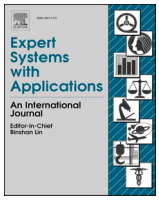
Expert Systems With Applications 252 (2024) 124175

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Occupational groups prediction in Turkish Twitter data by using machine learning algorithms with multinomial approach

Zeki Ciplak [a,1,*], Kazim Yildiz [b,2]

[a] Department of Computer Technologies, Gedik Vocational School, Istanbul Gedik University, Pendik, Istanbul, Turkey
[b] Department of Computer Engineering, Faculty of Technology, Marmara University, Maltepe, Istanbul, Turkey

ARTICLE INFO

Keywords:
Occupation prediction
Machine learning
Turkish twitter data analysis
Multinomial approach
Data mining

ABSTRACT

A lot of research has been done on personality and sentiment analysis, demographic and professional aspects using user shares in social networks. In particular, information extraction and value are produced based on Twitter data. This study aims to predict the users, occupational groups, who share in Turkish on Twitter, using machine learning methods. First, occupational groups and the Twitter accounts of the occupations in these occupational groups were determined manually and the tweets shared in these accounts were scraped. All tweets were then grouped by occupation into groups of one, five and ten, creating datasets with different characteristics, each containing more than 500,000 tweets. Some datasets were preprocessed using the Zemberek library, which is used in many Turkish NLP studies, and experiments were conducted out with a total 6 datasets. During the preprocessing phase, since the ready-made stopwords lists were not considered sufficient, unnecessary word lists consisting of single and binary words were created manually. Count and TF-IDF vectorizers are used to convert textual data into numerical. Since each word represents a variable in the text classification study, new variables were created by combining double and triple word phrases (ngrams) with feature extraction. In the experiments in which 24 different models were run, instead of using all the features created, the method of "determining the optimal number of features", which consists of the most valuable features, was used. It was found that the most successful model in the experiments using machine learning algorithms with a multinomial approach achieved 97.3% success in all calculated metrics.

1. Introduction

Occupation is an influential factor in many aspects of human life. A person's occupation affects first the way of thinking, then the personality, the speech and finally the writing. The status reflected in a person's occupation also affects the language that the person uses (Bernstein, 1960, 2003; Labov, 2006). Many research articles have been published showing the relationship between personality traits and occupation (Miller, 1962; Vernon, 1941).

The effect of educational level on language is an undeniable phenomenon. As people's level of education increases, the grammatical structure and vocabulary of the language they use improves. This development can also be seen in the messages shared on Twitter. People with more education tend to use the language more effectively by stricking to the rules of the language and avoiding meaningless abbreviations. There are other factors that influence language use and development. These include intelligence, gender, family, socioeconomic environment in which the individual grew up, parents' education level, bilingualism and health status (Uladı, Eryılmaz, Geyik, & Öztürk, 2019). When analysed as a gender factor, young girls generally respond more intensely to verbal stimuli, while young boys respond more intensely to visual stimuli (Temel, Bekir, & Yazıcı, 2014).

Twitter, one of the leading representatives of today's social media, is a technological platform where people reveal their whole selves, providing information about their personalities and lifestyles. Twitter gives its users the right to write a message called "tweet" of 280 characters. In addition, each user profile has a bio (biography) section with personal information and this section is 160 characters long (Twitter, 2023). Twitter users, use their accounts to express their opinions on almost every issue in the world.

* Corresponding author.
E-mail addresses: zeki.ciplak@gedik.edu.tr (Z. Ciplak), kazim.yildiz@marmara.edu.tr (K. Yildiz).
[1] https://orcid.org/0000-0002-0086-3223.
[2] https://orcid.org/0000-0001-6999-1410.

https://doi.org/10.1016/j.eswa.2024.124175
Received 28 January 2023; Received in revised form 22 February 2024; Accepted 6 May 2024
Available online 7 May 2024

Since a person's occupation covers a very large part of their life and a significant part of their day is spent at work, many of the tweets written on Twitter during this process may also contain information about that person's occupation. Given that there are millions of Twitter users, such a large amount of content would be suitable for use in many studies. Occupations can be inferred from the tweets they write, the biographies provided on their accounts, usernames, nicknames used as titles, and similar information that reflects the personalities of the users.

In the case of Turkey, according to Twitter is one of the most popular social networks. The report published by the communication office of the Presidency of the Republic of Turkey (Communications, 2022) shows this. In this sense, the Twitter social network contains sufficient and appropriate content for a job prediction study to be made over Turkish tweets.

In this study, Twitter users with certain occupational groups and tweeting in Turkish were identified, and then a large dataset was created with the shares of these users, an attempt to estimate the occupational groups of the relevant users. A literature review was conducted in related works. In the third part, information is given about the methods used in the research, the creation and processing of the dataset, feature selection, the classification algorithms and parameters used are given. The results of the study are presented and the obtained information is discussed with other research. Finally, the outcome of the work by interpreting the findings and possible future directions are given as a conclusion.

2. Related works

Preoţiuc-Pietro et al. (Preoţiuc-Pietro, Lampos, & Aletras, 2015) conducted a classification study for nine occupations using Twitter data. The nine selected occupations were taken from the system Standard Occupational Classification (SOC) system, which was developed by the UK government to classify occupations. A total of 18 features were created during the construction of feature set, such as the number of followers and friends, as well as tweets by users. As a result of the study, it was found that less than 20% of the users wrote their occupation directly in their account information and biographies. In many tests using Support Vector Machine (SVM), Logistic Regression (LR) and Gaussian Process (GP) algorithms, GP performed the best with an accuracy of 52.7%.

Tianran Hu et al. (Hu, Xiao, & Luo, 2016) revealed the characteristics of eight different professions (software engineer, manager, entrepreneur, designer, marketer, editor/author, public relations specialist, office clerk) in their occupation prediction study based on users' tweets. In addition to predicting occupations, the data collected was also used to analyse personalities. The study used a computer program called Linguistic Inquiry and Word Count (LIWC) for language analysis. The average F1 value of the predictions made for all occupations in the tests was found to be 0.78.

Kazi Zainab et al. (Zainab, Srivastava, & Mago, 2021) conducted a study to predict the health-related occupations of Twitter users working in the medical field. Using the bio information provided by users in their profiles, the study benefited from Word Embedding, Long Short-Term Memory (LSTM), Bidirectional LSTM (BLSTM), Gated Recurrent Unit (GRU), Bidirectional Encoder Representations from Transformers (BERT) and A lite BERT (ALBERT) algorithms. ALBERT was the best performing with an F1 score of 0.90. The accuracy, precision and recall scores 0.95, 0.92, and 0.91 respectively.

Jiaqi Pan et al. (Pan et al., 2019) carried out a new study using the dataset of Preoţiuc-Pietro et al.'s 2015 study. The 5191 accounts found in the previous study decreased to 4557 in the last period, due to the closure of some accounts and the setting of some accounts as private accounts. In the study, in addition to the tweets of these users, the users they followed by the users and the tweets of their followers were also included in the dataset. In the study using the Graph Convolutional Network (GCN) classifier, an accuracy rate of 61% was achieved which is a higher performance than in the study by Preoţiuc-Pietro et al.

Islam Mayda (Mayda, 2022) determined ten specific occupational groups in his job prediction study based on Turkish tweets. Five users were taken from each occupational group were taken when creating the dataset. For each user 500 tweets were collected. In this way, a total of 25,000 tweets with the occupation tag were included in the dataset. The highest accuracy performance of the study, in which LR and SVM algorithms were used, was obtained from the LR algorithm with 74.9%. Subsequently, by combining tweets in groups of 5 and 10, higher performances were obtained in the retests.

Shaojie Yan et al. (Yan, Zhao, & Deng, 2022) investigated the relationship between users' language use and their occupation by collecting data from Sina Weibo which is the largest microblogging site in China. The data of 20,452 active bloggers, who have published at least ten blog posts, are associated with 67 occupations. They were later reduced to six basic occupations. SVM, LR and Time Aware-Long Short Term Memory (T-LSTM) models are proposed for user occupation estimation. The best performance was achieved by the T-LSTM + LIWC method with an F1 score of 50.83% and an accuracy of 59.13%.

Shayan Vassef et al. (Vassef, Toosi, & Akhaee, 2022) created a dataset consisting of 1314 observations based on tweets and bio information of users on Twitter. Using TF-IDF Word Embedding and Deep Neural Network (DNN) techniques, they achieved an accuracy rate of 54% for professional titles in 9 categories.

Nikolaos Aletras et al. (Aletras & Chamberlain, 2018), estimate the socioeconomic characteristics, occupational class and income of Twitter users using information obtained from their extended networks. As a method, prediction models were developed using graph embeddings, which are low-dimensional vector representations of users. SVM-Graph + Topics, the best model for predicting occupational class, achieved 52% accuracy.

