Descriptive Analysis and Topic Modelling of X Posts to Detect Changes of Customer Trends in Pandemic Period: A Case Study of Jeans

Yıldırım Güneş[1], Murat Arıkan[2]

[1](*Graduate School of Natural and Applied Sciences, Department of Supply Chain and Logistcs Management / Gazi University, Turkey)*

[2](*Engineering Faculty, Department of Industrial Engineering / Gazi University, Turkey)*

ABSTRACT : *X (previously named Twitter), one of the social media platforms, is used for collecting data from customers and converting the data into valuable information for business. While X analyses are mostly conducted on current, political, cultural and daily issues that are of interest or rapidly spreading in society, sectoral analyses are conducted less frequently. The aim of the study is to reveal the effects of the pandemic on consumer trends and to extract the topics from consumers' tweets. This paper deals with analyzing dataset of X messages related to jeans before pandemic and during the pandemic. The study focused on a niche topic like the clothing industry, through X by applying these three approaches: 1. A statistical and time analysis of descriptive features of tweets, 2. Topic modelling with words of context, 3. Revealing the change between pre-pandemic and pandemic period. Here, a total of 28 265 tweets posted between December 15, 2019 and December 31, 2020 was collected and processed for analysis. At the end of the study, it is determined that the most appropriate method for content detection of our dataset is unigram and LDA, and pre-pandemic agenda topics are categorized under 8 headings. Four new topics are added to these categories during the pandemic period.*

KEYWORDS - *X (Twitter) analysis, topic modelling, descriptive analysis, statictical analysis, jeans (kot pantolon), clothing sectorial analysis, pandemic effect*

I. INTRODUCTION

X (previously named Twitter), as a social networking, text intensive and user-generated content (UGC) platform, is an important source for extracting innovative ideas from customers' texts for business. For this reason, X is the subject of many studies to gain insights of customers, especially those using text analysis methods. Text analysis has its own challenges. The difficulty of extracting information or topic modelling from short text micro blogs like X is due to the fact that they contain short and noisy data that may lead to incorrect inference [1].

Although there are many researches on X text analysis, there are still very few studies on specific areas other than popular topics such as politics and culture. X analysis focuses on topics such as health, politics, society and social media. Due to the Covid-19 outbreak, there has been an increase in research on health issues during the pandemic period. However, it is seen that sectoral research on the business domain remains quite low, and the total research rate on the business domain between 2009-2021 remained at the level of 2 % [2].

The research subject in the study is intentionally selected on a niche topic as jeans. To the best of our knowledge the topic has not been analyzed by using tweets' texts in Turkish literature.

Therefore, we think that the study will contribute to the domain specific researches on Turkish keyword and text-oriented studies. We also hope that it will encourage researchers in the analysis of such niche topics.

Increasing sectoral demands or directing these demands to its own organization is one of the main strategic goals of businesses. In order to achieve this goal, potential customer interests or trends must be determined correctly. The determinations made contribute to the increase in demand by directing the product development efforts of businesses. Statistical analysis and topic modeling tools are effective methods for determining customer trends and interests and may vary depending on the sectors and analyzed data sets. The main purpose of the study is to first reveal the sectoral trends and agenda of potential customers and to identify the changes in these trends and agendas in the clothing sector, especially for jeans, due to the pandemic. Secondly, to determine the methods that will give more meaningful results for the detection of these trends.

The pandemic outbreak occurred in 2019 significantly affected people's daily lives, habits or tendencies. Many studies have been conducted on how the pandemic period affected people's tendencies, such as working from home or ordering food to home [3,4]. Similar changes can be seen in other sectors. For example, identifying changes in people's shopping habits can provide insights for different sectors.

The use of social media channels to share or query various topics has increased during the Covid-19 period [3,5]. The same trend is expected to occur in the clothing industry. The changes in customer behavior can also be captured through social media platforms such as X. Augmenting textual sharing, especially in X, will increase the diversity of data and datasets for determining public interest in specific niche topics and will allow for more detailed textual analyses with a larger number of observations. Thus, it will be possible to obtain more accurate ideas about the direction of customer trends in sectoral issues.

Natural language processing techniques, text mining are the basic techniques used to extract information from human-written texts and UGC. Using tools such as descriptive analysis, sentiment analysis and network analysis, it is possible to extract information and other details of datasets from the metadata of tweets. Descriptive analysis can provide the opportunity to reach innovative ideas with less effort and less technical detail than the other types of analysis mentioned [3].

Descriptive analysis of tweets shared on X can be performed using text content, word and hashtag frequencies, users' activities, follower counts, original tweet sharing and retweeting, and statistics derived from correlating these numbers with time, and specific inferences can be made for the sector analyzed. By determining words, word groups and their frequencies and statistical operations performed on them, the subjects of the texts, the topic issues of the datasets and customer trends can be determined [1]. These analyses were conducted on pre-pandemic and pandemic period datasets, aiming to reveal the differences and changes between the two periods specifically for jeans, a clothing sector that is rarely encountered in tweet analyses.

The contribution of this study is to present the information in the dataset for jeans, a special area of the clothing industry, through statistical and descriptive analysis, and to determine the changes in the industry with the pandemic and the focal points of customer opinions.

The statictical analysis of text and metadata of tweets can provide business insights but there are some limitations. Firstly, the dataset used in the research is created with data obtained within the framework of the numerical limitations applied by the X platform in extracting data from publicly available tweets. Secondly, the datasets were collected by using specific keywords which limits the perspective of the research.

The paper is organized as follows: Studies on topic modeling and content analysis, mostly related to the pandemic period, are listed in the second section. The aims, methods, tools and contributions of the studies on the use of n-grams, LDA and TURKISHBERTweet methods, especially on X texts, are briefly explained. In third section, all phases of the analysis from data collection to information acquisition are explained in detail, along with the tools used. This section also covers topic modeling in the context of word analysis in detail. The fourth section discusses and evaluates statistical analysis and topic modelling/content analysis results in detail. The final section presents the conclusions, proposal for business and areas for future research.

II. LITERATURE REVIEW

Human-written texts can be beneficial for business insights. These business insights and useful information can be obtained by analyzing texts [6]. As a text-based UGC platform, X is a valuable resource for various statistical analyses with texts and metadata received from users. This valuable resource can be used for innovative business ideas targeting these users [7].

The posts in X, together with their text and metadata, are an important data source. In order to transform this data into meaningful and valuable information, Chae [8] presents a framework in the form of descriptive, content and network analysis. Among these, statistical analyses are basically performed on the user, tweet and tweet text. Bruns and Stieglitz [9] emphasized that combining these analyses with time periods can also provide practical information about the users' activities depending on time. It is possible to gain insights about all dataset by analysing selected part of it, instead of examining the entire dataset depending on the purpose and target audience of the research.

There is a 280-character limit for text in X. Depending on the subject of posts, the character limit of X for text writing may not be sufficient for users. The text language of the posts is adapted to user needs with additional shapes, symbols or abbreviations for effective use and to say more with fewer characters [10]. However, it is also seen that the number of characters in text writing in X can remain at low levels, depending on the topics of conversation, especially in a sectoral area [3]. On the other hand, these narrative conveniences in users' texts emerge as noise that needs to be cleaned up in text analysis.

Descriptive statistics such as word counts, tweets, retweets, and sentence lengths allow us to identify the dataset, users, and their features. For example, the number of original tweets sent by a user for the first time, called chat starting tweet, reflects the diversity of new ideas in the dataset. In addition, original tweets can help identify influential users. Retweets are also important tools for spreading any information in tweets and understanding the tendencies of the users [11]. When a user finds a topic interesting and retweets it for their followers to see, the visibility of that topic among users increases. Highly visible topics can provide insight into users' agendas and interests. It should also be noted that attractiveness is a subjective issue and can vary from user to user [12]. On the other hand, RTs can manipulate the content of the dataset (especially due to advertising or propaganda RTs) and therefore cause misleading effects in content-oriented analyses. It is also possible to artificially increase the volume of retweets through fake users [13].

Another important tool for discovering tweet texts in terms of content and detecting user tendencies is the analysis of the statistical values of words, keywords and hashtags. These analyses, performed with algorithms such as feature extraction and topic modeling, enable the classification of texts in terms of content and the understanding of prominent topics in the content. Studies on topic modeling and content analysis tools in the literature are listed in Table 1.

Table 1: Summary of literature review of topic modelling and content analysis

Paper Ref.No.	Data collection	Purpose (Related with COVID-19)	Methods, Tools Used	Results/Contributions
[14]	-Twitter API, keywords: COVID-19 and related with COVID-19; - March 1 and July 31, 2022;	to develop a technique for summarizing topics related with COVID-19	Latent Dirichlet Allocation (LDA), n-grams, K-means clustering, ROUGE metrics	Methods including stages of data analysis for topic summarization,

	- 100,000 tweets.			
[15]	-Twitter API, key-words to filter COVID-19 related tweets; - 3 -13 Apr 2020; -46 million tweets scraped over.	Identifying topics and notable topic trends among people	Sequential LDA	An understanding of the topics surrounding the COVID-19 pandemic and their evolution over time
[16]	-Articles from Croatian portal Tportal.hr, - 1 Jan-19 Feb 2021; -12,080 related articles	Identifying topics from the Croatian internet portal during the pandemic period	LDA, NRC-lexicon	The topics are vaccination and earthquake. All extracted topics are predominantly negative emotions (anticipation, surprise, sadness and fear)
[17]	-Collected articles -50 articles	Topic detection	Literature review	Classifying the algorithms and methods tracking real-world incidents from Twitter
[18]	-Tweepy; -March 24- April 9, 2020; -23,830,322 tweets	Assessing the distinctiveness of topics / key terms / features, speed of information dissemination and network behaviors for Covid-19 tweets by using five different technics	Pattern matching and topic modeling through LDA to generate twenty different topics	The methods identify the unique clustering behavior of different topics to obtain important themes in the corpus and help assess the quality of the generated topics.
[19]	- Twint project tool; - March 2020-November 2021; -3 000 000	To investigate the effects of lockdown, carantina measures	Sentiment analysis, LDA topic modeling on textual data	-Even with the continuity of lockdown measures in Malaysia, the sentiments expressed on Twitter were primarily positive. -The topics discussed among the people are highlighted in each lockdown and related keywords.
[20]	-Feb-May 2020; -39,073 words collected from sixteen speeches in two different periods of pandemic;	to understand the psychological functions of language affected during COVID-19 pandemic	Linguistic Inquiry Word Count (LIWC) program	A significant change was seen in verbal behaviors in 2020, especially in the most prominent words and psychological functions.
[21]	-Tweets Kaggle dataset, #Covid-19; -17 000 tweets; -25 July 2020	Conducting exploratory statistical analysis of posts of COVID-19 period	Statistical analysis, word frequency analysis , n-	Most of the tweets were found to have neutral sentiment polarity.

			grams, and sentiment analysis.	
[22]	-Keyword: "wabah corona", - 13,670 tweets; - January 9-May 11, 2020;	Analyzing tweets related to COVID-19 in Indonesia.	LDA for topic modelling.	13 ideal topic segments detected
[23]	- Twitterscraper API, keyword: "#pfw," "#mfw," "#nyfw," and "#lfw," - 33 525 records - 8 Feb-5 March 2019	Analyzing social media data from four cities during the 2019 Fashion Week.	Topic modelling, sentiment analysis.	It is confirmed that brands that embodied similar themes in terms of topics and had positive sentimental reactions were also most frequently mentioned by the consumers.
[24]	Twitter posts related with vaccine safety signals.	Evaluating topic models for identifying user posts related with vaccine safety signals	Topic modelling with Gensim Topological Data Analysis (TDA), java for LDA Dirichlet Multinomial Mixture (DMM) models	Method proposal to detect topics that best revealed documents.
[25]	-7,000 CBC COVID-19 related news articles and 100,000 research manuscripts - Jan 9-May 3,2020 and January 2-Aug 1 2020	to deal with a large and computationally intensive corpus for topic modelling	LDA	Proposal for LDA technics was developed and main topics were detected in datasets.
[26]	- 530 000 tweets on 2018 Worldcup, - 760 660 tweets on Games of Thrones, - 42 013 tweets on 2016 US Elections, - 179 108 tweets on COVID-19	Modelling user interactions on Twitter as a weighted and directed network by using social network analysis.	Topic modelling with LDA and network analysis (Pagerank algorithm, Greedy Modular Algorithm and the Leiden Algorithm) to detect influential users,	A four step process is proposed to connect the topics with topic modelling and the users with network analysis.
[27]	- 22-30 Mar 2020; - 43 000 000 tweets;	Helping organizations to make better decisions and prioritize tasks	N-grams (unigram, bigram and trigram)	Unigram terms were trended from 3 to 70 times more frequently than bigram and trigram terms.

