INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
S. E. Erol, Vol. 11, No. 2, pp. 69-84

Detecting Personal Health Data Disclosures in Turkish Social Data

Salih Erdem Erol[1] , Şeref Sağıroğlu[2], Mustafa Umut Demirezen[3]

[1]Gazi University, Department of Computer Forensics, Ankara, Turkey
[2] Gazi University, Department of Computer Engineering, Ankara, Turkey
[3] Huawei Technologies, R&D Center, Artificial Intelligence Enablement Department, Ankara, Turkey
Corresponding Author: salih.erdem@gazi.edu.tr

Research Paper Received: 09.05.2022 Revised: 15.06.2022 Accepted: 16.06.2022

Abstract—The number of users of social networking environments is increasing day by day. In parallel with the number of users, new social networking platforms are also taking place on the internet according to the wishes and needs of the users. Social networking environments, which are in an indispensable position with the instinct of socialization, also provide an environment for unconscious personal data disclosures. In this study, the health data disclosed by users in social networks due to lack of awareness has been focused on. By using the data collected from Twitter, it is aimed to identify the tweets that disclose health data. To achieve this purpose tweets collected from Twitter in accordance with search keywords about personal health experiences and labelled by a group of computer engineers. Created corpus preprocessed with natural language processing tool for Turkic languages, named Zemberek, and classified with fastText library. With language model created, tweets containing personal health data disclosure were detected with %88 accuracy. The main contributions in this paper are mainly; being the first study to detect personal health data disclosures in Turkish language, creation of Turkish search keywords that will serve as a reference for obtaining data to meet the health data domain, instead of disease-specific approach seen frequently in literature a holistic perspective implemented by collecting tweets containing many distinct keywords about health experiences, and creation of Turkish data corpus by manually labelling around 4.500 tweets in personal health data domain.

Keywords—health data, disclosure, privacy, social network, social data, personal data

1. Introduction

With the developing technology and easier access to applications or services, social networking environments have become an indispensable part of people's daily lives. While these interests and developments increase the types and numbers of social networking platforms, they also increase the number of users day by day.

In today's digital life, the number of active users has reached 4.2 billion with a double increase in the last 5 years, the time spent in social media environments has increased by 1.5 times and reached an average of 2 hours and 25 minutes per day [1]. Although being online for long times, it is seen that the environment in which people trust the least is social media with %56 [2]. Moreover, users are not aware of the data they share on the internet, especially on social networking environments. They trust the least where they share their personal data most openly [2].

Despite the lack of trust people have towards social networking environments, the dominance of socialization instinct creates an environment for people to share on social networks that may cause them to violate their privacy and confidentiality [3]. In 2021, %56.3 of social network users stated that the purpose of using the internet is to be in contact with their friends and family, which shows the importance of the socialization instinct [1]. This instinct also triggers the desire to be online regardless of time and place, so mobile usage rates have reached %70.5 for Turkey and %55.7 as the world average [1]. Because of this mobilization and lack of awareness, personal data disclosures in social networks have become common.

When the subject is carefully examined in terms of individuals, it is seen that they are not aware of the importance of social data, which they are in the position of producer in their daily lives. Although social networking environments are considered private by users, they openly share their private data on these platforms [4]. However, unlike individuals, attackers have high motivation and awareness in this regard. In the report published by IBM about the costs of data leaks in 2021, attacks on the theft of customers' personal identifiable information (PII) accounted for %44 of all attacks that were the subject of the study, and with a value of 180$ for each record it is the most costly data [5].

When the cost of data leaks to the sectors is investigated, healthcare, finance, medicine, technology and energy took the first 5 places in 2021 [5]. In the healthcare sector, which ranks first, the average cost increased from $7.13 million in 2020 to $9.23 million in 2021, with an increase of %29.5 [5], and a %10.9 increase has taken place in attacks on health data in 2021 compared to the previous year [6].