Margaret L. Kern et al. conducted a study (Kern, McCarthy, Chakrabarty, & Rizoiu, 2019) to predict ideal occupations for individuals based on linguistic information collected from social media. Using the linguistic analysis of 128,279 Twitter users, personality profiles were automatically evaluated and the personality profiles of different professions were visually mapped. This analysis showed that similar professions are grouped closely together. The model obtained using the Extreme Gradient Boosting (XGBoost) machine learning algorithm showed the best performance and showed that based on the individual's digital traces, their occupations can be predicted with more than 70% accuracy.

Joseph Michael O'Carroll (O'Carroll, 2023) investigated gender bias in occupation classification in Natural Language Processing (NLP) models and evaluated the effectiveness of various methods to reduce bias in semantic representations. In the study, five semantic representations and models of varying complexity were used to classify occupations from Twitter biographies. The study found that there was a trade-off between performance and bias reduction and that no bias reduction technique reduced gender bias in a statistically significant way. The study found that bias reduction efforts add noise that can affect overall model performance, but gender bias persists across different datasets and reduction techniques.

In the literature, there are also occupational, demographic, gender and personality analyses based on the profile information of Twitter users, the images they share and the tweets they write. For example, Jennifer Golbeck et al. (Golbeck, Robles, Edmondson, & Turner, 2011) a conducted personality analysis from Twitter. Zahra Riahi Samani et al. (Samani, Guntuku, Moghaddam, Preoţiuc-Pietro, & Ungar, 2018) also examined personality analysis on Twitter and Flickr images. Jacob Levy Abitbol et al. (Abitbol, Karsai, & Fleury, 2018) published a study based on socioeconomic status based on location, occupation, and semantics on Twitter. Priya Gaur et al. (Gaur, Vashistha, & Jha, 2023) performed sentiment analysis using a Naive-Bayes (NB) based machine learning algorithm on Twitter data. Miftahul Qorib et al. (Qorib, Oladunni, Denis, Ososanya, & Cotae, 2023) also conducted a sentiment analysis study on

COVID-19 hesitation using text mining and machine learning techniques on Twitter data. The five most recent studies mentioned here are visual or sentiment analysis based studies. As this study is entirely text-based and aimed at profession assessment, it was not necessary to go into the details of these studies.

3. Material and methodology

This section provides information on which occupations were included in the study, how the Twitter accounts of the occupations were identified, how the dataset was created, the preprocessing applied to the dataset, the feature selection process, the classification methods used and the parameter details of the machine learning algorithms. Fig. 1 summarizes of all the processes carried out during the study. The grey area in the figure was run separately for each model.

3.1. Determination of occupations and occupational groups

Since the main purpose of this study is to try to predict which occupational group users who tweet in Turkish belong to, it was preferred to create *Occupational Groups* with the occupations deemed suitable for the study before creating the dataset. While some occupational groups consist of several occupations, some occupational groups consist of only one occupation. Thus, a prediction made about an occupational group consisting of a single occupation was also a prediction made about the occupation itself. The occupations to be included in the occupational groups were determined manually. The basic list used for this purpose is the Turkish Dictionary of Occupations (İşkur, 2023) published by the Turkish Employment Administration (TEA-İŞKUR). For some of the lesser known professions in this dictionary, which defines nearly 10.000 occupations, the research didn't find any Twitter account with professional sharing. Moreover, it is not possible to find a Twitter account for every occupation in İŞKUR's dictionary of occupations.

The well known dictionary of İŞKUR was tried to be made as simple as possible with the methods described below. All occupations and occupational groups used in the study are shown in Table 1.

In determining the occupations to be added to the list of occupational groups, we also took into account how many people in Türkiye might be practicing that occupation and how many might have opened a Twitter account. For this purpose, before including an occupation is included in the list of occupational groups, it is checked whether there is sufficient and quality content on Twitter about this occupation. The control process involved searching both Twitter and Google separately for each

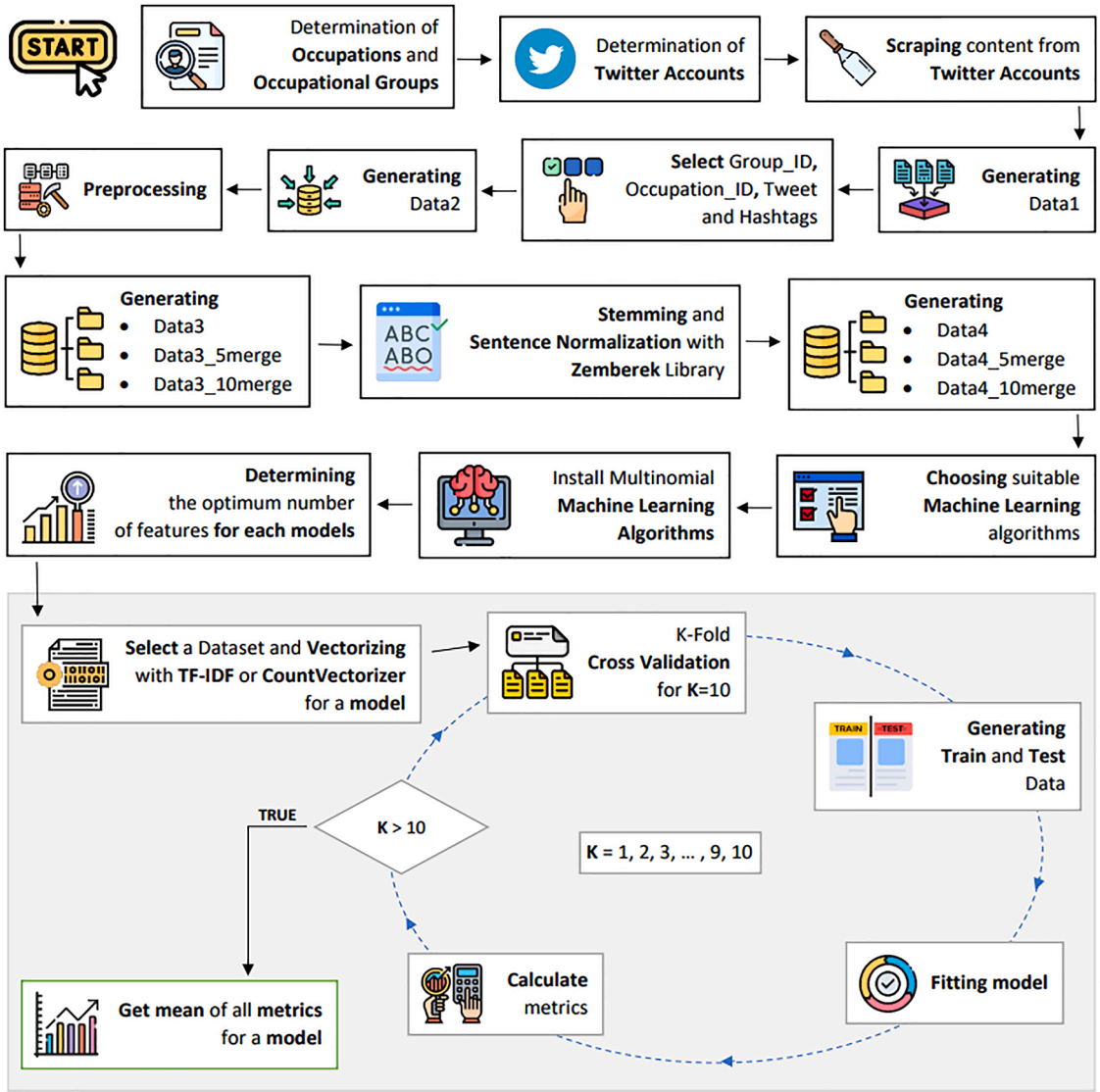


Fig. 1. Flowchart of the proposed study.

Table 1
Occupations and occupational groups.

Groups	Occupation(s)
0	Animator, Illusionist, Organizer, Clown
1	Archaeologist, Historian, Geographer
2	Chef, Gastronome, Food Engineer, Nutritionist
3	Basketball Player, Footballer, Volleyball Player, Coach
4	Computer Engineer, Software Engineer, Software Developer
5	Biologist, Bioengineer, Geneticist
6	Farmer, Agricultural Engineer
7	Diver, Swimmer
8	Dentist, Doctor, Nurse, Midwife, Pharmacist
9	Accountant, Stockbroker, Economist
10	Lawyer, Judge, Prosecutor
11	Physicist, Mathematician, Statistician, Chemist
12	Architect-Interior Architect, Civil Engineer
13	Occupational Health and Safety
14	Furniture Maker
15	Insurer
16	Fashion Designer
17	Geologist
18	Astrologer
19	Babysitter-Nanny
20	Barber-Hairdresser
21	Translator
22	Locksmith
23	Detective
24	Real Estate Agent
25	Photographer
26	Optician
27	Nuts Shop
28	Imam-Preacher
29	Psychologist
30	Jockey
31	Bookstore
32	Fireman
33	Courier
34	Jeweler
35	Vet

occupation. The Google search results on Twitter are important because they make it easier to find popular user accounts belonging to the relevant profession. In addition, when searching for a particular occupation, it was found that Twitter and Google searches often gave different results and highlighted different accounts. This has proven to be a valuable method in terms of reaching different content and users in the same occupation. Occupational accounts found in searches were stored for later analysis.