III. DATA AND METHODS

Data analysis process with all phases and the tools for analysis methods applied in the study is shown in Figure 1. In the study, the descriptive features of the datasets were manipulated by statistical analysis method, and for this purpose, Python programming language and libraries were utilized in all phases. The main purpose of the study is to reveal the sectoral trends and topics of the customers and to identify the changes in these trends and topics due to the pandemic. These trends/topics from potential customers can be utilized for product improvement/development in order to increase demand for the product. Detection of changes in the sector caused by unexpected situations such as pandemics that change the flow of daily life can shed light on similar situations in the future for businesses in the sector.

In particular, inferences can be made about the content and agenda of datasets through word analysis. Accordingly, among the phases shown in Figure 1, three different topic modelling methods were applied in the word analysis to extract the content and the main topics of the datasets. The main goal of the word analysis here is to reach the following conclusions: (i) Identify the method closest to manual selection of words in the datasets, (ii) Identify similarities in the clothing business (specifically on the topic of "jeans") between the pre-COVID and COVID period, (iii) Whether there is any similarity in the content when the retweets are not extracted from the datasets?

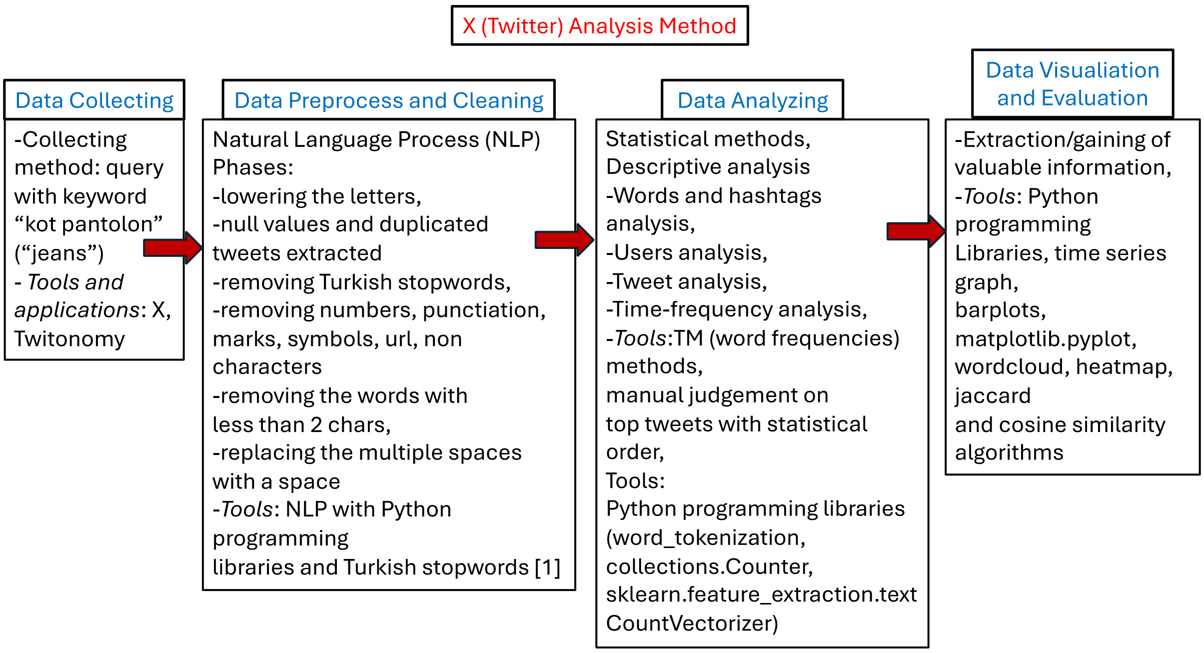

Fig. 1: Data analysis process/phases and tools for analysis methods used in the study

The original dataset was created using the Turkish keyword "kot pantolon" (jeans) from X with the Twitonomy application. The dataset was seperated into two parts due to the date of pandemic. In the following sections, we use the terms "dataset", "dataset-1" and "dataset-2" to refer to "the entire dataset including all tweets", "tweets posted during the pre-pandemic period" and "tweets posted during the pandemic period", respectively. Dataset-1 includes tweets between December 15, 2019 and March 15, 2020. Dataset-2 includes tweets between March 16, 2020, when the government announced pandemic measures in Turkey, and December 31, 2020.

One of the problems that can be encountered in selecting such niche topics is the low tweet shares on this topic. It has been observed that this situation continued during the pandemic period. Therefore, in order to increase the number and diversity of tweets in the dataset, the data collection period was set to be more than one year. The fact that the dataset covers an annual period also made it possible to make periodic determinations in detecting the user trends.

It was applied traditional Natural Language Process shownin Figure 1 with Python programming to clean the datasets. During the cleaning process, stop words utilized in the literature were removed from the tweets of the dataset [3]. No stemming and lemmization was performed for the words in the cleaning operations. As a result, statistical values were determined based on the current form of words in tweet sentences.

It is stated that users tend to share and retweet more on topics related to their interests and expertise. One of the factors that make a tweet retweeted is its content [28]. In this respect, retweets will have significant contributions to the dataset analysis results, especially on content analysis. On the other hand, it is also possible that retweets have directed, manipulative effects on datasets. Retweeting sometimes can be a political activity or marketers want to be a part of the chats by retweeting [29]. In these sitiuations, retweeting can be supported by political fans or marketers' business partners/collabrotars/followers. Such retweeting can have a directive effect on the topics of the datasets. Therefore, depending on the purpose of the analysis and the study, it may be decided to remove RTs from the datasets or the RTs may be subjected to further analysis [3,30]. Based on this, while retweets were taken into account in the analyses in some parts of the study, they were not taken into account in some other parts. In order to see the effects of RTs on the analyses of the study, the datasets in which RTs were included were also analyzed and comparisons were made.

Statistical results from the metadata of tweets allow discovering valuable information about the dataset. The analysis stages applied in the study are shown in Figure 1. For the word analysis shown in Figure 2, three different feature extraction/topic modeling algorithms that can provide more detailed information about the content of the datasets were used, and the results obtained from here were compared with the results obtained manually. At the interpretation phase of the analyses, barplot, scatterplot, wordcloud and heatmap and time graphs were used to make comparisons at the word analysis phase.

Topic modeling algorithms that examine the content of datasets may have different levels of performance on datasets from different domains. To identify the effective topic modeling algorithm on our dataset, examining the content of tweets was performed using a process shown in Figure 2. A word list was created by determining the word frequencies of the datasets. Words that are not related to the subject of the datasets were removed from this word list. This manually prepared list was used to measure the performance of the word lists obtained with the n-gram, LDA and TURKISH BER Tweet algorithms. For this purpose, the overlap rates of the words obtained by the algorithms in question with the manually detected words were determined. The proportional method Jaccard and the weighted method Cosine algorithms were used to determine the overlap rates. Heatmap graphics were used to visualize the similarity rates between all the word lists obtained with different algorithms.

In the analysis section of the study, we deployed the algorithms and methods shown in Figure 2. These methods focus on words and aim to capture the salient topics of our datasets. Among these methods, n-grams are used to extract features from text-based databases and to classify text-based databases. N-grams (sequence of characters/words) are a preferred method due to their ease of use and success in small datasets, which reveal the distinguishing power of words. N-grams also do not require any additional features other than the tweet text and can be employed in broad perspectives like gender detection [31,32]. This method is mostly seen in the literature as unigram, bigram and trigram. N-grams extract simple statistics of some sequential word/character combinations and the detected words/characters based on these statistics are successfully utilized for text classification [33]. In our study, word frequencies utilized for descriptive and statistical analyses were determined by unigram and bigram method which are expressed as word-gram1 and word-gram2, respectively, in the following sections. The word lists obtained word-grams (word-gram1 and word-gram2) were compared with the word lists obtained by other methods.

Laureate et al. [34] examined 189 articles on topic modeling. They reported that LDA was used in 154 of the articles they reviewed. LDA, which is widely employed in topic modeling and gives successful results, works with a method similar to n-grams based on frequencies. In our study, LDA was utilized to obtain the best word lists for content detection of tweet texts and to compare them with other methods.

A third method we applied in the study is the Turkish BER Tweet [35] method. Turkish BER Tweet basically runs the same architecture as the BERT [36] model, takes into account semantic contexts, and operates by bringing similar words together. It is stated that Turkish BER Tweet can produce faster results for Turkish texts [35].

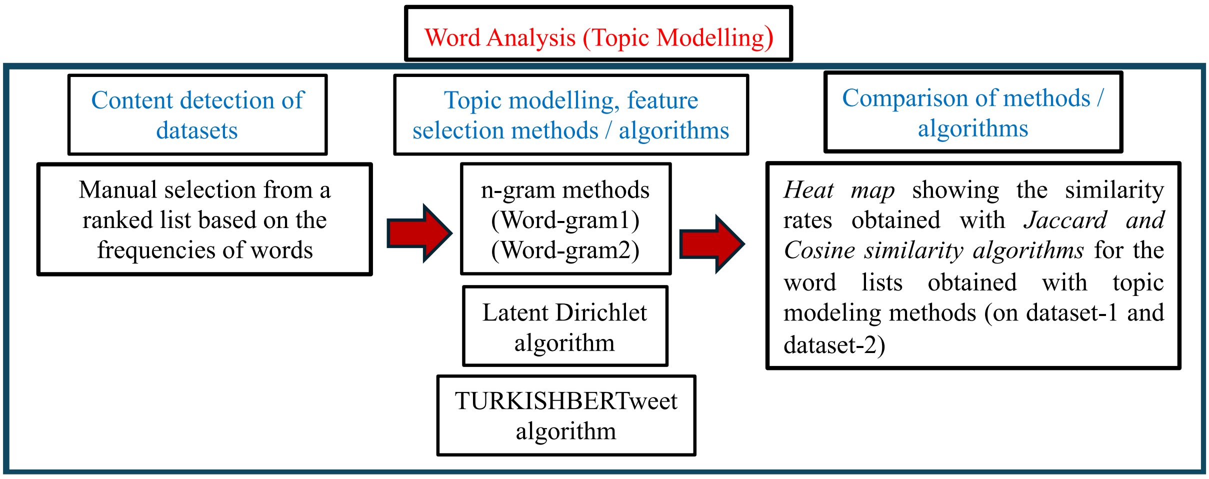

Fig. 2. Word analysis methods applied on the research

IV.RESULTS AND DISCUSSION

The dataset-1 and dataset-2 consists of 6 108 and 21 157 tweets respectively after removing null values. The statistical values of the dataset-1 and dataset-2 are shown at the graph in Figure 3. The percentage of tweets and RT's in dataset-1 is 41% and 59%. Additionally, the percentages of tweets containing mentions, hashtags, and URLs in dataset-1 are 70%, 0.2%, and 0.7%, respectively. The corresponding values for dataset-2 are 55% (tweets), 45% (RTs), 63% (mentions), 0.44 % (hashtags), 0.6% (URLs). Dataset-2 has a 14% higher percentage of tweets than dataset-1. This means that more new topics of conversation and new ideas were added to the social network chat environment after the pandemic measures were announced. Although the hashtag rate was low in both datasets, it doubled in dataset-2.