One of the important points in the increase of attacks targeting healthcare sector is that personal health information does not change over time. In general, information circulating or captured on social networks or from internet can be evaluated in two groups as temporary data such as credit card information, phone number, e-mail address that may change over time, and permanent information such as health experiences, date of birth, family informations [7]. For this reason, personal health information that will not change once acquired such as psychological state of the person, the operations they undergone, and the results of the analysis are great importance. Also, it is inevitable that permanent personal data will form a very important basis for social engineering attacks.

In a data breach report published by the Personal Data Protection Authority in Turkey in 2022, it was stated that 2 million 500 thousand records belonging to an estimated 500,000 people were seized by cyber attackers [8]. Considering the number of attacks targeting the healthcare sector in recent years, the amount of affected users and costs, the value and importance of health data is clearly seen. Given that 85% of data leaks are caused by the human factor [9], it is clear that data leaks cannot be prevented only by technological measures. Moreover, personal data is not only obtained from corporate environments, but also publicized by individuals by sharing it on public platforms such as social networking environments and blogs without any privacy setting. For this reason, there is a need to increase personal awareness in addition to the measures taken and increased security measures in the healthcare sector. In this context, a study has been carried out to detect unconsciously disclosed personal data in the field of healthcare, where costly data loss and attacks are frequently seen in recent years.

The rest of this paper is structured as follows: Section 2 introduces related works, Section 3 explains material and method, Section 4 provides

experimental setup and results, Section 5 presents findings and evaluations about the conducted study.

2. Related Works

To be used in the study [10] aimed at extracting information from the posts related to heart attack on Twitter a total of 7000 tweets, 770 informative and 6230 non-informative, were used. The results obtained with DNN (Deep Neural Networks), SVM, Naive Bayes and J48 Decision Tree algorithms were compared and the best results were achieved using the DNN algorithm. An accuracy of %95.2, and F1-score values of %73.6 and %97.4 for informative and non-informative classes, respectively, were obtained.

In the study, the Majority Voting-based Ensemble Deep Learning (MVEDL) model was used to detect tweets containing information about COVID-19 shared during the pandemic [11]. For model training, the dataset [12] presented by the organizers of W-NUT, containing a total of 10,000 tweets with 4719 informative and 5281 non-informative tweets, was used. In order to increase the model performance, experiments were conducted with combinations of different machine learning and deep learning models. The best results were obtained with the COVID-Twitter BERT, BERTweet, and Roberta deep learning models. It is stated that the proposed model outperforms traditional machine learning and deep learning models with an accuracy of %91.75 and an F1-score of %91.14.

An ontology-based study was carried out on Turkish posts to automatically identify tweets related to public health [13]. The data collected from Twitter in two different periods is semi automatically labelled with the ontology-based system, and tests were carried out. With the different methods and models applied, results ranging from %65-%99.1 were obtained.

In the study on the detection of sensitive and non-sensitive personal data [14] a system called Tweet-Scan-Post (TSP) has been developed. The system is designed to classify the posts made in personal, corporate and health domains. Rule-based, dictionary-based approaches and machine learning methods are used together, and the results are shown through an application. The best results for the classification of sensitive data were achieved with Naive Bayes, and accuracy rates of %83.39, %68.47 and %73.87 were obtained for personal, corporate and health classes, respectively.

In order to detect the posts about personal health, with the classification made by tagging 2000 tweets about cancer, depression, hypertension and leukaemia [15] 34 high-impact health status detection were made with an accuracy of %77. One of the important findings in the study is that people share about themselves or others according to the type of disease. For example, while %69 of the posts about insomnia is related with themselves, only %21 of the tweets about Down Syndrome belong to the person who posts. Another important finding is that %44 of tweets containing keywords related to health status contain disclosures about health information.