When creating occupational groups, occupations in similar fields have been included in the same group. The reason for this is that some occupations are closely related in nature and at the same time there are similar messages in the surveyed posts.

For example, users who are nurses or doctor, both work in hospitals and both are on duty. The professional messages they will share on Twitter will be very similar. From this point of view, it was considered reasonable to include these two professions in the same occupational group. Another example, the shares that people who are lawyers, judges or prosecutors will make on Twitter may be very similar. However, given that users in these occupations will share posts, that are mostly about court cases, legal terms, and anecdotes about complainants and defendants. It will be more difficult to distinguish known posts according to a particular occupation. Therefore, these three occupations are included in the same occupational group.

The fact that sub-branches of some professions have become professions in their own right and that is difficult to distinguish between these occupations on the basis of shared tweets alone, has led to the inclusion of many occupations with in similar area (such as *Biologist*, *Bioengineer* and *Geneticist* are in the same occupations). Some occupations were not included because they are rare encountered, and some of them included because there was no user who clearly stated in their profile that they were in that occupation. Some well-known occupations are not included in the list of occupation groups due to a lack of sufficient and quality content on Twitter.

By applying various criteria in this way, a list consisting of 65 different professions and 36 occupational groups in total was obtained. This list was taken into account in all the tweet messages to be collected on Twitter. Once the list was created, separate Twitter accounts were searched for each occupation on the list. After the occupational groups were identified, the Twitter accounts of the occupations were determimed.

3.2. Determination of Twitter accounts

The process of detecting Twitter accounts was carried out manually, as was the process of identifying occupations and occupational groups. As some Twitter user accounts were set aside during the identification of occupational groups, user searches on Twitter and Google were only carried out for occupations whose users were not identified at this stage.

Firstly, care was taken to have at least five user accounts for each occupation in the list of occupational groups. However, this number has been further reduced as no suitable accounts can be found for some occupations. On the other hand, although some occupations may have less than five accounts but they may have shared more than one profession with five accounts. In this case, it is not the total number of user accounts that is taken into account but how much a professional account shares. When identifiying the Twitter accounts of the professions, it was observed that there are accounts that users already follow and interact with. For example some users shared (*Retweet or Retweet via Quote*) the messages of other users in the same occupation, which also played a role in identifying these users.

This made it easier to access and share Twitter accounts related to a profession or occupational group. In addition, the accounts of users who stated in the bio section of their profile that they belonged to a certain profession, but who posted little or nothing about their profession on Twitter, were not included. In the latter case, a total of 304 Twitter accounts belonging to 65 professions were identified. All collected accounts are matched to the corresponding professions in the list of professions. The content scraping phase was then started from these accounts.

3.3. Scraping content from Twitter accounts

After determining which Twitter accounts to scrape content from, the process of collecting the last 5000 tweets shared by each determined Twitter user was started by using the Python library named *Snscrape* (JustAnotherArchivist, 2022). All tweets received from a user are tagged with the occupation of that account.

In addition to the tweets, we also collected a lot of metadata such as the time of the tweet was shared, the list of hashtags used in the tweet, the biography section of the user profile, the information about whether the user account was approved or not, the website shared by the user on their profile, and location data, if shared, were also collected. In the preprocessing phase, which will be explained later, some of these data were excluded from the main dataset as they were outside the scope of this study and would not be used in the machine learning algorithms.

It has been observed that the total number of tweets is less than 5000 in some Twitter accounts whose content is desired to be collected. However, due to the fact that the tweets shared by the user contain professional information and quality content, it was decided to keep the Twitter accounts of the relevant user in the list of occupational groups. The tweets collected from each user were recorded separately in CSV (*Comma Seperated Values*) files during the collection phase. When the collection of tweets from all users was complete, all the separate CSV files were merged into a single CSV file. The first dataset to be used in the study was created. Once the first dataset was created, the preprocessing phase began.

3.4. Generating datasets and preprocessing

The dataset, in which all other data belonging to a user, including tweets, were recorded in the rawest form, without any preprocessing, was called the *First dataset*. The first dataset contains two labelled columns in addition to the Twitter data. One is the information that specifying the occupational group, and the other is the information is the occupation itself. Each tweet in the first dataset is tagged with these two pieces of information. This results in a dataset that can be used to classify both of occupational groups and in the classification of individual occupations. The first dataset consists of a total of 770,189 rows and 17 columns.

The *Second dataset* is taken from the first dataset; it consists of columns containing occupational group information, occupational information, tweet and the list of hashtags used in the tweet. It is the same as the first dataset in terms of the number of rows. The second dataset is free of unnecessary features and contains more appropriate data for the purpose of the study. Most of the preprocessing was done on the second dataset. The first preprocessing on the second dataset was to remove the links in the tweets. Then the usernames that were mentioned in the tweet and were not related to the topic of the tweet were also removed from the tweets. Finally, the hashtags in the tweet were passed. Since the information about which hashtags are in each tweet is in a separate column in the second dataset. This information is used to remove the excess hashtags from the tweets with more than four hashtags.

This is to prevent spam tweets and to avoid situations that could disrupt the optimization of the dataset. The reason why such tweets are not completely removed is that some of them contain more hashtags but on the other hand, have quality content that gives information about the relevant profession. This was done in order not to lose the information that is already exists in the dataset and is suitable for use.

The preprocessing of the second dataset was continued by reducing all the characters. All characters except for a specific group of letters (*aâbcçdefgğhıiîjklmnoöprsştuüiûvyzwxq*) were removed from the tweets. Special characters such as *â*, *î*, *û* were later converted to their regular forms *a*, *i*, and *u*. In addition, space characters that were more than a single space were reduced to one and many invisible characters, such as *whitespace*, such as jumping to the bottom line, were removed from the tweets.

Many unnecessary messages such as memorials, celebrations, congratulations, follow-up requests and the like have been removed from tweets. The aim is to highlight professional topics in tweets rather than other topics. An important pre-processing in the second dataset is the removal of words called *stopwords*, usually prepositions, conjunctions, adverbs, and similar words from tweets.

Since the Turkish stopword list of the Natural Language Toolkit (NLTK) library (Bird, Klein, & Loper, 2009), which is widely used in NLP studies in the literature, was found to be insufficient, it was necessary to manually create a new stopword list. This list is called the *Redundant Words* (RW) list. For this purpose, a list was created by counting every single word and every two compound words in the second dataset were counted one by one and ordering them according to the number of repetitions, and trying to obtain a sufficient amount of RW. The process of obtaining RW was carried out manually and two lists were created. The first list is a list with a total of 729 elements, consisting of one-word words. The second one is a list of 1646 elements of two-word words. These lists were saved so that they can be used in future Turkish-based NLP projects. Then all RW from tweets were removed with the help of these lists.

Apart from redundant words, unnecessary syllables added to words were also removed from tweets. For example, since the quotation mark' in the sentence "*İstanbul'da kutlamalar devam ediyor*" (EN: *Celebrations continue in Istanbul*) was replaced with a space character in the preprocessing, the suffix "da" remained alone. This and similar suffixes do not have the aspect of specifying any profession by themselves. In this respect, the removal of all suffixes in this situation was carried out with lists of RW. Thus, with this post-processing and previous pre-processing, each tweet in the second dataset was optimized for a specific occupation. After all these processes, many tweets became messages of three words or less. These messages may consist of messages that do not provide information about an profession, such as "*merhaba, günaydın arkadaşlar*" (EN: *"Hello, good morning friends"*). These messages may consist of messages that do not provide information about any profession. For this purpose, tweets with less than four words were completely excluded from the dataset in order to avoid such situations as much as possible and to further increase the optimization. As a result of the preprocessing done so far, some tweets have turned into lines of empty data, and some tweets have become repetitive. At this point, tweets containing empty data and tweets that were identical to each other (*duplicate tweets*) were excluded from the dataset. In addition, all indices were rearranged, as all preprocessing has disrupted the order of the index numbers in the dataset. As the second dataset is very different from the first one and no further preprocessing is done; *The third dataset* was created by selecting the columns containing tweet, occupational group information and occupation information.