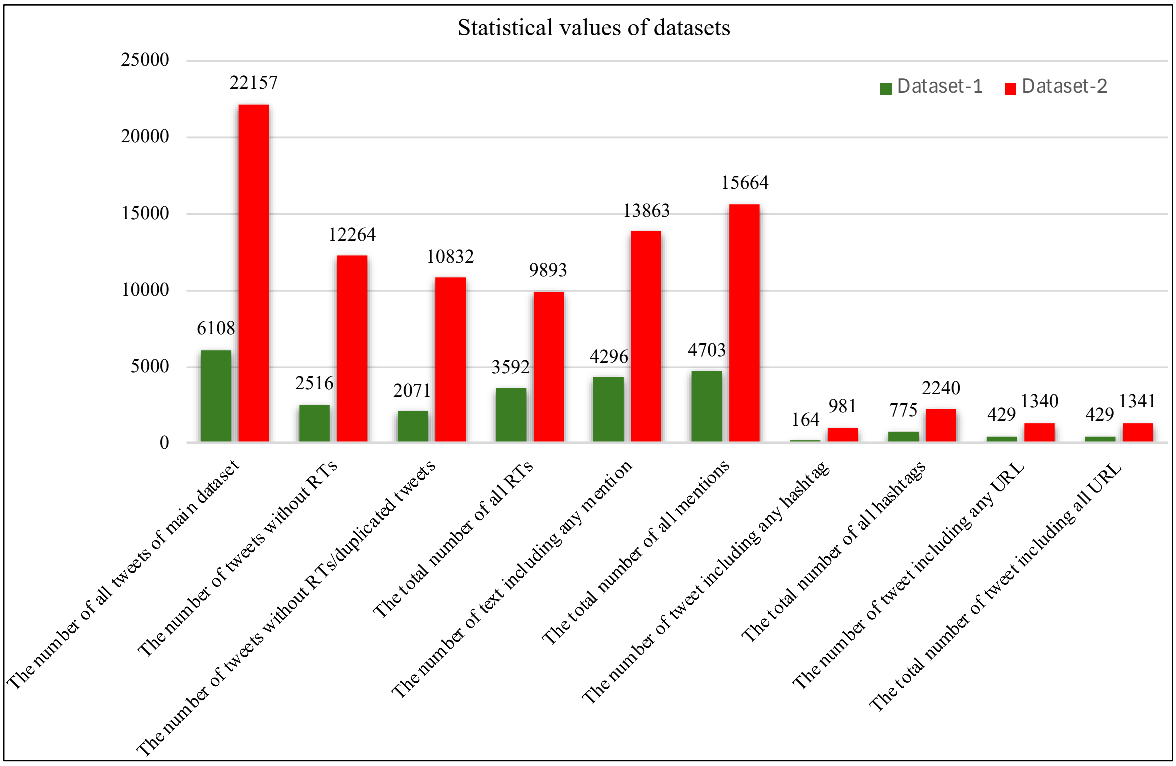

Fig. 3: Statistical values of datasets

The average values of dataset-1 and dataset-2 are close to each other as seen in Figure 4. The average number of characters in a tweet in both datasets is over 90. The average length of a tweet is approximately 14 words. Considering X's 280-character limit, we can say that the tweets in the datasets consist of short sentences. The 280-character tweet length limit in X makes it easier for users to express their ideas [10]. However, shorter texts attract more attention and receive more RTs from users. Such tweets stand out because they have more concise, to-the-point, more understandable and readable content [37,38]. As with other sectoral topics, tweet texts related to the clothing industry consist of short sentences [3]. Hashtag usage among X sers is very low. The average number of hashtags per tweet is 4.73 and 2.28 for both datasets. It was determined that the rate of hashtag usage is low among users chatting about jeans. The same applies to URL usage. Although the rate of hashtag usage has increased among all users in dataset-2, the number of words used as hashtags in a tweet in dataset-2 is less than in dataset-1.

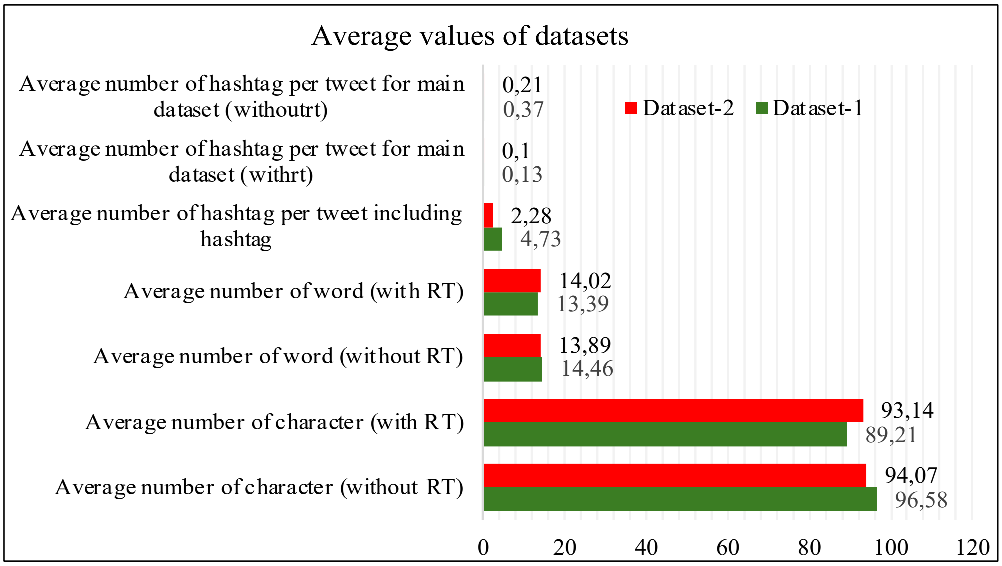

Fig. 4: Average values per tweet in datasets

4.1. Word and hashtag analysis

4.1.1. Word analysis

In this section, word analysis was applied on datasets with differents methods. These methods are manual selection of words as a heuristic approach, n-grams (word-grams), TurkishBERTtweet and LDA. The number of words to be included in the tables in this section was determined by taking into account the ratio of approximately 1/3 between the number of tweets in data set-1 and data set-2. The phrases "kot pantolon-jeans" used as keywords during the collection of data sets were not included in the tables. The first 40 words manually selected from the word lists in the datasets based on their frequencies are listed in Table 2. Table 2 is used as a basis for comparisons with the word lists in the following sections.

Tablo 2: Word list selected manually depending on frequencies of words

dataset-1	siyah-black, ceket-jacket, gömlek-shirt, giyen-wearer, dar-tight, mavi-blue, ayakkabı-shoes, tişört-t-shirt, kazak-sweater, güzel-beaitiful, kumaş-fabric, erkek-man, yırtık-tear, spor-sport, giyip-wearing, elbise-cloth, yeni-new, kız-girl, giydim-wore, bugün-today, bel-waist, gün-day, zaman-time, mont-coat, bol-loose, etek-skirt, beden-size, uzun-long, moda-fashion, açık-open, kombin-combination, deri-leather, gri-green, paça-trouser, aldım-got, renk-colour, boy-tall, tayt-leggings, giymiş-dressed
dataset-2	siyah-black, giyen-wearer, beyaz-white, tişört-t-shirt, gömlek-shirt, dar-tight, giydim-wore, ceket-jacket, giyip-wearing, ayakkabı-shoes, mavi-blue, giymek-wear, yaş-age, evde-at home, güzel-beaitiful, spor-sport, kumaş-fabric, bugün-today, gün-day, aletini-dick, erkek-man, zaman-time, etek-skirt, uzun-long, giymeyi-towear, kilo-weight, belli-particular, elbise-clothes, bol-baggy, yırtık-tear, tshirt-t-shirt, giymiş-dressed, kız-girl, şort-shorts, kesim-cutting, yeni-new, kadın-woman, tek-only, eşofman-sweatsuit, adam-man

Taking into account the agenda topics in Table 1, eight categories were identified for the pre-pandemic period. These categories are; color, other clothes for combination with jeans (such as sweaters, shirts), style (such as ripped, tight), expressions of apperance (such as beautiful), parts of jeans (such as waist, leg), type of goods (such as leather, fabric), action (such as wearing), gender. Four new categories were added to this list during pandemic period. These are; jeans model (such as boot-cut) model, sexuality, time, casual clothes (casual, sweatsuit etc.). The categories identified here can be used as label name for text classes in text classification and supervised/unsupervised machine learning studies for the clothing industry. The word lists above can be utilized as seed word lists for domain-specific and lexicon-based classification studies. On the other hand, this change in topics and categories can show businesses a direction to improve their product features, and these topics can also be used by businesses in advertising content posts to influence potential customers.

The word frequencies of the datasets determined by the n-grams method are given in Table 3 and Table 4 below. The main difference of Table 3 from Table 4 is that the calculation was made without removing RTs from the datasets. In contrast, the frequency values in Table 4 were obtained from datasets that do not include RTs. The purpose of making two separate calculations in this way is to determine whether RTs cause changes in terms of content in the datasets through similarity comparisons. The semantic analysis of the datasets yielded the following results regarding the content.

In Table 3, it is seen that 21 words out of 46 (yırtık-tear, kirli-dirty, samanlı-fodder, etek-skirt, elbise-cloths, giymek-wear, hanım-lady, deneme-try, kahve-cafe, zengin-rich, fakir-poor, beden-size, tensel-sensual, kumaş-fabric, boy-size, fiyat-price, siyah-black, kız-girl, makyaj-make-up, ürün-product, kodu-code) in data set-1 and 18 words out of 44 words (ayakkabı-shoe, converse, bez-fabric, maaş-salary, jean, kumaş-fabric, pantolon-pant, yaş-age, blue, giymek-wear, dar-tight, siyah-black, kilo-kilo, kesim-cutting, yakışıklı-handsome, tarz-style, gömlek-shirt, beyaz-white) in data set-2 that are related to the subject or can be a source of innovative ideas.

Table 4 shows that all but 4 (yok-absent, değil-not, olan-be, bugün-today) of the 24 words in dataset-1 and all but 18 (yok-absent, değil-not, evde-at home, bugün-today, gün-day, alet-tool, erkek-man, zaman-time, belli-particular, olan-be, tek-only, adam-guy, olur-be, Beylikdüzü, vardı-there was, tayt-tights, aynı-same, olsun-let it be of the 56 words in dataset-2 are directly related to the topic.