In the study [16], which aims to detect sensitive data about health by using social data, tweets collected from Twitter were filtered in accordance with a 20-word dictionary. 11.647 tweets filtered by keywords were labelled in line with Regret Theory. Primary (me, my) / Secondary tweet score (he, his, she, her), Named Entity Recognition score of tweets, Term Frequency-Inverse Document Frequency (TF-IDF), Cyber-Keyword Ratio, hashtag mentions and user mentions used as features for classification. Accuracy rates ranging from %50.01 to %67.06 were obtained in the classification tests performed with different algorithms.

FLU2013 dataset which was created in a previous

study [17] to detect personal health posts in social media environments by using Influenza reports and the PHM2017 datasets created within the scope of the study were classified and compared with different methods and algorithms [18]. The PHM2017 dataset was created by tagging 7.192 English tweets covering 6 different types of diseases, including alzheimer's, hearth attack, parkinson's, cancer, depression and stroke. Tweets are labelled as self-mention, other-mention, awareness, and non-health. It has been revealed that the method developed within the scope of the study, called WESPAD, obtained higher results than the compared models. F1-measure values between 0.818-0.851 for FLU2013 dataset and 0.652-0.695 for PHM2017 dataset were achieved.

In the study [19] conducted to create a corpus for physical health data disclosures, researchers focused on information about positive test results, physical symptoms and health history during the pandemic period. Individuals' self-mentioned posts and posts about their relatives were included in the scope. Corpus was created by labelling 1.494 tweets randomly selected from 22.331 tweets collected from Twitter.

Assuming that personal health experiences may contain important findings about the diseases or treatments of the person, it is aimed to detect personal health experiences by collecting posts containing 103 pre-defined drug names in Twitter posts [20]. 12.331 tweets randomly from the created data set were labelled by evaluating whether they contain personal experience. Accuracy values ranging from 0.637 to 0.815 were obtained in the classifications performed with different algorithms within the scope of the study. The best results were achieved in the classifier where Word Embedding and LSTM were used together.

In the study [21] aiming detection of personal sensitive data in unstructured texts, the data is primarily classified as sensitive and non-sensitive. Afterwards, the data classified as sensitive were categorized under 14 different classes in line with General Data Protection Regulation (GDPR) and studies examined in the literature. For the categorization of the data, SVM, DT, LR, NB and RF algorithms were used and F1-Measure values between 0.79-0.88 were obtained.

Content analysis study was conducted in line with the posts of health-related users, and 700 tweets were analysed in depth [22]. While the general purpose is to determine whether the contents are health-related or not, 5 categories for health status posts and 7 groups for users who post are suggested. According to analyses conducted on the data set, %17.6 posts related to personal health was detected.

In the analysis study carried out for the personal data disclosures made by health professionals, 43.374 tweets posted under the hashtag [23] #ShareAStoryInOneTweet were collected. Analysis was carried out by selecting 1206 tweets that were found to be posted by doctors, nurses or other health professionals. According to analysis conducted, it is seen that healthcare professionals disclose personal health data of others, %74.1 of which belong to doctors and %14.9 to nurses, the rate of personal data disclosure of family members or close friends is %32.1, and %2 of tweets include disclosures directly related to the patient's name.

In the study which aims to identify fake news about COVID-19 in Turkish language [24], a new dataset was created by gathering data from 5 various sources. Total of 2110 samples were collected, labelled, and preprocessed. In order to identify fake news by using created dataset, 5 conventional machine learners, 9 deep learners and advanced neural language transformers were implemented and accuracy values between 0.636 and 0.985 were obtained. The best results were achieved with neural trans-

formers particularly the BertTURK transformer.

Apart from studies detecting personal health data disclosures, disease prediction [25], [26], disease detection [27], disease trends [28], health-related subject classifications [29] and studies such as analysis and sentiment analysis of health conditions/outbreaks/diseases [30], [31] is frequently included in the literature and the interest in these issues is increasing day by day.