The third dataset is the most basic one in this study. It is and other datasets created from this dataset are the main sources of the machine learning algorithms used in the study. This dataset consists of 551,577 rows and five columns. The first five and the last five records of the third dataset are shown in Table 2. For a better understanding of the topic, English translations have been added under the Turkish tweets in Table 2. There are no English tweet translations of the tweets in the actual dataset.

The roots of the Turkish words in the tweets of the third dataset were found and replaced with their original form (*Stemming*). This dataset is also called the *Fourth dataset*. In the fourth dataset, the *Zemberek* library (Akın & Akın, 2007), which helps to find word roots and is often preferred in Turkish NLP studies, was used. The words for which the Zemberek library could not find the root of were removed from the relevant tweet. In this way, since some tweets contain less than 4 words of content, some tweets become null content, and some tweets become duplicate tweets and are not included in the fourth dataset. In short, approximately 12,000 tweets from the third dataset were not added to the fourth dataset. This reduced the number of records obtained in the fourth dataset decreased to 539,407. The order of some tweets in the third dataset has changed when the indices were reconstructed. Table 3 shows the first five and the last five records of the fourth dataset.

As can be seen from Table 2 and Table 3, the Zemberek library has made tweets expressible with shorter words. If the last tweet of the third and fourth datasets is examined, the tweet, which was included as "*sarıyer hayvan toplanıyor iddialarına yanıt…*" (EN: *"response to the allegations that sarıyer is collects animals…"*) in the third dataset, was transformed into a word consisting only of root words as "*sarıyer hayvan topla iddia yanıt…*" (EN: *"sarıyer collect animals claim response…"*) in the fourth dataset. Words that have the same meaning but different spellings in the third dataset are considered to be the same word in the fourth dataset. For example, the words *"iddialarına, iddiasına, iddiaları, iddiası…"* (EN: *"to their claims, to the claim, their claims, his/her claim…"*) and similar words were all replaced by the word "*iddia*" (EN: *"claim"*) alone.

The environment and cultural background in which an person grows up has a significant impact on language use. Nowadays, abbreviations of words and phrases, known as internet slang, are a common phenomenon on platforms such as Twitter. Such language use has been the subject of many academic studies (Barseghyan, 2013; Dixon, 2011; Manuel, Indukuri, & Krishna, 2010; Pasechnaya & Shcherbina, 2020). Although these abbreviations sometimes facilitate communication, they can sometimes cause confusion in meaning. They can also lead machine learning models being trained on misleading data. It is possible to see this kind of language use in different languages around the world. For example, many people in English prefer to write "*how r u*" instead of "*how are you*". This requires a more detailed analysis of the relevant

Table 2
A snippet from the third dataset.

Occupation_TR	Occupation_EN	Tweet	Group_ID	Occupation_ID
Organizatör	Organizer	mehteranlı sünnet organizasyonu…(EN: circumcision organization with mehteran…)	0	45
Organizatör	Organizer	asır organizasyon sünnet düğün…(EN: asır organization circumcision wedding…)	0	45
Organizatör	Organizer	yaz aylarında yapacağınız kır sünnet…(EN: rural circumcision that you will do in summer…)	0	45
Organizatör	Organizer	yeşil ailesinin kıymetli evlatları yiğit…(EN: precious sons of the green family, yiğit…)	0	45
Organizatör	Organizer	sünnet tahtı kiralamak ulaşınız istanbul… (EN: rent the throne of circumcision, reach istanbul…)	0	45
…	…	…	…	…
Veteriner	Vet	birleşmiş milletler tarım örgütü… (EN: united nations agricultural organization…)	35	51
Veteriner	Vet	burdur erzincan eskişehir istanbul… (EN: burdur erzincan eskişehir istanbul…)	35	51
Veteriner	Vet	moderatörlüğünü başkanımız prof… (EN: moderated by our president prof…)	35	51
Veteriner	Vet	prof kemal altunatmaz köpeklerde… (EN: prof kemal altunatmaz in dogs…)	35	51
Veteriner	Vet	sarıyer hayvan toplanıyor iddialarına yanıt… (EN: response to the allegations that sarıyer is collecting animals…)	35	51

Table 3
A snippet from the fourth dataset.

Occupation_TR	Occupation_EN	Tweet	Group_ID	Occupation_ID
Organizatör	Organizer	mehteran sünnet organizasyon…(EN: mehteran circumcision organization…)	0	45
Organizatör	Organizer	kuran mevlit semazen sünnet paket… (EN: quran mevlit whirling dervish sunnah package…)	0	45
Organizatör	Organizer	sünnet düğün organizasyon kalite…(EN: circumcision wedding organization quality…)	0	45
Organizatör	Organizer	eylül pazar gelin tuğba hanım…(EN: september sunday bride tuğba lady…)	0	45
Organizatör	Organizer	kına gece organizasyon iletişim…(EN: henna night organization contact…)	0	45
…	…	…	…	…
Veteriner	Vet	ibb veteriner hizmet müdür dünya…(EN: ibb veterinary service manager world…)	35	51
Veteriner	Vet	esenler mücavir alan sınır göktürk…(EN: esenler contiguous area border gokturk…)	35	51
Veteriner	Vet	iklim değişik vektörel larva mücadele…(EN: climate change vector larva fighting…)	35	51
Veteriner	Vet	bölge sahip güç düş hayvan zor kış…(EN: region having power dream animal hard winter…)	35	51
Veteriner	Vet	sarıyer hayvan topla iddia yanıt…(EN: sarıyer collect animals claim response…)	35	51

messages on Twitter and, if necessary, special handling in the pre-processing stage. In this research, word and sentence abbreviations are carefully analysed, especially those that occur under the influence of cultural background, are carefully analyzed. The Zemberek library for Turkish was used to correct expressions caused by regional or cultural differences or spelling mistakes. This approach prevented data loss and minimised the negative effects of incorrect data usage.

As a result, the fourth dataset has become a more optimized dataset in terms of meaning than the third dataset. In addition, if there are spelling errors in the tweets in the fourth dataset, these are also corrected to the *sentence normalization* feature of the Zemberek library. For example, a sentence such as *"yrn okula gidicem"* was adapted to the Turkish spelling rules as "*yarın okula gideceğim*" (EN: *"I will go to school tomorrow"*), and the errors in the words were corrected. So the, tweets have been optimized both in terms of meaning and spelling.

In the next step, the third and fourth datasets were regrouped with five and ten tweets from the same occupation, and two more datasets were created for each dataset. In the last case, a total of six datasets were obtained, and experiments were carried out with these six datasets. Table 4 shows the list of all datasets used in the study.

The effect of the datasets on the machine learning algorithms and the information obtained after running the algorithms is explained in detail in the results section. After creating the datasets to be used in the classification experiments were created, the classification algorithms to be used in the study were determined.

3.5. Choosing classification algorithms

In a text classification research, each word represents a feature, and the greater the number of features, the greater the runtime complexity of the machine learning algorithm (Ali, Khan, Anwar, & Asif, 2019). When each different word in the third dataset was counted, it was found that there were 414,438 different words. This number is also the total number of features. However, in a text classification study, it is not enough to identify the each word as a feature. It has been stated in some studies that counting double and triple phrases as a feature by itself contributes to obtaining more accurate results (Gaydhani, Doma, Kendre, & Bhagwat, 2018; Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016). In this study, when binary and triple word groups (*ngram_range*) were counted as variables in the same dataset, it was seen that the total number of features increased to around 8.7 million. For the fourth dataset, this number was found to be around 6.5 million. In order to be able to work properly with such large datasets, fast working machine learning algorithms were needed.

As the process of estimating occupational groups is a text classification process as a result, various experiments have been carried out with the algorithms most commonly in such studies. The aim of these experiments is to find the fastest possible solution. Machine learning algorithms called Multinomial Naive Bayes (MNB) (Losada & Azzopardi, 2008) and Multinomial Logistic Regression (MLR) (Böhning, 1992) were chosen as the algorithms that gave the fastest and best results in the

Table 4
Dataset contents prepared for the study.

Name of Dataset	CSV File	Explanation	Zemberek
Third Dataset	Data3	All rows contain 1 tweet.	0
	Data3_5merge	In all rows, 5 occupation-based tweets are combined.	
	Data3_10merge	In all rows, 10 occupation-based tweets are combined.	
Fourth Dataset	Data4	All lines contain 1 tweet.	1
	Data4_5merge	In all rows, 5 occupation-based tweets are combined.	
	Data4_10merge	In all rows, 10 occupation-based tweets are combined.	

experiments. The main reason for choosing these algorithms is that they can work quickly on large datasets. The implementation of these two algorithms was done using *Scikit-learn* (Pedregosa et al., 2011), a machine learning library written in the Python programming language.

When using of the MNB algorithm, the MultinomialNB class in the Scikit-learn library is used. The default values of the parameters in the class are used. MNB is widely used for large datasets because it is simple and efficient. Unlike other NB classification algorithms, it takes into account word frequencies in text documents. In cases where the number of words in text documents is important, MNB can produce more accurate results (McCallum & Nigam, 1998).