Tablo 3: The (word-gram1) words frequencies of datasets including RTs

Dataset-1 Frequencies higher tha 250			
Word-gram1	Frequencies	Word-gram1	Frequencies
yırtık	900	köy	422
belirtisi	891	tensel	418
kirli	775	hayaller	411
samanlı	768	orhanosmanoglu	396
etek	622	kumaş	395
elbise	621	boy	389
yere	613	fiyat	387
giyiyorum	589	göstergesi	356
giderken	580	olup	347
hanım	570	binen	343
deneyip	569	fakirliğin	338
bünye	568	siyah	337
müsait	568	kız	327
olmaya	567	yumurtası	326
hanımcık	564	kıymete	324
değill	560	eskiden	320

Dataset-2 Frequencies higher than 750			
Word-gram1	Frequencies	Word-gram1	Frequencies
adı	4362	getiren	2160
ayakkabı	2756	türkiyeye	2160
alan	2324	bizdeki	2143
converse	2323	sedaözen	2116
sovyetler	2311	giyen	1908
verip	2306	dar	1822
kırık	2305	aletini	1765
bez	2301	siyah	1418
dağılınca	2295	beylikdüzü	1229
kıçı	2293	belli	1216
maaşını	2293	kilo	1158
plastik	2293	kargo	1104
varşova	2293	kesim	1096
analizbabe	2283	yakışıklı	1082
pakt	2283	tarzda	1043
jean	2282	gömlek	1021

mizantrop ii	537	şimdi	320
olan	526	sweat	307
kahve	466	makyaj	305
zenginlik	446	gelen	285
şimdilerd e	445	içmeye	279
fakirlik	445	ürün	268
beden	436	kodu	262

kumaşı	2191	pasifim	991
pantolonu n	2189	uyan	978
yaş	2177	aktifler	970
biliyor	2170	arayışıma	970
dünyada	2169	arayışım	930
blue	2162	beyaz	890

Tablo 4: The (word-gram1) words frequencies of datasets not including RTs

Dataset-1 Frequencies higher than 40			
Word-gram1	Frequencies	Word-gram1	Frequencies
siyah	265	kız	54
ceket	150	giydim	53
beyaz	121	bugün	52
gömlek	119	bel	51
yok	98	zaman	50
giyen	97	gün	50
dar	96	etek	48
mavi	83	mont	48
tişört	79	bol	48
ayakkabı	79	beden	48
değil	76	uzun	46
kazak	75	moda	46
güzel	74	deri	44
kumaş	71	kombin	44
erkek	70	açık	44
yırtık	67	tane	44
spor	64	paça	40
giyip	59	gri	40
olan	57	boy	40
elbise	56	aldım	40
yeni	54	renk	40

Dataset-2 Frequencies higher than 150			
Word-gram1	Frequencies	Word-gram1	Frequencies
siyah	881	belli	232
giyen	699	olan	225
beyaz	645	elbise	224
tişört	644	bol	224
gömlek	642	yırtık	223
dar	615	tshirt	215
giydim	592	giymiş	211
ceket	508	kız	210
yok	472	şort	206
giyip	409	kesim	203
ayakkabı	374	yeni	201
mavi	366	kadın	197
giymek	344	tek	192
değil	343	eşofman	181
yaş	341	adam	180
evde	323	olur	178
güzel	318	beylikdüzü	176
spor	305	giyim	174
kumaş	301	vardı	173
bugün	298	yakışıklı	172
gün	291	beden	171
aletini	273	tayt	172
erkek	271	kazak	170
zaman	261	rahat	166
etek	247	aynı	165
giymeyi	240	açık	162
uzun	240	olsun	160
kilo	235	tarzda	159

After removing RTs from the dataset, changes in prominent topics and words emerged.

Table 4 reflects more sectoral characteristics than Table 3. From Table 3, it is seen that RTs have the effect of changing the direction of the content of the datasets. This result is also confirmed by comparing different topic modeling methods and similarity of datasets in the following paragraphs. Based on these results, datasets with RTs removed were used in all subsequent analyses. In sectoral studies such as this one, it is recommended to remove RTs from datasets in order to explore domain specific content in the content analysis of textual datasets created from the X platform.

The main topics of the data sets were identified by word-gram1 method and presented in a word cloud in Figure 5. Data set-2, which belongs to the Covid period, does not differ much from data set-1 in terms of content when we look at the wordclouds.

Data set-1 Data set-2

Fig. 5: The wordclouds of frequencies of single words (word-gram1) of datasets

In order to determine the characteristic features of the domain, the word-gram2 method was also applied and the words most identified with the word "jeans" and the word frequency lists are given in Table 5. In the binary word groups, words similar to those obtained with the word-gram1 method were detected and it was seen they were gathered around the keyword "jeans" used in data collection.

Table 5: Binary word group frequencies of datasets not including RTs

Dataset-1 Frequencies higher than 20			
Word-gram2	Frequencies	Word-gram2	Frequencies
kot ceket	77	düşük bel	29
pantolon giyen	61	pantolon siyah	28
pantolon kot	58	pantolon aldım	27
mavi kot	48	bel kot	26
yırtık kot	48	pantolon giyip	25
dar kot	42	pantolon tişört	24
pantolon beyaz	40	pantolon gömlek	23

Dataset-2 Frequencies higher than 60			
Word-gram2	Frequencies	Word-gram2	Frequencies
pantolon giydim	470	kilo yaş	141
kot ceket	259	giyen aktifler	141
siyah kot	255	beylikdüzü kilo	137
pantolon giyip	246	kumaş pantolon	136
pantolon giymek	244	pantolon giymiş	132
dar kot	235	pasifim arayışım	129
pantolon giymeyi	213	pantolon siyah	126

spor ayakkabı	37	deri ceket	23
siyah pantolon	36	pantolon kazak	23
pantolon giydim	36	pantolon giyiyorum	21
kumaş pantolon	33	pantolon giyiyor	21
gömlek kot	29	pantolon giymek	21

pantolon beyaz	187	beyaz gömlek	124
pantolon tişört	184	beyaz tişört	120
pantolon kot	182	uyan aletini	118
mavi kot	181	gömlek kot	101
kesim kot	176	pantolon giyince	97
dar kesim	170	siyah pantolon	95
evde kot	155	ceket kot	93
tarzda dar	154	pantolon giyiyorum	92
aletini belli	154	pantolon spor	91
yakışıklı aletini	154	aletini okşatacaklara	89
yaş yakışıklı	154	okşatacaklara harçlık	88
belli tarzda	154	harçlık verebilirim	88
spor ayakkabı	151	deri ceket	83
yırtık kot	147	defa kot	83
arayışıma uyan	146	mini etek	73
tişört kot	146	pantolon giyme	71
aktifler arayışıma	145	pantolon aldım	69
arayışım yaş	142	sıcakta kot	66
pantolon gömlek	142	pantolon üzerine	65
yaş pasifim	141	paça kot	62

In general, the results obtained for the dataset content using the word-gram method are as follows. The words reflecting the topics and characteristics of kot pantolon-jeans in dataset-1 are; siyah-black, beyaz-white, ceket-jacket, gömlek-shirt, ayakkabı-shoe, dar-tight, güzel-beaitiful, mavi-blue, kazak-sweater, elbise-clothes, tişört-tshirt, kumaş-fabric, spor-sport, giymek-wearing, yırtık-tear, mont-coat, moda-fashion, bel-waist, kombin-combination, paça-leg, renk-colour, kız-girl, tayt-leggings, bol-baggy, deri-leather. In addition to these words in dataset-1, the following words are also present in dataset-2: eşofman-tracksuit, kesim-cutting/fitting, şort-shorts, zaman-time, erkek-man, marka-brand, mini-mini, rahat-casual, kısa-short, cinsellik-sexuality.

LDA and TurkishBERTtweet algorithms were also used to determine the content of the datasets. The results obtained with these algorithms are given in Table 6 and Table 7, respectively. As can be seen from Table 6, the word similarities between dataset-1 and dataset-2 are high even though their rankings are different.

Table 6: Top 40 words of datasets excracted by LDA topic-modelling algorithm

Dataset-1				Dataset-2			
Wods	Frequencies	Words	Frequencies	Words	Frequencies	Words	Frequencies
siyah	255	yeni	54	siyah	881	gün	291
ceket	150	kız	54	giyen	699	aletini	273
beyaz	121	giydim	53	beyaz	645	erkek	271
gömlek	119	bugün	52	tişört	644	zaman	261
yok	98	bel	51	gömlek	642	etek	247
giyen	97	zaman	50	dar	615	uzun	240
dar	96	gün	50	giydim	592	giymeyi	240
mavi	83	bel	48	ceket	508	kilo	235
ayakkabı	79	beden	48	yok	472	belli	232
tişört	79	mont	48	giyip	409	olan	225
değil	76	etek	48	ayakkabı	374	bol	224
kazak	75	uzun	46	mavi	366	elbise	224
güzel	74	moda	46	giymek	344	yırtık	223
kumaş	71	tane	44	değil	343	tshirt	215
erkek	70	açık	44	yaş	341	giymiş	211
yırtık	67	kombin	44	evde	323	kız	210
spor	64	deri	44	güzel	318	şort	206
giyip	59	renk	40	spor	305	kesim	203
olan	57	gri	40	kumaş	301	yeni	201
elbise	56	aldım	40	bugün	298	kadın	197

When Table 7 is examined, it is determined that the word lists obtained from the data sets with the TurkishBERTweet method generally differ.

Table 7: Top 40 words of datasets exctracted by TurkishBERTtweet topic-modelling algorithm

Dataset-1				Dataset-2			
Word-gram1	Frequency Ratio	Word-gram1	Frequency Ratio	Word-gram1	Frequency Ratio	Word-gram1	Frequency Ratio
siyah	0.0174	topuklu	0.0114	twittercomdunkofkimst	0.0221	giyme	0.022
ceket	0.0126	kelebek	0.039	üniversiteye	0.0076	özler	0.0204
yok	0.0118	sweatshirt	0.0371	twittertelecomtrsta	0.0064	özlemişim	0.0152
kazak	0.0109	altın	0.0365	kızın	0.0059	unuttum	0.014

gömlek	0.0109	moda	0.0333
ayakkabı	0.0102	jeans	0.0325
dar	0.0101	denim	0.0292
giyen	0.0101	mesih	0.029
sıyah	0.0466	ceket	0.0252
gömlek	0.0233	gün	0.01336
ceket	0.0215	güneş	0.0359
beyaz	0.0202	gun	0.0256
tişört	0.0196	gözlüğü	0.0244
giydim	0.0168	tişört	0.0196
dar	0.0163	gömlek	0.0172
giymek	0.0	olan	0.0161
giyiyor	0.0121	siyah	0.0159
giymiş	0.0117	blog	0.062
yıkıldı	0.0116	pinterestcom	0.062
batı	0.0116	twittercomal	0.062

değil	0.0058	giymem	0.0117
yok	0.0057	gün	0.0702
kadın	0.0056	güneş	0.0162
fındık	0.0055	geçen	0.0144
giymek	0.0377	gözlüğü	0.0133
giymeyi	0.0261	giydim	0.0095
giyme	0.0184	günler	0.0086
giymeye	0.0165	gün	0.007
giymesi	0.0137	sabah	0.0061
özledim	0.0124	ceket	0.0215
istiyorum	0.0105	mavi	0.0208
giymesin	0.0099	okulun gözdesi	0.0127
giymesine	0.0096	neydi	0.012
giymeyi	0.0799	bel	0.0117
giymek	0.0592	hasanmese	0.0105
özledim	0.0402	tukenmis	0.0105

In order to determine the topic modeling method that reveals the content/main topics of the tweets with the highest similarity to the manually prepared word lists, the comparison results of these methods are shown on a heat map in Figure 6 and Figure 7. Jaccard and Cosine algorithms were used in the similarity comparison simultaneously in order to see whether there is an unusual error in the results obtained with the topic modeling algorithms. The similarity rates in Figure 6 and Figure 7 show that there is no unusual difference between the results of the two similarity algorithms. The results obtained here are discussed and compared from different perspectives.

Although both algorithms yield similar results, the average similarity value from the Cosine similarity algorithm is 0.1977 times higher than that obtained from the Jaccard algorithm.

In order to see how RTs affect the content detection of datasets, two groups of word lists belonging to datasets with RT, shown as Wordgram_1_Dataset1_with_RT and Wordgram_1_Dataset2_with_RT in the heat maps, were used in the comparisons. These two groups of word lists prepared using the Word-gram1 method have very low similarity rates in all comparisons. According to the Jaccard and Cosine similarity algorithms, the similarity rates of the two groups are 0.013 and 0.025, respectively. Since the comparison in question was made between two different periods, pre-corona and corona, such a low rate may be normal. However, when two datasets belonging to other groups without RT are compared, it is seen that the similarity rate is much higher. For example, the Jaccard and Cosine similarity rates of Manual dataset-1 and Manual dataset-2 are 0.48 and 0.65, respectively. These rates show how directly and strongly the inclusion or exclusion of RTs can affect the content of datasets.

When analyzed in terms of methods, the method that obtained the content of Dataset-1 closest to the manual method was the LDA algorithm with values of 0.82 and 0.9 according to the Jaccard and Cosine indexes, respectively. The method that obtained the content of Dataset-2 closest to the manual method was word-gram1 and LDA algorithm with values of 0.86 and 0.93 according to the Jaccard and Cosine indexes, respectively. In determining the content of the dataset using descriptive analysis features, LDA was found to be the most compatible algorithm with our study topic and purpose.