When the studies in the literature are evaluated in general, it is seen that there are no studies about detection of personal health information disclosures through Turkish tweets, there are limited studies on the detection, analysis and classification of health information in Turkish, and the main difficulties of the studies in healthcare field are time-consuming data set preparation phase and manual labelling requirement. It has been observed that the performances in the classification studies for the detecting personal health data disclosures on other languages accuracy values between 0.50-0.88 achieved.

3. Material and Method

One of the most important steps in classification problems that are tried to be solved using natural language processing and machine learning methods is collecting and labelling the data set. Errors and incomplete applications to be made at this stage will be among the factors that primarily affect the performance of the model. The data set preparation process carried out within the scope of this study was conducted in 5 main steps, and each step was explained in subsections. By using data corpus prepared, tweets were classified with fastText. Research methodology followed in this study is presented in Figure 1.

3.1. Deciding data gathering platform

OSNs (Online Social Networks) create a medium for users to create their own profiles, express themselves, communicate with others, share their own content, view and comment on content created by other users [32]. In order for the collected data to be labelled and the language model to be created accurately, first of all, the scope of the data to be studied should be well defined and its structure should be understood. The data gathered through social media, in the literature, is called Social Data [33] and the conceptual model of social data is presented in Figure 2.

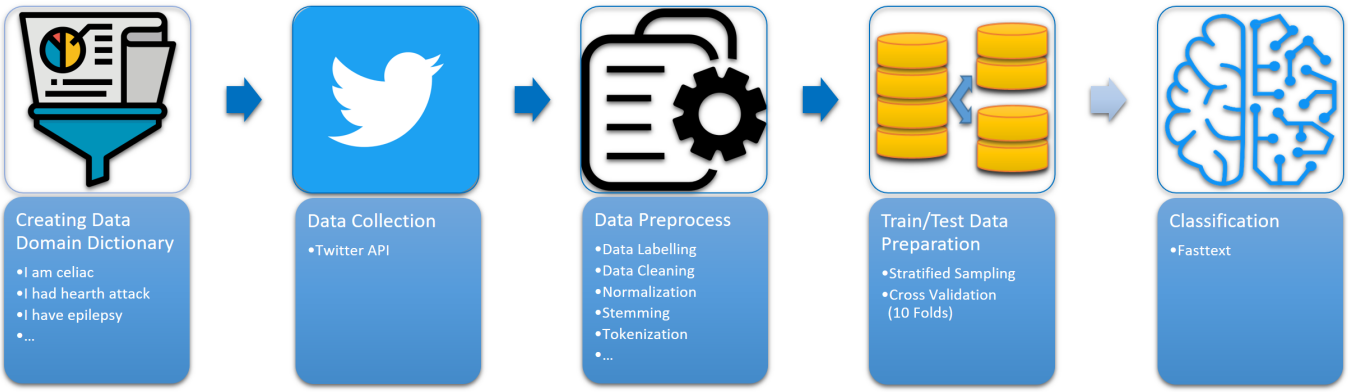

Figure 1. Research methodology.