According to the Twitter users's tweet, the study's hypothesis was that the user's employment might be inferred from the Word frequencies used in the text. Upon converting a tweet to the word vector $\langle word_1, word_2, word_3, \cdots, word_m \rangle$, MNB tries Eq.(1) to ascertain the occupational group (Jiang, Wang, Li, & Zhang, 2016) that tweet belongs to, using Eq. (1).

$$c(Tweet) = \text{arg}max_{c \in C} P(c) \prod_{i=1}^{m} P(word_i|c)^{frequency_i} \quad (1)$$

The expression *c(Tweet)* is the occupational group prediction made by the MNB algorithm about the relevant tweet. *C* refers to all occupational groups. The *m* in the equation is based on all tweets; It is the number of different words consisting of single, double and triple phrases. These different words are also the number of features. Assigning different weights to the features can improve the performance of the MNB algorithm (Sucar & Sucar, 2021). $word_i$ is the i-th word in a *Tweet*. $frequency_i$ is the number of occurrences of the $word_i$ in a *Tweet*.P(c) is the probability that the *Tweet* is in class *c*. $P(word_i|c)$ represents the conditional probability of the $word_i$ when class *c* is known. For example, when the profession of doctor is known, the conditional probabilities of the words hospital and apple would be $P(hospital|doctor)$ and $P(apple|doctor)$.

Of these possibilities, the probability of $P(hospital|doctor)$ is higher than the other. P(c) and $P(word_i|c)$ are expressed by Laplace transform as in Eq. (2) and Eq. (3).

$$P(c) = \frac{\sum_{j=1}^{n} \delta(c_j, c) + 1}{n + s} \quad (2)$$

$$P(word_i|c) = \frac{\sum_{j=1}^{n} frequency_{ji} \delta(c_j, c) + 1}{\sum_{i=1}^{m} \sum_{j=1}^{n} frequency_{ji} \delta(c_j, c) + m} \quad (3)$$

n is the total number of *Tweet* s. This number varies according to the dataset used in the study. However, the total number of classes indicated by the *s* expression is constant in each scenario and its value was 36. c_j represents the labeled class value of the j-th *Tweet* in the training data. $frequency_{ji}$ is the number of occurrences of the $word_i$ in the j-th *Tweet* in the training data. $\delta(c_j, c)$ is a binary function and takes the value 1 if the values given in its parameter are equal to each other, and 0 otherwise.

MLR is also an LR algorithm that is often used in multi-text classification problems. The implementation of this algorithm uses the *LogisticRegression* class from the Scikit-learn library. The *solver* parameter of this class is used as *Newton-CG*. The Newton-CG method is effective in solving optimization problems in high-dimensional parameter spaces. Newton-CG is a combination of the Newton method and the Conjugate Gradient method. The Newton method is used to find the minimum value of a function, and the Conjugate Gradient method is used to find the point where the gradient of the function is zero. By combining of these two methods, the Newton-CG solver quickly solves the MLR problem involving a large number of variables (Fung, Tyrväinen, Ruthotto, & Haber, 2019; Genkin, Lewis, & Madigan, 2007; Shewchuk, 1994; Wright, 2006).

The *penalty* parameter in the *LogisticRegression* class is used to prevent overfitting of the model. The term L2 penalty is known as *Ridge Regularization*. When the Penalty parameter is set to L2, the coefficients of the model go towards zero and overfitting is prevented (Conroy & Sajda, 2012; Ghojogh & Crowley, 2019).

After selecting the classification algorithms to be used in the study, the total number of features obtained from the datasets was reduced. For this purpose, the *optimal number of features* was determined separately for each dataset and machine learning algorithm.

3.6. Determining the optimum number of features

The size of the total number of features was considered to be one of the main problems encountered in the study. Therefore, instead of using the millions of features mentioned above, it was decided to use the features (max_features) that are mentioned most often in the texts or have the most valuable weight. On the other hand, the experiments showed that using all the features obtained from the dataset led to worse results. In this sense, reducing the total number of features was considered important in terms of both to obtain more accurate results and to design applications that run faster. In determining the optimal number of features, the process of digitizing the textual information in the existing datasets was also applied.

For this purpose, vector space methods (Lee, Chuang, & Seamons, 1997) named Count and TF-IDF (Term Frequency-Inverse Document Frequency) were used. These two methods are included in the *feature extraction* section in Scikit-learn, a machine learning library, and have been used together in many text classification studies in recent years (Demir & Tepecik, 2022; Patel & Meehan, 2021; Raza, Butt, Latif, & Wahid, 2021).

In this study, vectorization operations were performed on Jupyter Notebook by using *CountVectorizer* (CVec) and *TfidfVectorizer* (TVec) classes in Scikit-learn. CVec and TVec digitize text data and aim to build a machine learning model using this numeric data.

CVec uses the word frequencies in the text document when converting a text data into numerical data. In short, it finds the number of occurrences of a word in a text. Each tweet in this study is counted as a text document. In this way, it measures the frequency of each word in the text document and converts these frequencies into a matrix. This matrix is used by machine learning algorithms.

TVec, on the other hand, assigns weights to the words in the text document when converting a text data to numerical data. This weight takes into account, the frequency of use of a word both in the local text document and in all other text documents is taken into account. If we follow this study, what TVec does is to consider the frequency of use of a word in each tweet and in all tweets together. In this way, the weight of valuable words will be higher, while the weight of meaningless words will be lower. These weights are used to create a matrix that is used by the machine learning algorithms. Table 5 shows the experimental values for the third dataset, created in 10 tweets based on occupation, and obtained when using the CVec and MNB algorithms.

CVec and TVec in the scikit-learn library are tools that work with

Table 5
Data3_10merge dataset with CVec and MNB algorithm determining the optimum number of features.

MaxFeatures	Accuracy	Precision	Recall	F1 Score
5000	0.916	0.931	0.916	0.919
15,000	0.936	0.945	0.936	0.937
25,000	0.944	0.950	0.944	0.945
35,000	0.949	0.954	0.949	0.949
45,000	0.952	0.957	0.952	0.953
55,000	0.954	0.958	0.954	0.954
65,000	0.956	0.960	0.956	0.956
75,000	0.956	0.960	0.956	0.956
85,000	0.956	0.960	0.956	0.956
95,000	0.956	0.960	0.956	0.956
105,000	**0.957**	**0.961**	**0.957**	**0.957**
115,000	0.956	0.960	0.956	0.956
125,000	0.956	0.960	0.956	0.955

similar parameters. In this study, the *max_features* and *ngram_range* parameters were used in both tools. The max_features parameter in CVec represents the most commonly used variables. In the machine learning algorithm, instead of using millions of variables, fewer variables are identified and used with max_features. The max_features parameters were tested separately for each dataset and algorithm. The presence of the optimal number of features was found where most of the evaluation metrics of the machine learning model (Accuracy, Precision, Recall, and F1 Score) started to decrease or stagnate together.

The values obtained in Table 5 show that it is appropriate to take the optimal number of features at around 70,000 for the Data3_10merge dataset. As this value is increased, the metrics do not change. A stagnation is observed up to 105,000. The increase in the metrics at 105,000 was not taken into account because it was only 0.001. For values after 105,000 there was a rapid return to the metric values of 70,000. To put it briefly, it has been shown that the success achieved with 125,000 features is also possible with 70,000 features. The graph of Table 5 can be seen in Fig. 2.

Similarly, the process of determining the number of specific optimal variables for each different scenario, such as using CVec or TVec, using or not using the Zemberek library, using one of the third or fourth datasets, has been done. Details of these models are given in the results section.

3.7. Building a machine learning model

All the steps after this point were repeated separately for each model. Firstly, the dataset to be used in the model was selected from the datasets created in the previous stages. Then the vectorisation tool to be used was determined. These tools were TVec or CVec, as previously mentioned. Both the ngram_range and max_features parameters are used. The ngram_range parameter is taken as (1, 3). This means that all the words in the texts are counted as separate features, first one at a time, then double and then triple word groups. In addition, the max_features parameter is predefined for each model, so it can be used out of the box in this tool. Then for each model the *K-Fold Cross-Validation* (KFCV) method (Allen, 1974; Stone, 1974; Zhang & Yang, 2015) was used to divide the dataset of the relevant model was divided into two as test and training data. KFCV is a method that divides the dataset into more than one piece and uses these pieces in different ways as training test data. This method ensures that the data in the dataset is more evenly split as training test data, so that the model can be trained and tested more accurately. For example, in a dataset divided into 10 pieces, the performance of the model can be evaluated by using each piece as test data once. In this way, all the data in the dataset is used once as test data once and the actual performance of the model can be predicted more accurately. In this study, the dataset used in each model was divided into 10 parts (K = 10) and 1 of them was used as test data each time and the corresponding model was trained. For such a model, 10 different training test data sets were generated and 10 different metric scores were obtained.