In terms of datasets, the similarity ratios between dataset-1 and dataset-2 were analyzed to determine how the emergence of an isolated lifestyle with Corona led to a change in users' conversations about this sector. When the manual processing results of dataset-1 and dataset-2 are compared, it is seen that the similarity ratios are 0.48 and 0.65 according to the jaccard and Cosine indexes, respectively. These similarity ratios also indicate that the difference between the two datasets is 0.52 or 0.35. When we look at the highest similarity ratio between dataset-1 and dataset-2 (which also shows us the lowest difference between them), we see that the similarity ratio between the two datasets determined by the word-gram1 method and the LDA method have the same values, which are 0.54 (Jaccard) and 0.7 (Cosine). These similarity results also tell us that the difference between the two datasets is 0.46 or 0.3. As a result, for the analyzed dataset, it can be said that there was a change of at least 0.3 in the content shared between users during the corona period compared to the pre-corona period.

It is observed that unexpected pandemics or disasters that affect daily life such as corona also cause changes in the interests of society in specific sectors. Such changes may be detected by businesses in the sector from social media platforms at the beginning of disaster. Obtaining these changes early and directly from potential customers will guide the product/service development. Businesses that provide early product/service developments in line with potential customer expectations will be able to gain an advantage over other businesses by directing the demands.

When examined in terms of topic modeling/feature extraction algorithms, in general, in all comparisons of dataset-1 and dataset-2, the similarity rates of word-gram1 and LDA methods are high. The highest similarity rate was obtained when these two algorithms were used on dataset-2. On the other hand, the similarity rate of TurkishBERTweet method with other methods and manually detected word lists is generally low. The most important reason for this difference is that the systems of LDA and n-grams methods are similar to each other and they are frequency based methods. TurkishBERTweet, unlike the other two algorithms, employs a different methodology that brings together the most similar words using their semantic meanings.

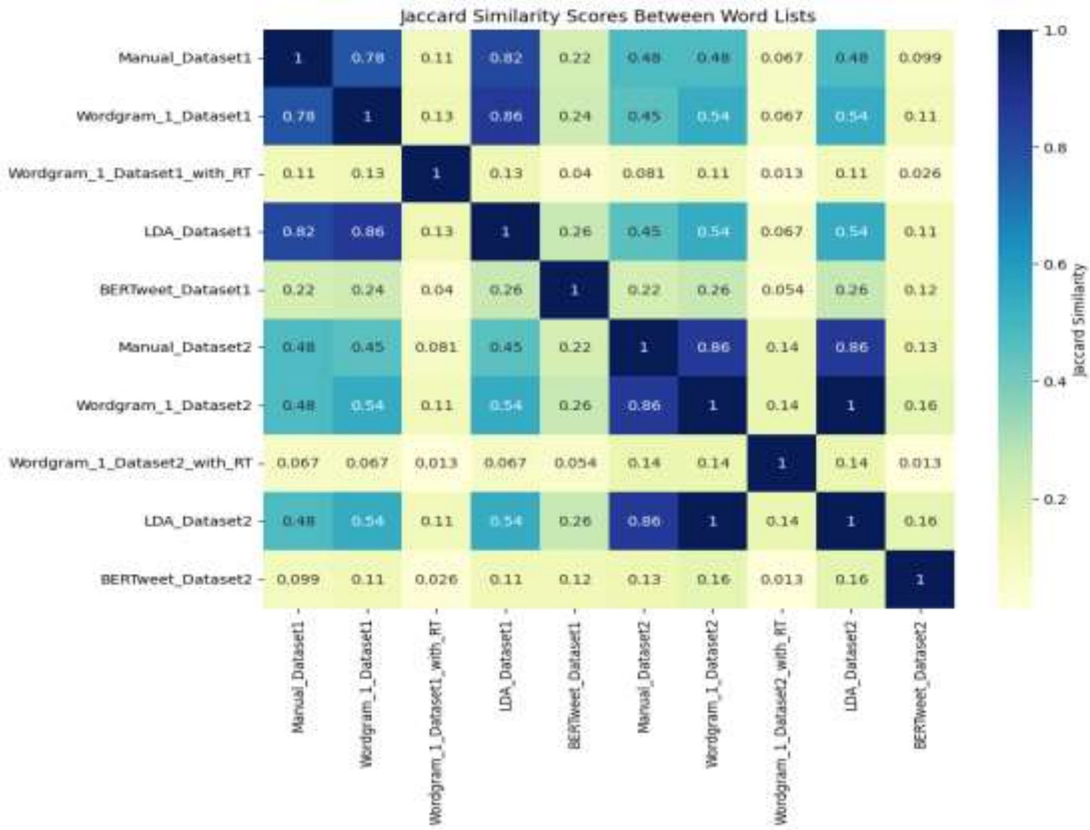

Fig. 6: Heatmap representation of Jaccard similarity scores of word lists extracted by different topic modeling algorithms

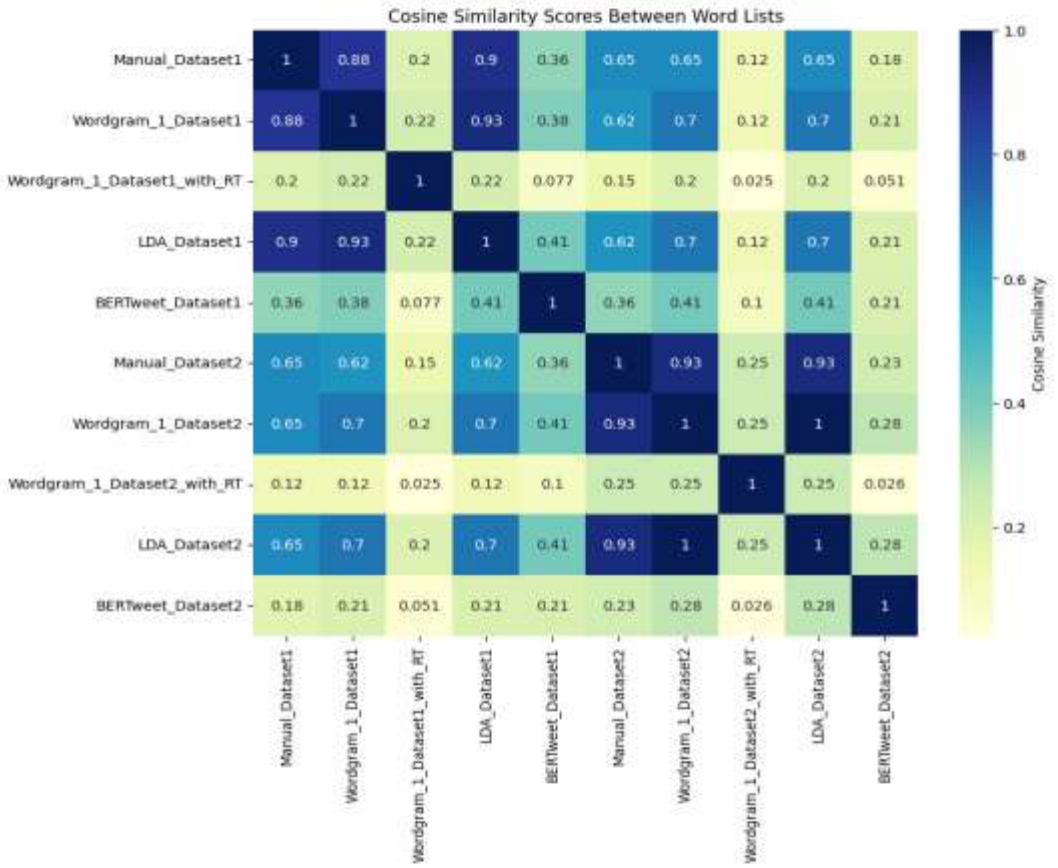

Fig. 7: Heatmap representation of Cosine similarity scores of word lists extracted by different topic modeling algorithms

4.1.2. Hashtag analysis

Figure 8 shows the hashtag words of the datasets on word clouds. Hashtag words generally do not include the features/content of jeans, which is the main topic, but only the words "kot-jeans, pantolon-pants, giyim-clothes, mavi-blue, alışveriş-shopping, kombin-combination, jeans, denim, fashion, moda-fashion, toptan-wholsale". In dataset-2, hashtag words reflecting topics related to students and mottos (#MaliyedenOnay60BnOgr#ApprovedByFinance60kStu, #EvdeKal-#StayHome, EvdeHayatVar-#LifeAtHome,#yokuniversiteleriacin-#opentheuniversities) related to the corona period come to the fore.

The hashtag words in Figure 8 do not reflect product features and they are only general expressions compared with the words obtained from word analysis. Morover Figure 4 shows that the hashtag usage rate is low for both datasets. When both information is considered together, it may be stated that querying, analyzing and content exploration of datasets based on only hashtag words will not produce sufficient results to reach information about user opinions.

Fig. 8: The wordcloud of hashtag of tweets of datasets

4.2. Users analysis

The average values of different users of the datasets are shown in Figure 9. Users tweeted approximately 1.2 times. This average decreases when RTs are removed from the datasets while dataset-2 has a higher average than dataset-1. Tweets and RTs show that communication about jeans among users during the corona period is higher than before the pandemic. More than half of different users have mentions in their tweets, while hashtags and URLs are close to zero for the same users. When this information is evaluated together with the information on the semantics of hashtag words in the datasets in the previous sections, hashtag words are insufficient to determine the content of the datasets. When we look at the average URLs in tweets that contain URLs, we see that there is only one URL per tweet.

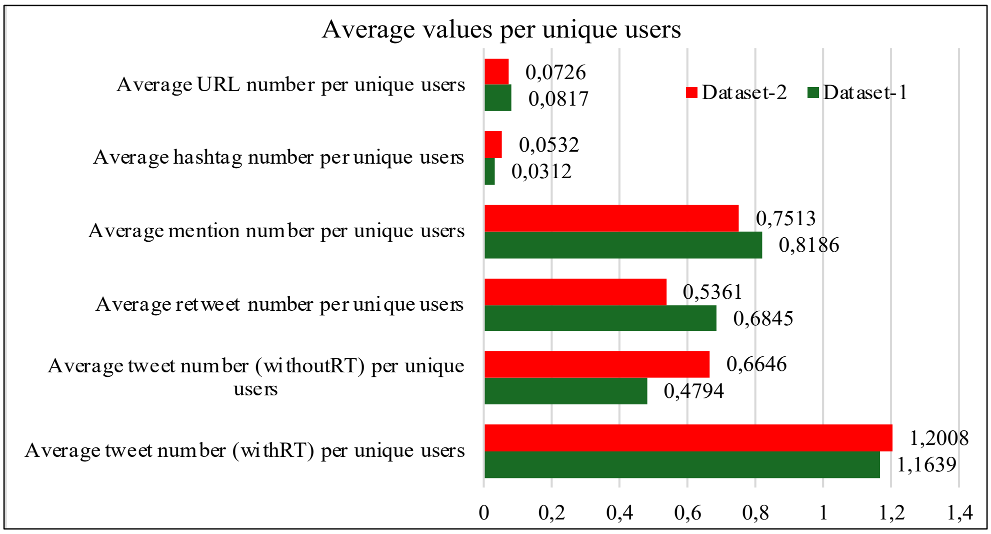

Fig. 9: Average values of tweets of datasets

The first 24 most influential users in the datasets according to the number of followers are shown in Table 8. It is seen that the users at the top of the datasets in terms of follower numbers have completely changed during the pandemic period (dataset-2), except for two (@duygusalterorm, @MrEgosuz, @), and the clear superiority of national news/broadcast channels in sharing has emerged.