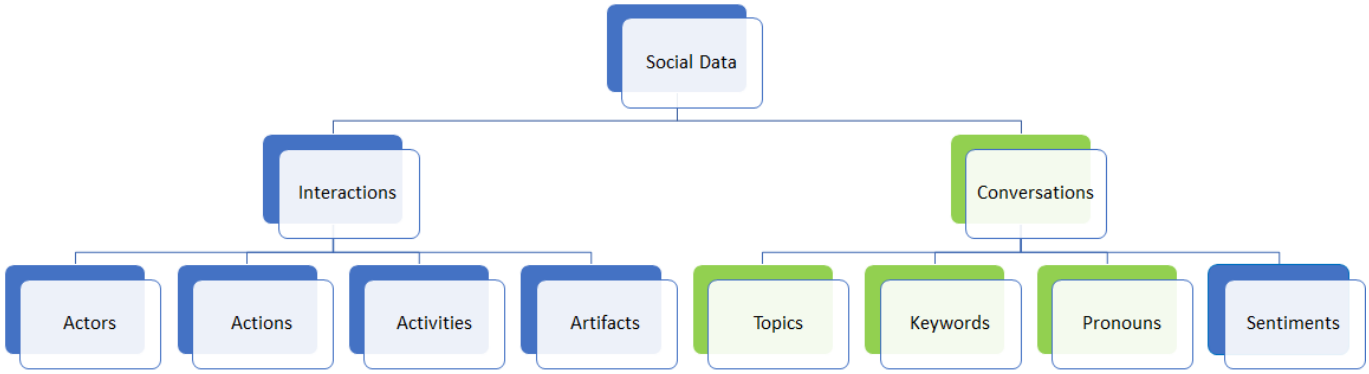

Figure 2. Social data conceptual model [33].

Since it is aimed to detect disclosure of health data information through tweets of individuals, topics, pronouns, and keywords classified under the title of conversations are focus of this study. In order to obtain the data in the focus area of our study, in the literature, it is seen that the Twitter platform is preferred due to its advantages. Main advantages can be summarized as mentioned below.

- Easy access to individuals' data on Twitter with API [10], [19] [21].
- It enables content and subject-based studies to be carried out easily with hashtag structure [10], [19].
- Twitter users are more likely to share their personal information with an unknown community than other social networks such as Facebook [21].
- Unlike many other social networking environments, the anonymity provided in Twitter encourages users to share their sensitive data [19], [34].
- Allowing a large population to share their personal experiences in a less biased and natural way thanks to anonymity [34].
- No need for mutual follow-up of users to access or view other user's data.

Considering the aforementioned advantages, especially because of easy access to data and the higher probability of containing personal data of individuals in posts, Twitter platform was preferred.

3.2. *Data collection*

In order to obtain data set for classification, violation notifications of the Personal Data Protection Agency, the most common health problems in Turkey, attacks on health information and the affected data, and judicial decisions were examined and chosen search keywords in Turkish presented in Table 1 with similar meanings in English.

Today, the privacy of personal data is expressed as the ability to control one's own data. The right of the individual to control his own data is protected by the concept of explicit consent. Explicit consent means that the person consents to the processing of his/her data, either voluntarily or upon request from the other party [35]. Therefore, in order to be able to talk about personal data, the data must first be produced by and should be belong to the individual. Considering posts and replies in Twitter environment, since it is directly shared by individuals, its belonging to the person is undisputed. Replies are also important in terms of personal data disclosure.

Table 1.
Personal health data keywords and labels list used for corpus preparation.