The averages of these metrics were then calculated and taken as the values of the model. The information and other details obtained from all the models created in the study are discussed in the results section. Fig. 3 shows the *pseudocode* of the model function used in this study.

There are several reasons for choosing the K value of 10. When the K value is chosen as 10, the dataset is divided into sufficiently large subsets. This makes it possible to create balanced samples for both training and validation. Increasing the K value too much would increase computations and the time required for these calculations. For example, if a K value of 100 is choosen, due to the high segmentation, training would be required for each segment, significantly increasing the computational burden. Choosing K = 10, on the other hand, reduces this intensive load.

From the perspective of variance, a smaller K value, such as K = 3 or K = 5, can induce a higher variance. This potentially leads to the model demonstrating different performances across various subsets.

In this study, when K was set to 10, it was observed that the variation in the metrics calculated for the subsets formed was minimal. Only a 0.05% change was observed in each iteration. Conversely, a larger K value could also result in a large bias. Considering these factors, the selected of K = 10 was considered to be a balanced choice. Many studies in the literature recommend choosing a K value of 5 or 10, further

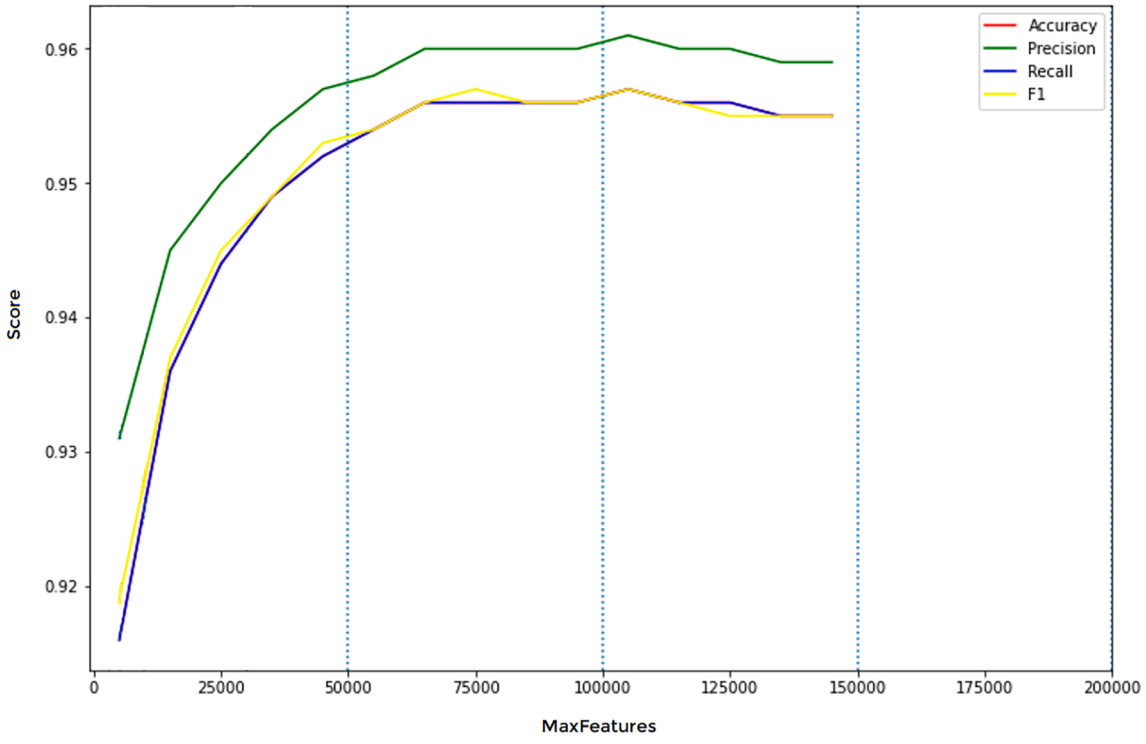


Fig. 2. Evaluation metrics graph according to different feature values using Data3_10merge dataset, CVec and MNB algorithms.

```
FUNCTION model (dataFile, vectorizer, model, splitKFold, optimum_features)
BEGIN FUNCTION

  SET df TO (READ dataFile)
  SET Xt TO df["Tweet"]
  SET y TO df["Group_ID"]

  IF vectorizer EQUALS "cv":
    SET vect TO (CALL CountVectorizer with max_features=optimum_features and ngram_range=(1,3))
  ELSE:
    SET vect TO (CALL TfidfVectorizer with max_features=optimum_features and ngram_range=(1,3))
  ENDIF

  SET X_vect TO (CALL vect.fit_transform with Xt)
  SET acc TO [], pre TO [], rec TO [], f1s TO []
  SET skf TO (CALL StratifiedKFold with n_splits=splitKFold and shuffle=True and random_state=42)

  FOR train_index and test_index IN (CALL skf.split with X_vect and y)
     SET X_train and X_test TO X_vect[train_index] and X_vect[test_index]
     SET y_train and y_test TO y[train_index] and y[test_index]

     CALL model.fit with X_train and y_train
     SET y_pred TO (CALL model.predict with X_test)

     SET accuracy TO (CALL accuracy_score with y_test and y_pred)
     SET precision TO (CALL precision_score with y_test and y_pred and average="weighted")
     SET recall TO (CALL recall_score with y_test and y_pred and average="weighted")
     SET f1 TO (CALL f1_score with y_test and y_pred and average="weighted")
     PRINT accuracy, precision, recall, f1

     SET acc[test_index] TO accuracy
     SET pre[test_index] TO precision
     SET rec[test_index] TO recall
     SET f1s[test_index] TO f1
  ENDFOR

  PRINT round(mean(acc), 3))
  PRINT round(mean(pre), 3))
  PRINT round(mean(rec), 3))
  PRINT round(mean(f1s), 3))

  RETURN model, vect

END FUNCTION
```

Fig. 3. Model function's pseudo code.

establishing this value as a standard practice (Anguita, Ghelardoni, Ghio, Oneto, & Ridella, 2012; Marcot & Hanea, 2021; Nti, Nyarko-Boateng, & Aning, 2021).

4. Results and discussion

In this study, a total of 24 models were created from two separate machine learning algorithms, two separate vectorizing tools, six different datasets with or without Zemberek library. The KFCV method was applied on all datasets. Information on the models created, the tools used and the metric values obtained are shown in Table 6.

The most successful model in the study is model 18, shown in Table 6. The model built on the dataset using the MLR algorithm, the CVec vectorizer and the Zemberek library achieved a success rate of 0.973 in all metrics with only 3,500 features. The MLR algorithm was found to be more successful than the MNB algorithm, depending on many factors. However, although the MLR algorithm was successful in terms of results, it took longer to train than the MNB algorithm. In addition, it was found that the models created with the datasets where the tweets are in groups of ten were more successful than the others. The numbers of these models in Table 6 are respectively; 3, 6, 9, 12, 15, 18, 21 and 24.

These models, like the most successful model, produced more successful results with fewer variables than the others. In this sense, it was an important step to improve the results by aggregating together the tweets. The success of the model that could not be achieved with hundreds of thousands of features, could be achieved with fewer features by using tweet groups of 5 and 10. Accordingly, obtaining 10 shares of a Twitter account will be sufficient to detect the occupational group information of that account with a probability of 97.3%, after going through the appropriate pre-processesing.

When evaluating the vectorizing tools, it was found that the CVec vectorizer gave better results than TVec vectorizer in all models. It was found that the use of the Zemberek library had an effect on increasing success when all models were considered.

One of the main reasons for the 97.3% success rate of the study is the meticulousness of the preprocessing stage. Especially for the Turkish language, the stopword lists published in many places were not considered sufficient, so creating a list of redundant words specific to this study and removing these redundant words from the tweets in the datasets ensured that each tweet share was further optimized for professional content.

The sentence normalization feature of the Zemberek librar which replaces the incomplete or misspelled words with the correct ones, also enabled the information in the datasets to be better evaluated by machine learning algorithms. The fact that similar professions are grouped together can be seen as another reason for this success. As the number of occupations increased, it was deemed appropriate to include similar

Table 6
Metric values were calculated for 24 models that were created using varying combinations (K = 10).