Table 8: The most influential top 24 users of datasets

Dataset-1				Dataset-2			
Users	Followers	Users	Followers	Users	Followers	Users	Followers
@birdahabak	1046616	@Damlaazgin1	211210	@gazetesozcu	2534433	@kadrajmagazin	252193
@zekikayahan	632512	@tarihianlar	208422	@bilio muydunuz	1530134	@burak	194737
@Cokiyibence	406397	@tarihianlar	208419	@BirGun Gazetsei	1517660	@sporcope	192892
@Cokiyibence	406221	@wannartcom	196254	@t24comtr	1504827	@duygusalterorm	190211
@eskiseyler	379097	@KurtogluKagan	190700	@trtspor	1430601	@avukatergunn	179993
@erkekterimi	332798	@KadincaTweets	176333	@onediocom	667922	@MrEgosuz	169935
@erkekterimi	326516	@KelimePerileri	175862	@dayagiyedin	537700	@ MrEgosuz	169463
@pratikbilgi	305945	@ KelimePerileri	175550	@buzzspor	528303	@roleximyok	168984
@ pratikbilgi	305945	@MrEgosuz	171274	@futbolarena	495289	@kediefy	164557
@ pratikbilgi	305917	@ MrEgosuz	170394	@hikmetgenc	482992	@kenan kiran	157260
@AslihabElif	269625	@ MrEgosuz	170394	@eskiseyler	452128	@kenan kiran	157260
@agresifprofil	228001	@duygusalterorm	165929	@LutfuTurkkan	353540	@AcimasizTweets	150903

Note: The number of followers of X users may change over times. Users with the same name in the table are included in the list with different follower numbers due to the change in the number of followers in tweets posted on different dates.

The top 24 most influential users (based on chat starter tweet numbers of users) are shown in Table 9. When influential users are examined from this perspective, it is seen that national news/broadcast channels with high follower numbers are not included in Table 9. Seven users who were influential before the pandemic (@DenizEkrem, @pantolon dar, @hayalimserii, @umudumdan, @Umutbitmeyecek, @medyaglizencisi, @Capsvidorg) continued their activities during the pandemic, and apart from this, new influential users emerged with their chat starter tweets during the pandemic.

Table 9: The most influential top 24 users of datasets

Dataset-1				Dataset-2			
Handles	Tweet Frequencies	Handles	Tweet Frequencies	Handles	Tweet Frequencies	Handles	Tweet Frequencies

@hayalmiserii	9	@DenizEkrem4	3
@Capsvidorg	9	@OceansButik	3
@MobilyeZerk	9	@görmelisin	3
@baharcabutik	5	@birblognet	3
@AlpaslanNamlı	5	@hauntedysiv	3
@utuku27	4	@pantolondar	2
@Umutbitmeyecek	4	@medyaglizencisi	2
@canancli	4	@umudumdan	2
@warren marques	4	@Drjekyl09376201	2
@aptalamaesmer	3	@Ebruogretmenim	2
@patatesist	3	@unvillage1992	2
@sanasozulan	3	@evrenselbakis	2

@DenizEkrem4	159	@Capsvidorg	7
@modasarigiyim	26	@AnteplioglutT	7
@pantolondar	19	@gulergulk	6
@hayalmiserii	15	@AnailGuler	6
@NGamiotea	15	@squezelis	6
@umudumdan	14	@mochiminxq	6
@Umutbitmeyecek	12	@emilycanke	6
@ardaakgul2	11	@ArifAta28576710	5
@madyaglizencisi	10	@LousaGlenda1	5
@ibrahimduymaz21	10	@uurelik	5
@ANKARAAKYURT	9	@modaikracom	5
@esgulbutik	7	@forummaras	5

In order to reach real users and their opinions of pandemic period, the active user ranking given in Table 9 is considered to be more useful than the influential user ranking based on the number of followers. It is normal for news/broadcast channels to have higher follower numbers than other users. It is seen that news/broadcast channels that are not in dataset-1 are at the top of the ranking in dataset-2 in Table-8. The reason for this is news posts of these organizations increased with the pandemic and these posts include the keywords we used to create our datasets. These posts may affect potential customers (users) they reach. However, these posts' contents are not the opinions of potential customers/users that will be the source of product development [3]. The users who shape/affect chat environment seen in Table 9 are active users who start chat with new tweets (starting chat tweets) that allow user ideas to enter the chat environment. Sectoral businesses may shape/affect chat environment by primarily influencing these active users. For this reason, these active users can be chosen as the primary target users by businesses. In terms of our datasets, it would be appropriate to determine the seven users who continue to exist in both dataset-1 and dataset-2 in Table 9 as the primary target users by taking into account their continuity.

The frequency of mentions of users is showed in Table 10. Except for three users (@hayalmiserii, @Umutbitmeyecek, @roleximyok), it is seen that there are no users who can maintain their visibility in terms of mention frequencies both before and during the pandemic.

Table 10: The frequency of mentions of users of datasets

Dataset-1			
Users	Mentions	Users	Mentions

Dataset-2			
Users	Mentions	Users	Mentions

	Frequencies		Frequencies
@mizantropi	537	@roleximyok	64
@ORHANOSMANOGLU	396	@kalemdar	61
@hayalmiserii	236	@mehmettuluce	55
@izellyilmaz	204	@euthanasism	53
@ulvisaran	193	@BerdaVuran	51
@jafferson	164	@eskitvitler	50
@erkekterimi	161	@serkancelik1994	49
@KurtogluKagan	122	@UfukDemiray	49
@baharcabutik	117	@nakitbahislive	36
@Umutbitmeyecek	111	@yildirimhasret	36
@AslihanElif	102	@UgurBeyiniz	34
@duygusalterorm	84	@catwithaunicom	34

	Frequencies		Frequencies
@analizbabe	2283	@umudumdan	89
@Seda Ozen	2116	@roleximyok	84
@s3lcukluomer	576	@muuratipek	69
@hayalmiserii	437	@bilginizolsu	67
@Umutbitmeyecek	374	@milliveyerli25	66
@dayagiyedin	163	@Dunkofkim	64
@kenan kiran	150	@kediefy	63
@ aSLI	121	@Alicemyaa	59
@pantolon dar	120	@Adigeebey	54
@LutfuTurkkan	112	@TurkGayPlatform	53
@nimesay24948705	99	@EmreSungur 12	53
@gayarabul	94	@muharremkc	53

The top 24 most liked users in the datasets without RTs are shown in Table 11. It is seen that completely different users stand out in terms of users' likes before and during the pandemic. There is no similarity in likes between the pre-pandemic and pandemic periods.

Table 11: The most liked top 24 users in datasets without RTs

Dataset-1				Dataset-2			
Handles	Likes Frequencies	Handles	Likes Frequencies	Handles	Likes Frequencies	Handles	Likes Frequencies
@jafferson	4312	@uugurgülmez	253	@roleximyok	2825	@TC Aynasız	778
@catwithaunicom	3587	@yzbprice	244	@kadrajmagazin	2532	@Salihizimm	768
@harikasincupiya	1541	@Damlaazgin1	234	@Seda Ozen	1556	@canndari	764
@nakitbahislive	1032	@kagit kız	219	@sahiranevirane	1298	@sulisindunyası	745
@dionysosdiyorki	1002	@lamelamagic	195	@Adigeebey	1131	@de1vinm4llory	684
@wannartcom	814	@duygusalterorm	179	@LutfuTurkkan	1028	@necirvann12	665

@gulsennaydın	749	@irokrom	179
@faniyizfani	466	@hiya1986	163
@ahlaksiizim	369	@esmanurtilki	158
@yildirimhasret	291	@fahrenhiitt	146
@zekikayahan	270	@UgurBeyiniz	146
@derunehanim	264	@amudtakla	145

@kerimhandumann	975	@bidkbenkonuscam	660
@kubrabasknn	894	@peradakinoob	629
@avukatergunn	858	@benverayim	580
@bayanteror	810	@goldenclosettr	567
@atavratzlatan	808	@kerimhandumann	513
@buzzspor	786	@ aSLI	509

The statistical values of the likes are given in Table 12. The number of tweets in dataset-2 is about three times the number of tweets in dataset-1. The number of users receiving likes in dataset-2 is about four times higher than the number of users receiving likes in dataset-1. This shows that engagement on the topic increased strongly during the pandemic period.

Table 12: The statistical values of likes of datasets

	Dataset-1	Dataset-2
The total number of likes	31 128	111 596
The total number of users got any like	1 475	7 920
Average values of likes of users got any like	21	14

The average values of users are shown in Figure 10. The mean values of users are similar before and during the pandemic. Although the average values of tweets per user were similar in both periods, the number of original (chat starting) tweets and hashtag usage nearly doubled during the pandemic compared to the pre-pandemic period. The increase in active participation of users in the chat environment with original tweets during pandemic indicates that businesses may also need to increase their effectiveness in terms of effecting potential customers on such platforms.

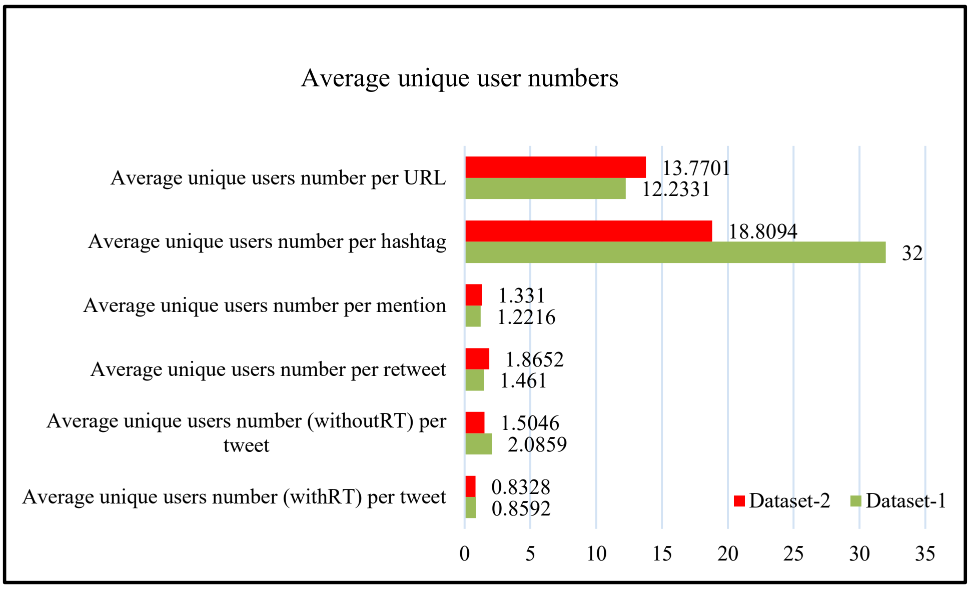

Fig. 10: The average values of users of datasets

4.3. Tweet analysis

The ratio of original/chat-starter tweets in dataset-1 and dataset-2 is 34% and 49%, respectively. There was a 15% increase in the number of original tweets during the pandemic period. Tweet-RT frequencies, tweets of the most influential users, the most liked tweets of the datasets were ranked and the top 25 tweets in the ranking were analyzed to determine the pre-pandemic and pandemic period topics and are given in Table 13.