Label Name	Search Keywords (English-Turkish)	Number of Labelled Data
operationtest (ameliyattahlil)	My surgery, my tests, my values are low, my values are high, my blood results, my test results (Ameliyat oldum, tahlillerim, değerlerim düşük, değerlerim yüksek, kan sonuçlarım, tahlil sonuçlarım)	272
asthma (astim)	I have asthma (Astım hastasıyım, astımım var, astım oldum, astım olduğum, astım rahatsızlığım)	48
celiac (colyak)	I'm celiac, I have celiac disease (Çölyak hastasıyım, çölyak rahatsızlığım, çölyak olduğumu)	6
covid (covid)	My test is positive, I have covid, I have corona, I am in quarantine (Testim pozitif, covid oldum, korona oldum, karantinadayım)	237
adhd (dehb)	I have ADHD, We have ADHD (DEHBli olduğumu, DEHBli olduğumuzu, DEHBliyim)	13
diabetic (diyabet)	I am diabetic (Diyabet hastasıyım, şeker hastasıyım, şeker hastası olduğumu, şekerim)	79
disabled (engelli)	I am disabled (Engelliyim, engelli olduğum, engelli olmam)	365
epilepsy (epilepsi)	I have epilepsy (Epilepsi hastasıyım)	12
hernia (fitik)	I have a hernia, I had a hernia surgery, I am a hernia patient(Bel fıtığım var, fıtık ameliyatı oldum, fıtık hastasıyım)	58
pregnant (hamile)	I'm pregnant (Hamileyim, hamile olduğumu, hamile olduğum)	55
disease (hastalik)	I got disease, I have a problem, I'm sick (Hastalığına yakalandım, rahatsızlığım var, Hastasıyım, hastası olduğumu)	408
hepatitits (hepatit)	I have hepatitis (Hepatit hastasıyım, hepatit olduğumu, hepatit olduğum)	23
hearth (kalp)	I'm a heart patient, I had a heart attack (Kalp hastasıyım, kalp krizi geçirdim)	272
cancer (kanser)	I have cancer, I beat cancer, I have leukemia (Kanser hastasıyım, kanseri olduğumu, kanseri yendim, Lösemi hastasıyım, lösemi olduğumu)	88
migraine (migren)	I have migraine, my migraine (Migren hastasıyım, migrenim var, migrenim)	30
msals (msals)	I have MS, I have ALS (MS hastasıyım, ALS hastasıyım, MS hastası olduğumu, ALS hastası olduğumu)	15
autistic (otistik)	I am autistic individual, I am autistic (Otistik bireyim, otistiğim)	15
rheumatism (romatizma)	I have rheumatism, I have rheumatism complaints (Romatizma hastasıyım, romatizma şikayetim var, romatizma hastası olduğumu, romatizmam)	6
psoriasis (sedef)	I have psoriasis (Sedef hastasıyım, sedef hastası olduğumu)	5
bloodpressure (tansiyon)	I'm a blood pressure patient, my blood pressure (Tansiyon hastasıyım, tansiyon hastası olduğumu, tansiyonum)	73
treatment (tedavi)	My treatment, I'm getting treatment, my treatment has started, I've started my treatment (Tedavim, tedavisi görüyorum, tedavim başladı, tedavisine başladım)	102
diagnose (teshis)	Diagnosed (Teşhisi kondu, teşhisi koyuldu, teşhisi koydu, teşhis edildi)	114
thyroid (tiroid)	I have a thyroid disease, I have a thyroid condition (Tiroid hastasıyım, tiroid rahatsızlığım)	43
failure (yetmezlik)	Heart failure, kidney failure, failure (Kalp yetmezliği, böbrek yetmezliği, yetmezliği)	37
nonpersonal (nonkvkk)	In the data set obtained for all search keywords, the posts that are considered to be non-disclosure are labelled as nonkvkk (Tüm arama kelimelerine ilişkin elde edilen veri setinde ifşa içermediği değerlendirilen paylaşımlar nonkvkk olarak etiketlenmiştir).	2720

Because when the collected data examined it is seen that people frequently ask questions to health professionals about their health problems under posts, people with similar health problems support each other, and share their health experiences.

Classification can be performed successfully with datasets containing small amounts of but distinctive types of samples [15]. Therefore, it is aimed to identify a wider set by differentiating search keywords related to health information as much as possible. In order to collect health information that is not covered by search keywords, keywords that can be used with all diseases containing general expressions such as ”I am sick”, ”diagnosed”, and ”my treatment” were also used in data collection process.

In line with the chosen search keywords, 9.476 tweets and replies were collected weekly via Twitter streaming API between 21.12.2021 and 23.03.2022. After checking and deleting duplicate records and highly similar posts from spam accounts, a data set consisting of 5096 tweets in total was created. The word cloud formed from the words in the obtained data set is shown in Figure 3.

3.3. Data labelling

It is very difficult to automatically process and make sense of texts with natural language processing. Because the meanings of the texts can be independent of the content [36]. For this reason, automatic high-level semantic tagging is still not used with sufficient accuracy in many areas today [37]. Thus, manually labelled datasets, called Gold Standard Corpora, are frequently used in natural language processing applications.