No	Model	Vectorizer	Zemberek Library	Tweet Groups	Optimum Features	Accuracy Score	Precision Score	Recall Score	F1 Score
1	MNB	CVec	0	1	250,000	0.770	0.787	0.770	0.770
2				5	100,000	0.931	0.937	0.931	0.931
3				10	**70,000**	**0.956**	**0.960**	**0.956**	**0.956**
4			1	1	200,000	0.747	0.768	0.747	0.747
5				5	50,000	0.927	0.933	0.927	0.928
6				10	**40,000**	**0.958**	**0.961**	**0.958**	**0.958**
7		TVec	0	1	60,000	0.731	0.767	0.731	0.725
8				5	11,000	0.892	0.906	0.892	0.889
9				10	**6000**	**0.924**	**0.932**	**0.924**	**0.921**
10			1	1	30,000	0.715	0.750	0.715	0.710
11				5	7000	0.902	0.913	0.902	0.900
12				10	**3000**	**0.940**	**0.945**	**0.940**	**0.939**
13	MLR	CVec	0	1	310,000	0.774	0.780	0.774	0.773
14				5	160,000	0.945	0.946	0.945	0.944
15				10	**30,000**	**0.971**	**0.971**	**0.971**	**0.971**
16			1	1	33,000	0.735	0.740	0.735	0.733
17				5	11,000	0.947	0.948	0.947	0.947
18				10	**3500**	**0.973**	**0.973**	**0.973**	**0.973**
19		TVec	0	1	31,000	0.746	0.758	0.746	0.745
20				5	19,000	0.933	0.935	0.933	0.932
21				10	**13,000**	**0.962**	**0.963**	**0.962**	**0.961**
22			1	1	130,000	0.765	0.772	0.765	0.763
23				5	4000	0.934	0.936	0.934	0.934
24				10	**3000**	**0.962**	**0.963**	**0.962**	**0.962**

occupations in the same group so that the success rate did not decrease.

In addition, as a result of the using of the KFCV method, the close metric values obtained in each iteration proved that the preprocessed data used in the study were quite robust. An example of this is shown in the table below. Table 7 shows the metrics calculated in the KFCV iterations of the most successful model in the study and the average metrics obtained as a result are given together.

When Table 7 is examined, although a different test data was used in each iteration of the KFCV process, the change in the metric values obtained was at most five per thousand. To summarize briefly; It has been shown that no matter what data is selected for testing from the relevant dataset, the success of the model does not change significantly. Achieving an average accuracy value of 0.973 means that, for any K value between 1 and 10, the metric consistently has a value greater than 0.97. The primary objective of this calculation is to highlight that the accuracy rate above 0.97 is maintained regardless of changes in the K value, highlighting the minimal in the metric even when K is 5 or 8.

To enhance the comprehension of the subject, additional experiments were carried out after the study, evaluating different K values to determine new average accuracy rates. In these experiments, average accuracy was recalculated for K values of 5, 10, 15, 20, 25, and 30. In particular, there was a noticeable decrease in accuracy for all models where K value was less than 10.

For models with a K value greater than 10, an increase in the accuracy rate was observed in certain instances. However, this increase was marginal, not exceeding 0.1%. Throughout the study, only six models

Table 7
Comparison of the metric values of model number 18 obtained in KFCV iterations with the average metric values.

Iteration	Accuracy	Precision	Recall	F1
1	0.974	0.975	0.974	0.974
2	0.972	0.972	0.972	0.972
3	0.972	0.972	0.972	0.972
4	0.971	0.971	0.971	0.971
5	0.975	0.975	0.975	0.975
6	0.974	0.974	0.974	0.974
7	0.975	0.975	0.975	0.975
8	0.975	0.975	0.975	0.975
9	0.972	0.972	0.972	0.972
10	0.970	0.971	0.970	0.970
Mean	**0.973**	**0.973**	**0.973**	**0.973**

Table 8
The average accuracy rates of the models that achieve accuracy above 0.95, vary depending on the changing K values.

Model No	K = 5	K = 10	K = 15	K = 20	K = 25	K = 30
18	0.972	**0.973**	0.973	0.973	0.973	0.973
15	0.970	0.971	0.971	**0.972**	0.972	0.972
24	0.961	0.962	**0.963**	0.962	0.961	0.961
21	0.960	**0.962**	0.959	0.958	0.956	0.956
6	0.957	**0.958**	0.957	0.958	0.958	0.958
3	0.955	**0.956**	0.956	0.956	0.956	0.956

managed to achieve an accuracy rate greater than 95%. The results of the additional tests for these six models, considering different K values, are presented in Table 8.

It's crucial to strike a balance when selecting the K value: ideally, it should be as small as possible to speed up the training process, but at the same time, it should maximize accuracy. In Table 8, it is clear that a K value of 10 is optimal for most of the best performing models. Only models 15 and 24 deviate from this trend, with their optimal K values being 15 and 20, respectively. However, it is important to note that although models 15 and 24 achieved slightly higher accuracies with a larger K value, the improvement over the K = 10 baseline was minimal no more than 0.1%. To illustrate, while model 15 achieved an accuracy rate of 0.971 with K = 10, it reached 0.972 with K = 20, an increase of only 0.1%. Considering that the training time increases as the value of K increases, the 0.1% increase in accuracy can be considered negligible. On the other hand, as K increases, the proportion of the test data in the model decreases. For instance, when K is 10, the test data represents 10% of the total dataset, whereas when K is 20, it represents 5%. The decreasing proportion of test data as K increases potentially compromises the robustness of the model in testing. In order to obtain more accurate accuracy values from experiments, a substantial test data size is always required. Therefore, for models numbered 15 to 24, a K value of 10 can be considered optimal. Table 9 shows the comparison of the results of this study with other studies in the literature conducted in the same area.

This study was more successful than previous similar studies. First of all, when compared to Islam Mayda's study on Turkish tweets, this study includes more occupations and more tweets. Mayda used 25,000 tweets in his work, and the estimation of the profession was limited to ten

Table 9
Comparison of the our proposed model with state of the art studies.

Studies	Method	Accuracy	Precision	Recall	F1
Our Study	**MLR**	**0.973**	**0.973**	**0.973**	**0.973**
İslam Mayda	SVM	0.990	NA	NA	0.990
Kazi Zainab et al.	ALBERT	0.950	0.920	0.910	0.900
Jiaqi Pan et al.	GCN	0.610	NA	NA	NA
Shaojie Yan et al.	T-LSTM + LIWC	0.591	NA	NA	0.508
Shayan Vassef et al.	DNN	0.540	NA	NA	NA
Preot¸iuc-Pietro et al.	GP	0.527	NA	NA	NA
Tianran Hu et al.	LIWC	NA	0.780	NA	NA

occupations, and 99% success was achieved. In our study, over 500 thousand tweets, 65 occupations and 36 occupational groups were estimated, and 97.3 success was achieved in all metrics. This study also outperformed similar studies in different languages. From the study by Preot¸iuc-Pietro et al., who made predictions for nine professions and obtained 52.7% accuracy; From the study by Tianran Hu et al., who made predictions for eight occupations and achieved a precision of 78 %; From the study by Kazi Zainab et al. who estimated the health-related occupations of Twitter users working only in the medical field, with an F1 score of 90%; from the study by Jiaqi Pan et al., which used 4,557 Twitter accounts and achieved an accuracy rate of 61%; Higher success was achieved in the study by Shayan Vassef et al., which consisted of 1,314 observations and estimated occupational titles in nine categories and achieved accuracy rate 54%.

In addition to the successes of the study, there are also some weaknesses. The first of these is that some professions cannot be predicted at all, and some professions can only be predicted by the occupational group in which they are found. When Table 1 is examined, it is seen that the first 13 occupational groups from top to bottom are made up of a combination of many occupations, while the other groups contain only one occupation. As each of the 24 models created in this study only predicts the occupational group, the occupations within these 13 occupational groups cannot be predicted individually. Another weakness of the study is that the datasets generated take too long to process and train in some machine learning algorithms, as they contain hundreds of thousands of tweets and millions of words.

5. Conclusion

This study aims to predict the occupational groups of users who share on Twitter by using machine learning methods. In the experiments, 24 different models were created using machine learning algorithms called MNB and MLR. The most successful of the models created is the model using the MLR algorithm, which consists of ten occupation-based tweet groups and uses a dataset preprocessed with the Zemberek library. This model achieved 97.3% success in all calculated metrics.

The results of the study show that the Count vectorization method outperforms the TF-IDF vectorization method among the most efficient models. It was also found that selecting the most important features (max_features) helps both to reduce the overall size of the model and to improve its performance. In this study, a number of problems arising from the unique structure of Turkish or the different uses of Turkish due to cultural differences were successfully solved in the data preprocessing phase. The accuracy rate was improved with innovative methods applied in the study. One of these innovations is the extension of the stopword list, which is already used in many Turkish NLP projects. Many unnecessary expressions, which were not recognized in previous studies and caused the models to shift to lower accuracy, were removed from the main texts.

Another innovative approach is the detailed analysis and removal of phrases that are not helpful in identifying a particular profession from the relevant Twitter messages. In this way, this study has gained a pioneering and guiding character that will pave the way for future research in the same field.