Table 13: Topics related with jeans among users from top 25 lists

	The topics exctracted from top 25 of lists	
	Dataset-1	Dataset-2
Tweet-RT frequency list	Casual wearing, ripped denim jeans, wearing with pleasure by everyone, hip-hugging/fitted jeans, sexuality, drinking caffe with jeans, uncomfortable wearing, ministeries wearing jeans, sleeping on sofa with jeans, italian jeans with american sport shoes, shirts with jeans, conradiction of makeup and jeans, low waist jeans, sensual fabric, tassel jeans, hijab jeans	Converse shoes and jeans, blue jean, sick wearing shirt and jeans, handsome wearing skinny jeans, uncomfortable wearing, black dress and jeans, plain black t-shirt and plain black jeans, tunic with jeans, chic-casual and sensual jeans, ripped jeans, missing to wear jeans, drinking caffe with jeans
List of the most influential (depending on followers' numbers and starting chat	wearing with pleasure by everyone, leather jacket with jeans and black jeans, drinking caffe with sweat and jeans, beacuse of the discomfortable, cannot wear jeans, wearing jeans without panties, blue jeans with brown colour shoes, sexuality, Jeans style,	Sweater-stubble beard, black jeans-an inseparable part of the wardrobe, blue jeans, ladies wearing jeans and Istanbul Convention, sweatsuit is comfortable and uncomfortable wearing with shirt and jeans, jeans style, leather jacket with jeans and bandanna, handsome men with white linen shirt and jeans, suitable jeans colours for elderly people, stylish stonewashed jeans, very compatible with brown shirt-linen

tweets) users' tweets		jacket and jeans, missing shopping and buying jeans, pleasure with wearing jeans, jeans with flip-flops is ugly, plain black t-shirt and plain jeans, white shirt with jeans
List of most liked tweets	Sexuality, body suit-low waisted jeans, knee trackless jeans, jeans with shirt, prefering to wear sweatsuit, not wearing plaid shirt-jeans and patterned socks, ripped denim jeans, leather jacket with black jeans, can't find nice jeans, jeans better than fabric pant, leather jeans with dutch style, wearing jeans without panty, obsession of black tayt-black pant-black jeans, girls with white sweat-jeans and necklace, brown shoes and jeans, men with ripped jeans is ugly, light colour jeans with black sweat, blue jeans, black coat-white sweat-blue jeans, skinny/tight jeans on thin leg is unbearable, black jeans is suit for workers in store	plain black t-shirt and plain jeans and cheap/healtfull life, handsome men with white linen shirt and jeans, participating a wedding party with jeans and shirts, missing to wear jeans, ripped jeans, skinny/tight jeans and colourful shirts not suitable for men, missing shopping and buying jeans, black pant is better jeans, nijerian beatifuls with striped shirt combined with loose jeans, shirt with jeans is uncomfortable, sport shoes-jeans with equipment are wonderful, adolescent-jeans-high heels, white shir and jeans, combining t-shirt-jeans-tight instead of chic satin and dress, thinking girl wearing jeans is bad, missing to buy jeans, knee-trackless jeans, girls with crop top-jeans-miniskirt-sweatshirt-abusing, tight pants do not fit the hijab, muscular calf with ripped jeans is sexy, black dress-black skirt-bustier-jeans, gri sweat with jeans

The content of the most influential users' tweets and the most liked tweets shown in Table 13 overlap with the content listed in Table 2. However, Table 2 is more comprehensive and includes 8 categories before the pandemic and 4 additional categories for the pandemic period.

4.4. Time series analysis

The monthly frequencies of tweets are shown in Figure 11. In the pre-pandemic period, the highest frequency of conversation initiation tweets (red color) on X was in January, while in the pandemic period it was in April and June. The highest frequency of RTs (blue color) on X in the pre-pandemic period was February, while the highest frequency was May and August during the pandemic.

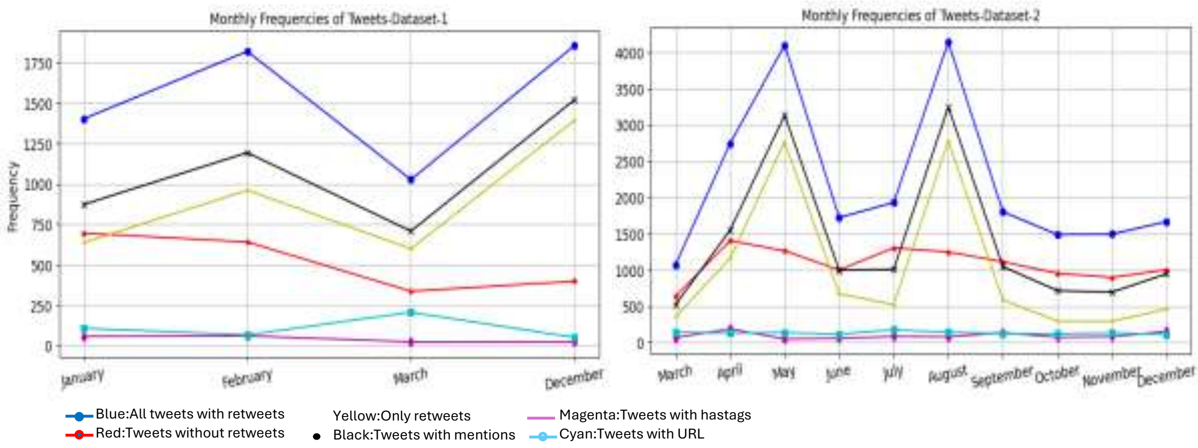

Fig. 11: Monthly frequencies of dataset tweets

Weekly frequencies of tweets are shown in Table 14. The highest frequency of the tweets is in the last week of the month in both the pandemic and pre-pandemic periods.

Tablo 14: The weekly frequencies of dataset tweets

	Dataset-1 Frequency	Dataset-2 Frequency
First week	1513	3995
Second week	1285	6549
Third week	1465	4154
Fourth week	1845	7459

Daily frequencies of tweets are shown in Figure 12. In the pre-pandemic period, two days of the month (21st and 26th days) had significantly higher engagement compared to other days. The highest values for the pandemic period are on the 9th and 28th days of the month. On these two days, although the interaction in terms of original tweets is similar, the number of interactions increases with retweets. Looking at the distribution of daily tweets, it is seen that the posts during the pandemic period have a more stable structure in general, that is, users who join the chat environment during this period are permanently present in the environment.

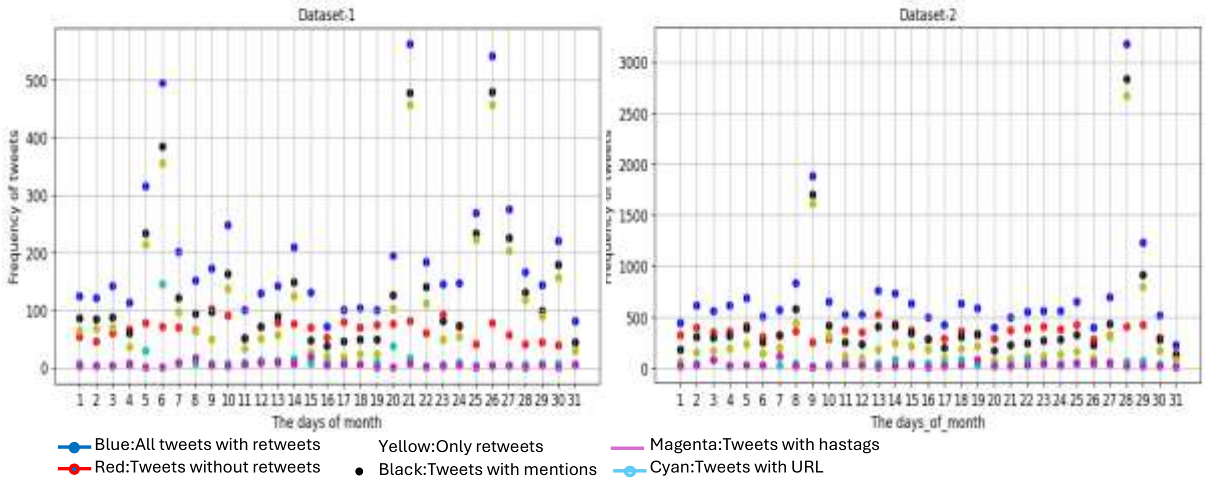

Fig. 12: Daily frequencies (according to the days of month) of dataset tweets

Daily (according to the days of week) frequencies of tweets are shown in Figure 13. It is seen that communication between users took place on Saturdays and Thursdays in the pre-pandemic period, and on Fridays and Saturdays during the pandemic period. In both periods, it was determined that the interest in the subject was more intense on Saturdays compared to other days, while the most intense interest was experienced on Fridays during the pandemic period.

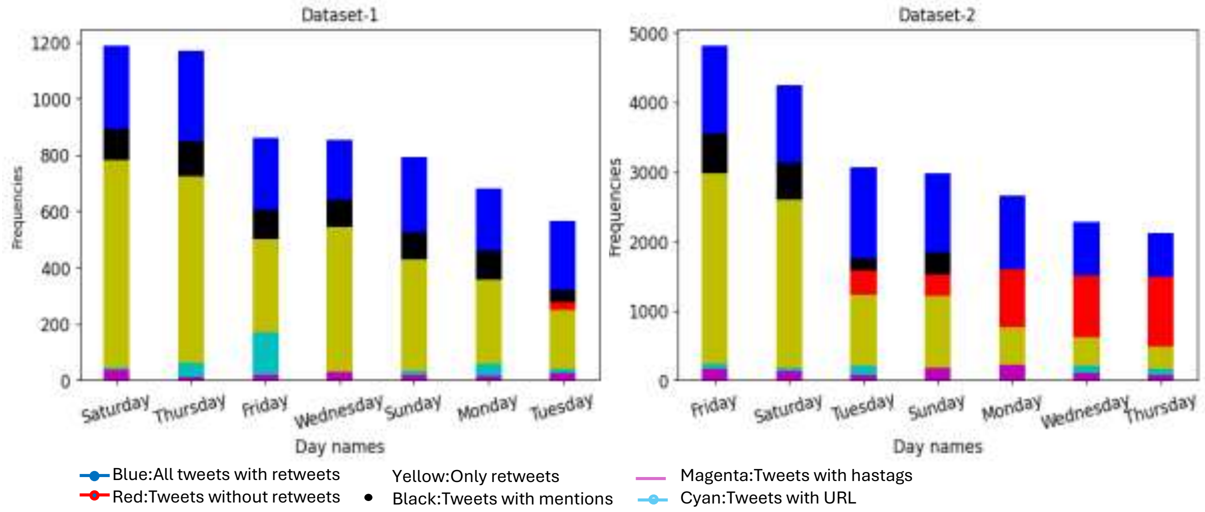

Fig. 13: Daily frequencies (according to the days of week) of dataset tweets

The hourly frequencies of tweets are shown in Figure 14. In terms of time zone, conversations intensify from 17:00 onwards in the pre-pandemic period and reach their highest level between 19:00-20:00. During the pandemic period, conversations are concentrated between 11:00-13:00 and 16:00-18:00. It is observed that conversations, which were concentrated in the evening hours before the pandemic, shifted to the day during the pandemic period.

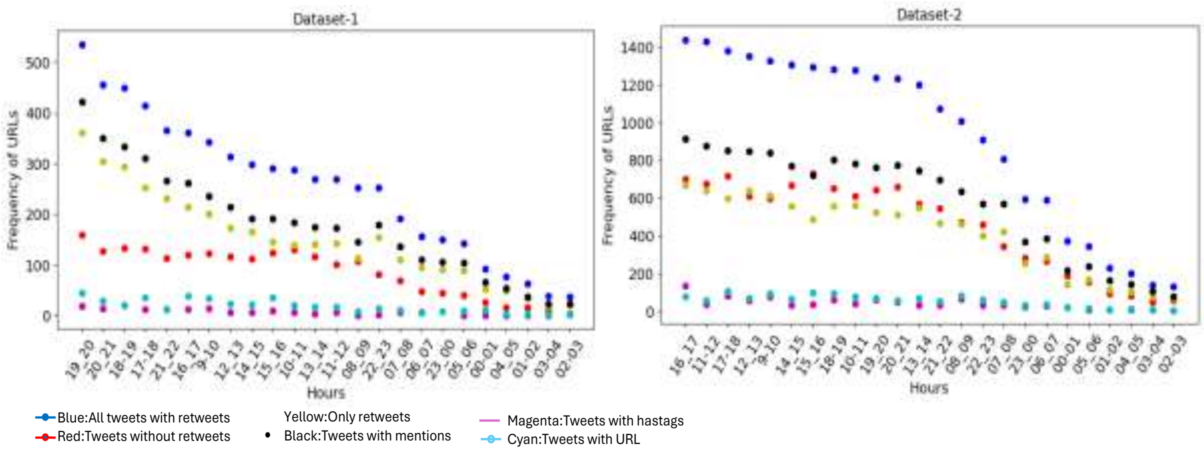

Fig. 14: The hourly frequencies of tweets and components of datasets

It is anticipated that the time periods mentioned above, when the conversations on the subject increase, may be reflected in the sales of the businesses. During these time periods, the tendencies of the customers in the conversation topics can be used by the businesses to direct potential customers in their social media advertisements and marketing policies.