A high quality Gold Standard Corpora can only be achieved with the opinion of more than one expert and the harmony between multiple labellers. Labellers' decisions are affected by the uncertainty of the data, the labelling rules, and the capabilities of the labeller [36]. Performing labelling by more than one labellers and manually makes the process very costly.

According to studies carried out to reduce the data set labelling costs, more than %90 similarity was achieved in the labels made by more than one field expert, therefore, a single domain expert labelling will be sufficient, a team of 4 people who are not field experts can perform labelling with more than %90 similarity, therefore it will be sufficient to form groups without field experts from 4 people [38].

In this direction, in order to create a Gold Standard Corpora, the rules and examples were explained to 4 volunteer labellers, who are computer engineers, and data labelling was carried out. In data labelling process basically whether the tweets directly included personal health data disclosure about individuals themselves or about their family members was checked.

Studies in the literature were generally conducted as disease-specific or general purpose text classification [18]. In this study, the general purpose point of view was adopted, since the health data are handled with a holistic perspective and binary classification was made. However, in order to enable disease-specific academic studies to be carried out using the created data set in the future, the data were first labelled with 25 different labels presented in the Table 1. Labels obtained from labellers were compared, necessary corrections were made and the data set was finalized.

3.4. Data preprocessing

In order to create a classification model, the data set must be preprocessed. In many classification studies better results achieved with preprocessed data [39]. To have smaller vector space and better performance with less dimension, preprocessing steps are implemented. Open source Zemberek Library [40] was preferred to be used, on natural language processing processes, and code developments were carried out in Java programming language.

