040.

[66] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," Knowledge-Based Syst., vol. 228, p. 107242, 2021, doi: 10.1016/j.knosys.2021.107242.

[67] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, "An analysis of COVID-19 vaccine sentiments and opinions on Twitter," Int. J. Infect. Dis., vol. 108, pp. 256–262, 2021, doi: 10.1016/j.ijid.2021.05.059.

[68] M. M. Rahman et al., "Socioeconomic factors analysis for COVID-19 US reopening sentiment with Twitter and census data," Heliyon, vol. 7, no. 2, p. e06200, 2021, doi: 10.1016/j.heliyon.2021.e06200.

[69] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA," Appl. Soft Comput., vol. 101, p. 107057, 2021, doi: 10.1016/j.asoc.2020.107057.

[70] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," IEEE Access, vol. 8, pp. 181074–181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[71] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf., vol. 1, pp. 1354–1364, 2015, doi: 10.3115/v1/p15-1131.

[72] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," Appl. Soft Comput. J., vol. 97, p. 106754, 2020, doi: 10.1016/j.asoc.2020.106754.

[73] A. Gregoriades, M. Pampaka, H. Herodotou, and E. Christodoulou, "Supporting digital content marketing and messaging through topic modelling and decision trees," Expert Syst. Appl., vol. 184, no. June, p. 115546, 2021, doi: 10.1016/j.eswa.2021.115546.

[74] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence," J. Infect. Public Health, vol. 14, no. 10, pp. 1505–1512, 2021, doi: 10.1016/j.jiph.2021.08.010.

[75] L. Wright, A. Burton, A. McKinlay, A. Steptoe, and D. Fancourt, "Public Opinion about the UK Government during COVID-19 and Implications for

Public Health: A Topic Modelling Analysis of Open-Ended Survey Response Data," medRxiv, p. 2021.03.24.21254094, 2021, [Online]. Available: https://www.medrxiv.org/content/10.1101/2021.03.24.21254094v1%0Ahttps://www.medrxiv.org/content/10.1101/2021.03.24.21254094v1.abstract.

[76] M. S. Satu et al., "TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets," Knowledge-Based Syst., vol. 226, p. 107126, 2021, doi: 10.1016/j.knosys.2021.107126.

[77] A. V. K. Kaila, R.P. & Prasad, "Informational Flow on Twitter - Corona Virus Outbreak – Topic," Int. J. Adv. Res. Eng. Technol., vol. 11, no. 3, pp. 128–134, 2020.

[78] S. Momtazi and F. Naumann, "Topic modeling for expert finding using latent Dirichlet allocation," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 3, no. 5, pp. 346–353, 2013, doi: 10.1002/widm.1102.

CURRICULUM VITAE

Name Surname : Lütviye Özge POLATLI

EDUCATION:

B.Sc. : 2020, Istanbul Medipol University, School of Engineering, and Natural Sciences, Industrial Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2015-2020, Istanbul Medipol University

OTHER PUBLICATIONS, PRESENTATIONS, AND PATENTS:

- E. Delice, L. O. Polatlı, H. Tozan, ve I. Argun Düzdar "Applications of Data Mining Algorithms for Customer Recommendations in Retail Marketing," The Future of Data Mining, c.3, no. 1, Nova Science, 2022, pp. 29–49.

- E. Delice, L. O. Polatlı, K. Abujbara, H. Tozan, and A. Erturk "Digitalization in Healthcare: A Systematic Review of the Literature" presented at the International Scientific Conference on Digital Transformation in Business: New Challenges and Opportunities. (ISCDT), 2022.

NATURAL LANGUAGE PROCESSING ANALYSIS OF COMMENTS ABOUT EDUCATION ON TWITTER DURING THE COVID-19

ORIGINALITY REPORT

14%	10%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	acikerisim.medipol.edu.tr Internet Source	1%
2	www.analyticsvidhya.com Internet Source	<1%
3	Submitted to University of the West Indies Student Paper	<1%
4	worldwidescience.org Internet Source	<1%
5	www.researchgate.net Internet Source	<1%
6	docs.edtechhub.org Internet Source	<1%
7	www.ncbi.nlm.nih.gov Internet Source	<1%
8	Momtazi, Saeedeh, and Felix Naumann. "Topic modeling for expert finding using latent Dirichlet allocation : Topic modeling for expert finding using LDA", Wiley	<1%