V. Conclusion

Text mining studies of data from X have generally focused on analyzing political, topical/current issues that are on the agenda of society. The novelty of this study is to reveal the effects of the pandemic through the analysis of data obtained through the X platform for a specific product (jeans) in the clothing industry. To the best of our knowledge, this is the first study using X data in the Turkish literature in this sector. Analyses were conducted with statistical and descriptive analysis methods.

As the pandemic process has affected many areas of life, it has also been determined that it has caused changes in customer trends in the clothing sector. In the pre-pandemic period, it was observed that customers' interest in jeans was mostly on topics such as colors, combinations, style, appearance, parts of jeans, product type, action and gender. During the pandemic period, new topics such as jeans model, sexuality, time, and casual clothes were added to the list. Among the topic modeling methods, the unigram and LDA method gave the closest result to reality in determining the agenda topics. Although there was an increase in conversations and interactions on the topic during the pandemic period, it was observed that these conversations continued in a more static and consistent structure compared to the pre-pandemic period.

While the conversations changed by 30% in terms of topic, the intensity of these conversations shifted from Saturday to Friday and from evening hours to daytime during the pandemic period. In terms of the topic of our dataset, hashtag words are insufficient to gain insight into the content of the datasets.

The domain specific sectoral studies on these topics will provide resources for research based on keywords and word frequencies. We hope that this study will motivate researchers to conduct studies on specific sectors. Businesses in other sectors can also use similar, low-cost methods in the study for customer interests and product development to gain sectoral advantage.

In future studies, X analyses for other sectors of daily life, such as cargo, automotive, real estate or internet sales platforms, can shed light on these business areas by taking the sincere opinions of the Users. Furthermore, the word frequency lists of datasets can be divided into a certain number of slices and the ability of each slice to represent the entire dataset can be measured with different algorithms.

References

[1] Albalawi R., Yeap T.H., & Benyoucef M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers Artificial Intelligence* 3, 42. https://doi.org/10.3389/frai.2020.00042.

[2] Cano-Marin, E., Mora-Cantallops, M., & Sánchez-Alonso, S. (2023). Twitter as a predictive system: A systematic literature review. *Journal of Business Research*, 157 (2023) 113561. https://doi.org/10.1016/j.jbusres.2022.113561.

[3] Güneş, Y., & Arıkan, M. (2023). Exploring Twitter Dataset Content by Descriptive Analysis: An Application on Online Food Ordering. *Journal of Information Technologies*, 16(2), 119-133. https://doi.org/10.17671/gazibtd.1190184.

[4] Hampshire A., Hellyer P. J., Trender W., & Chamberlain S. R. (2021). Insights into the impact on daily life of the COVID-19 pandemic and effective coping strategies from free-text analysis of people's collective experiences. *Interface Focus.* 12, 11(6). https://doi.org/10.1098/rsfs.2021.0051.

[5] Xiang, P., Chen, L., Xu, F., Du, S., Liu, M., Zhang, Y., Tu, J., & Yin, X. (2024). The impact of the COVID-19 pandemic on nostalgic social media use. *Frontiers in Psychology, Sec. Media Psychology,* 15, 1-7. https://doi.org/10.3389/fpsyg.2024.1431184.

[6] Shahi, G. K., & Majchrzak, T. A.(2023). An Exploratory Study and Prevention Measures of Mob Lynchings: A Case Study of India. *Lecture Notes in Business Information Processing,* 103-118. https://doi.org.10.1007/978-3-031-43590-4_7.

[7] Mishra N., & Singh, A.A. (2016). The Use of Twitter Data for Waste Minimisation in Beef Supply Chain. *Annals of Operation Reearchs*, 270, 337-359. https://doi.org/10.1007/s10479-016-2303-4.

[8] Chae, B. K. (2015). Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics,* 165, 247-259. https://doi.org/10.1016/j.ijpe.2014.12.037.

[9] Bruns, A. & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, 16(2), 91-108. https://doi.org/10.1080/13645579.2012.756095.

[10] Boot, A.B., Tjong Kim Sang, E., Dijkstra, K., & Zwan, R. A. (2019). How character limit affects language usage in tweets. *Palgrave Commun* 5, 76. https://doi.org/10.1057/s41599-019-0280-3.

[11] Firdaus, S. N., Ding., C., & Sadeghian. A. (2018). Retweet: A popular information diffusion mechanism – A survey paper. *Online Social Networks and Media*, 6 (2018), 26-40, https://doi.org/10.1016/j.osnem.2018.04.001.

[12] Webberley, W.M., Allen, S.M., & Whitaker, R.M. (2016). Retweeting beyond expectation: Inferring interestingness in Twitter. *Computer Communications*, 73 (Part B), 229-235, https://doi.org/10.1016/j.comcom.2015.07.016

[13] Maria, G., Despoina, C., Neil, S., Christos, F., & Athena, V. (2015). Retweeting Activity on Twitter: Signs of Deception. *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science*, 9077, 122-134. https://doi.org/10.1007/978-3-319-18038-0_10.

[14] Alabid, N., & Naseer, Z. (2023). Summarizing twitter posts regarding COVID-19 based on n-grams. *Indonesian Journal of Electrical Engineering and Computer Science,* 31 (2). http://doi.org/10.11591/ijeecs.v31.i2.pp1008-1015.

[15] Bogdanowicz A., & Guan C. (2022). Dynamic topic modeling of twitter data during the COVID-19 pandemic. *PLoS One,* 27, 17(5). e0268669. http://doi.org/10.1371/journal.pone.0268669.

[16] Maja, B.P., Jasminka, D., Slobodan, B., & Ana, M. (2021). Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal. *Information Society Conference*. Ljubljana, Slovenia.

[17] Asgari-Chenaghlu, M., Feizi-Derakhshi, M.R., Farzinvash, L., Balafar, M.A., & Motamed, C. (2021). Topic Detection and Tracking Techniques on Twitter: A Systematic Review. *Collective Behavior Analysis and Graph Mining in Social Networks,* 4, 1-15. https://doi.org/10.1155/2021/8833084.

[18] Ordun C., Purushotham S., & Raff E. (2020). Exploratory Analysis of Covid-19 Tweets using Topic Modeling. UMAP and DiGraphs. https://doi.org/10.48550/arXiv.2005.03082

[19] Alamoodi, A., Mohammed, B., Albahri, O.s, Bilal, B., Zaidan, A., Wing-Kwong W., Salem, G., Albahri, A.s., Miguel, A.P., Ali, J., & Baqer, M. (2022). Public Sentiment Analysis and Topic Modeling Regarding COVID-19's Three Waves of Total Lockdown: A Case Study on Movement Control Order in Malaysia. *KSII Transactions on Internet and Information Systems,* 16, 2169-2190. https://doi.org/10.3837/tiis.2022.07.003.

[20] Figueiredo, S. (2024).Topic modelling and sentiment analysis during COVID-19 revealed emotions changes for public health. *Scientific Reports, 14 (1),* 24954. https://doi.org/10.1038/s41598-024-75209-3.

[21] Rajput, N.K., Grover, B.A., Rathi, V.K., & Bansal, R. (2024). Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic. arXiv:2004.03925v2 [cs.IR].

[22] Rahman, P.B., & Habibi, M.. (2020). Topic Analysis for Coronavirus Disease (COVID-19) on Twitter Using Latent Dirichlet Allocation (LDA). *The 1st Universitas Muhammadiyah Yogyakarta Undergraduate Conference (UMYGRACE) 2020*, 1. Yogyakarta.

[23] Choi, Y.H., Yoon, S., Xuan, B., Lee, S. Y., & Lee, K.H. (2021). Fashion informatics of the Big 4 Fashion Weeks using topic modeling and sentiment analysis. *Fashion and Textiles,* 8. https://doi.org/10.1186/s40691-021-00265-6.

[24] Habibabadi, S.K., & Delir Haghighi, P. (2019). Topic Modelling for Identification of Vaccine Reactions in Twitter. ACSW 2019. *Proceedings of the Australasian Computer Science Week Multiconference*, 1-10. https://doi.org/10.1145/3290688.3290735.

[25] Du, Y. (2021). A Deep Topical N-gram Model and Topic Discovery on COVID-19 News and Research Manuscripts. *Electronic Thesis and Dissertation Repository*. 7797. https://ir.lib.uwo.ca/etd/7797.

[26] Logan, A.P., LaCasse, P.M., & Lunday, B.J. (2023). Social network analysis of Twitter interactions: a directed multilayer network approach. *Social Network Analysis,* 13 (65). https://doi.org/10.1007/s13278-023-01063-2.

[27] Arpacı, I., Alshehabi, S., Al-Emran, M., Khasawneh, M., Mahariq, I, Abdeljawad, T., & Ella Hassanien, A. (2020). Analysis of Twitter Data Using Evolutionary Clustering during the COVID-19 Pandemic. *Computers, Materials & Continua (CMC),* 65(1),193-203. https://doi.org/10.32604/cmc.2020.011489.

[28] Hung, Y.H., Hwu, D.S., & Arkenson, C. (2015). Designing for retweets - A study on Twitter interface design focusing on retweetability. *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015*, *Procedia Manufacturing*, 3, 5496-5503. https://doi.org/10.1016/j.promfg.2015.07.699.

[29] Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *2010 43rd Hawaii International Conference on System Sciences, Honolulu, HI, USA,* 1-10. https://doi.org/10.1109/HICSS.2010.412.

[30] Shi J, Lai K. K., & Chen G. (2023). Examining retweeting behavior on social networking sites from the perspective of self-presentation. *PLoS One,* 18(5). e0286135. https://doi.org/10.1371/journal.pone.0286135.

[31] Nieuwenhuis, M., & Wilkens, (2018). Twitter text and image gender classification with a logistic regression n-gram model. *In Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018).* https://api.semanticscholar.org/CorpusID:51940390.

[32] Hassan, N., Wael, G., Ghada, K., & Mohammed, H. (2020). Credibility Detection in Twitter Using Word N-gram Analysis and Supervised Machine Learning Techniques. *International Journal of Intelligent Engineering and Systems,* 13, 291-300. https://doi.org/10.22266/ijies2020.0229.27.

[33] Joachims, T. (1998). Text categorization with support vector machine: Learning with many relavant features. *In Proceedings of the 10th European Conference on Machine Learning,* In: Nédellec, C., Rouveirol, C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science, 1398, 137-142. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0026683.

[34] Laureate, C.D.P., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review,* 56, 14223–14255. https://doi.org/10.1007/s10462-023-10471-x.

[35] Najafi, A., & Varol, O (2023). TurkishBERTweet: Fast and Reliable Large Language Model for Social Media Analysis. *Expert Systems Analysis*, 255 (Part C). https://doi.org/10.1016/j.eswa.2024.124737.

[36] Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: Pre-tarining of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1,* 4171–4186, Minneapolis, Minnesota, USA.

[37] Postus. (08.10.2024). Ideal Social Media Posts Length: Facebook, X, Twitter. https://postus.ai/blog/ideal-social-media-posts-length-facebook-x-twitter/#:~:text=The%20Ideal%20X%20Post%20Length&text=However%2C%20if%20you%20want%20to,they%20will%20ge t%20more%20engagement.

[38] Wylie, A. (08.10.2024). *What's the Ideal Length of a Tweet?* https://www.prnewsonline.com/whats-the-ideal-length-of-a-tweet/#:~:text=You'll%20get%20more%20retweets,engagement%2C%20accor ding%20to%20Buddy%20Media.