Potential future studies aim to add new occupations to the list of occupations and to increase the content of existing datasets. Twitter is a social network that grows in content every day. In the future, new accounts may be opened for non-representative professions and new resources may be created that can benefit from their content. In this way, professions that are unpredictable today will become predictable tomorrow using the methods described in this paper. In addition, the aim is to predict all occupations directly rather than groups of occupations. To increase the success, it is planned to use various machine learning algorithms and deep learning techniques such as LSTM, which have not been tested in this study.

CRediT authorship contribution statement

Zeki Ciplak: Conceptualization, Methodology, Software, Writing – original draft. **Kazim Yildiz:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abitbol, J. L., Karsai, M., & Fleury, E. (2018). Location, occupation, and semantics based socioeconomic status inference on twitter. Paper presented at the IEEE International Conference on Data Mining Workshops (ICDMW).

Akın, M. D., & Akın, A. A. (2007). Türk dilleri için açık kaynaklı doğal dil işleme kütüphanesi: ZEMBEREK. *Elektrik mühendisliği, 431*, 38–44.

Aletras, N., & Chamberlain, B. P. (2018). Predicting twitter user socioeconomic attributes with network and language information. In Proceedings of the 29th on Hypertext and Social Media (pp. 20-24).

Ali, L., Khan, S. U., Anwar, M., & Asif, M. (2019). Early detection of heart failure by reducing the time complexity of the machine learning based predictive model. Paper presented at the International Conference on Electrical, Communication, and Computer Engineering (ICECCE).

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics, 16*(1), 125–127.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The'K'in K-fold Cross Validation. Paper presented at the ESANN.

Barseghyan, L. (2013). On some aspects of Internet slang. *Graduate School of Foreign Languages N, 14*, 19–31.

Bernstein, B. (1960). Language and social class. *The British Journal of Sociology, 11*(3), 271–276.

Bernstein, B. (2003). *Class, codes and control: Applied studies towards a sociology of language, (Vol. 2):*. Psychology Press.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit: " O'Reilly Media, Inc.".

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics, 44*(1), 197–200.

Communications, R. o. T. s. P. D. o. (2022). Twitter Use Report. Retrieved from http://sosyalagharitasi.gov.tr/report/download/95.

Conroy, B., & Sajda, P. (2012). Fast, exact model selection and permutation testing for l2-regularized logistic regression. Paper presented at the Artificial Intelligence and Statistics.

Demir, E., & Tepecik, A. (2022). Türkçe ses kayit verilerinin countvectorizer ve TF-IDFVectorizer yöntemleri ile BERT modelleri olarak google colab platformunda ve

rapidminer'da makine öğrenmesi algoritmalariyla analizi. *Fırat Üniversitesi Fen Bilimleri Dergisi, 34*(1), 19–29.
Dixon, H. B., Jr (2011). Texting, tweeting, and other Internet abbreviations. *Judges J., 50*, 30.
Fung, S. W., Tyrväinen, S., Ruthotto, L., & Haber, E. (2019). ADMM-Softmax: an ADMM approach for multinomial logistic regression. arXiv preprint arXiv:1901.09450.
Gaur, P., Vashistha, S., & Jha, P. (2023). Twitter Sentiment Analysis Using Naive Bayes-Based Machine Learning Technique. In Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022 (pp. 367-376): Springer.
Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. arXiv preprint arXiv:1809.08651.
Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics, 49*(3), 291–304.
Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv preprint arXiv:1905.12787.
Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. Paper presented at the IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing.
Hu, T., Xiao, H., Luo, J., & Nguyen, T.-v. T. (2016). What the language you tweet says about your occupation. Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.
İşkur. (2023). Türk Meslekler Sözlüğü. Retrieved from https://esube.iskur.gov.tr/Meslek/meslek.aspx.
Jiang, L., Wang, S., Li, C., & Zhang, L. (2016). Structure extended multinomial naive Bayes. *Information Sciences, 329*, 346–356.
JustAnotherArchivist. (2022). Snscrape: a social networking service scraper in Phyton. Retrieved from https://github.com/JustAnotherArchivist/snscrape.
Kern, M. L., McCarthy, P. X., Chakrabarty, D., & Rizoiu, M.-A. (2019). Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences, 116*(52), 26459–26464.
Labov, W. (2006). *The social stratification of English in*. New York city: Cambridge University Press.
Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *IEEE software, 14*(2), 67–75.
Losada, D. E., & Azzopardi, L. (2008). Assessing multivariate Bernoulli models for information retrieval. *ACM Transactions on Information Systems (TOIS), 26*(3), 1–46.
Manuel, K., Indukuri, K. V., & Krishna, P. R. (2010). Analyzing internet slang for sentiment mining. Paper presented at the Second Vaagdevi international conference on information Technology for Real World Problems.
Marcot, B. G., & Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics, 36*(3), 2009–2031.
Mayda, İ. (2022). Türkçe tweetlerden makine öğrenmesi ile meslek tahmini. *Avrupa Bilim ve Teknoloji Dergisi, 40*, 55–60.
McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *Paper presented at the AAAI-98 workshop on learning for text categorization*.
Miller, S., Jr (1962). Relationship of personality to occupation, setting, and function. *Journal of Counseling Psychology, 9*(2), 115.
Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Paper presented at the Proceedings of the 25th international conference on world wide web*.
Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. *J. Inf. Technol. Comput. Sci, 6*, 61–71.
O'Carroll, J. M. (2023). *Exploring Gender Bias in Semantic Representations for Occupational Classification in*. NLP: Techniques and Mitigation Strategies.
Pan, J., Bhardwaj, R., Lu, W., Chieu, H. L., Pan, X., & Puay, N. Y. (2019). Twitter homophily: Network based prediction of user's occupation. *Paper presented at the Proceedings of the 57th annual meeting of the association for computational linguistics*.
Pasechnaya, L. A., & Shcherbina, V. E. (2020). Internet neologizms as youth slang supplementation: The main ways of formation. *European Proceedings of Social and Behavioural Sciences*.
Patel, A., & Meehan, K. (2021). Fake news detection on reddit utilising CountVectorizer and term frequency-inverse document frequency with logistic regression, MultinominalNB and support vector machine. *Paper presented at the 32nd Irish Signals and Systems Conference (ISSC)*.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.
Preoţiuc-Pietro, D., Lampos, V., & Aletras, N. (2015). An analysis of the user occupational class through Twitter content. *Paper presented at the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications, 212*, Article 118715.
Raza, G. M., Butt, Z. S., Latif, S., & Wahid, A. (2021). Sentiment analysis on COVID tweets: An experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models. *Paper presented at the International Conference on Digital Futures and Transformative Technologies (ICoDT2)*.
Samani, Z. R., Guntuku, S. C., Moghaddam, M. E., Preoţiuc-Pietro, D., & Ungar, L. H. (2018). Cross-platform and cross-interaction study of user personality based on images on Twitter and Flickr. *PloS one, 13*(7), e0198660.
Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. In: Carnegie-Mellon University. Department of Computer Science Pittsburgh.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(2), 111–133.
Sucar, L. E., & Sucar, L. E. (2021). *Bayesian classifiers* (pp. 43–69). Probabilistic Graphical Models: Principles and Applications.
Temel, Z. F., Bekir, H., & Yazıcı, Z. (2014). *Erken çocuklukta dil edinimi*. Vize Publisher.
Twitter. (2023). How to customize your profile Retrieved from https://help.twitter.com/en/managing-your-account/how-to-customize-your-profile.
Uladı, G., Eryılmaz, D., Geyik, M., & Öztürk, M. (2019). 36–72 aylik çocuklarin dil gelişim özelliklerinin çeşitli değişkenler bakimindan incelenmesi. *Karabük Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9*(1), 265–277.
Vassef, S., Toosi, R., & Akhaee, M. A. (2022). Job Title Prediction from Tweets Using Word Embedding and Deep Neural Networks. *Paper presented at the 30th International Conference on Electrical Engineering (ICEE)*.
Vernon, M. D. (1941). The relationship of occupation to personality. *British Journal of Psychology. General Section, 31*(4), 294–326. https://doi.org/10.1111/j.2044-8295.1941.tb00996.x
Wright, S. J. (2006). Numerical optimization (T. V. Mikosch, S. I. Resnick, & S. M. Robinson Eds. Second ed.): Springer.
Yan, S., Zhao, T., & Deng, J. (2022). Predicting Social Media User Occupation with Content-aware Hierarchical Neural Networks. *Paper presented at the 8th International Conference on Big Data and Information Analytics (BigDIA)*.
Zainab, K., Srivastava, G., & Mago, V. (2021). Identifying health related occupations of Twitter users through word embedding and deep neural networks. *BMC bioinformatics, 22*(10), 1–16.
Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics, 187*(1), 95–112.