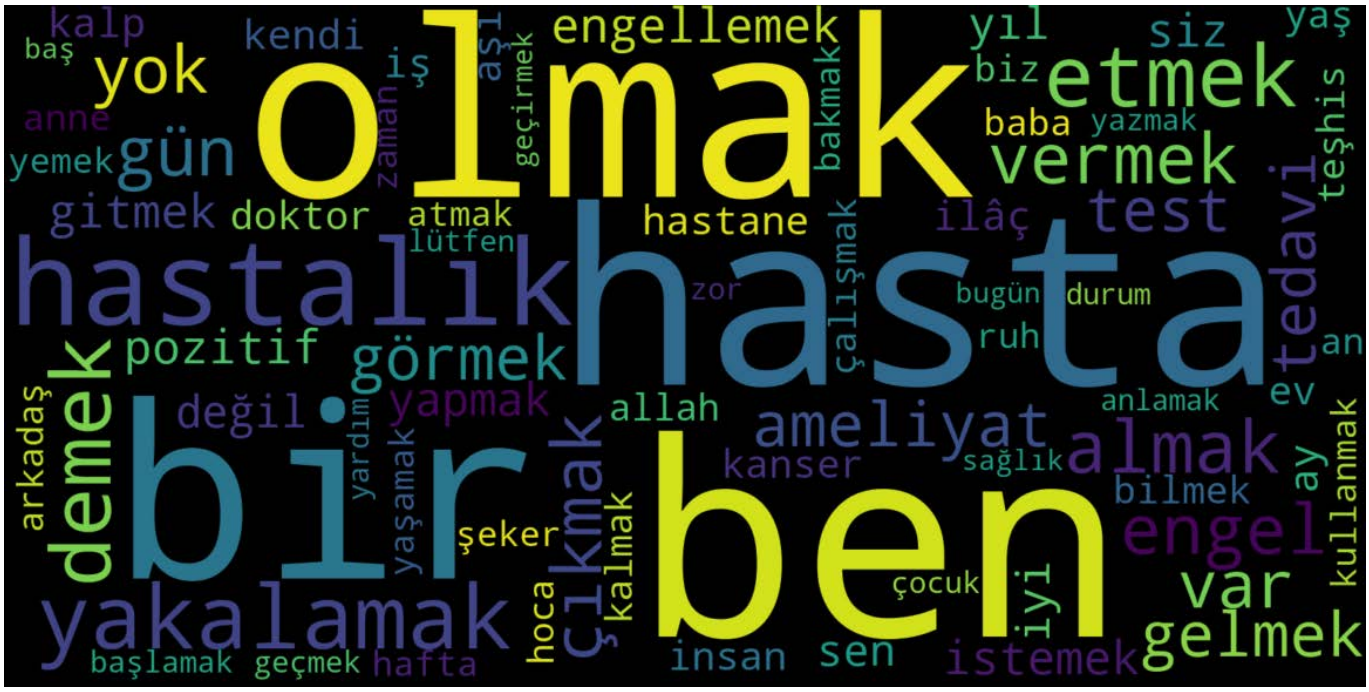

Figure 3. Word cloud of personal health data corpus (in Turkish).

After preprocessing process completed total of 4502 distinct tweets remained in health data corpus. Steps implemented to have preprocessed corpus listed below.

- Duplicated tweets were deleted.
- Special characters, punctuation marks, and emoticons that were found to cause erroneous records in natural language processing processes when written unified with words were cleaned.
- To eliminate errors made by users in tweets such as missing letters, normalization applied to corpus.
- In Turkish there are many words with more than one meaning. To eliminate this situation, semantic root extraction (disambiguation) was applied to find the most meaningful stem for tweets.
- Mentions and hashtags were deleted.
- URL link addresses were deleted.
- Stop words, which are considered to have no semantic contribution, were deleted.
- All words converted to lower-case.
- Duplicated tweets were deleted (After preprocessing some duplicate texts can occur).

Regarding the data preprocessing steps, sample data is presented in Table 2 as original tweet and preprocessed tweet with similar meanings in English.

3.5. *Train and test data preparation*

The step of creating training and test data while developing language models is very important in terms of obtaining high accuracy results, especially in studies with unbalanced data sets. There are many different sampling methods used to create training and test sets in accordance with the purpose of the research. In the literature, sampling methods are classified under two main headings as proba-

Table 2.
Preprocessed tweet samples.

Sample 1	Original Tweet	10 yıldır panik atak hastasıyım. Defalarca atak geçirdim. Bu süreçte çok ilginç müdahale yöntemleriyle karşılaştım. Ama bugün ilk defa burnumu ve ağzımı kese kağıdıyla kapatıp nefes almamı istediler. Hayır ölmem için değil müdahale için. Evet iyi geliyormuş
	in English	I have had panic attacks for 10 years. I've had multiple attacks. In this process, I encountered very interesting intervention methods. But today, for the first time, they asked me to cover my nose and mouth with a paper bag and breathe. No, not for me to die, for intervention. Yes it sounds good
	Preprocessed Tweet	yıl panik atak hasta defalarca atak geçirmek süreç ilginç müdahale yöntem karşı bugün ilk burun ağız kese kâğıt kapamak nefes almak istemek hayır ölmek değil müdahale evet iyi gelmek
	in English	year panic attack patient suffer multiple attack process interesting intervention method against today first nose mouth pouch paper close want breathe not die intervention yes come good
Sample 2	Original Tweet	@username Öncelikle çok geçmiş olsun bende kanser olduğumu yeni öğrendim gerçekten çok zormuş kabullenmek ama hayata tutunuyoruz sımsıkı yaşamak çok güzel moral moral moral bizlet güçlü kadınlarız https://url
	in English	@username First of all, get well soon, I just learned that I have cancer, it's really hard to accept, but we are holding on to life, it's good to live tight morale morale we are strong women https://url
	Preprocessed Tweet	öncelik geçmiş olmak ben kanser olmak yeni öğrenmek gerçekten zor kabullenmek hayat tutunmak sımsıkı yaşamak güzel moral moral moral biz güç kadın
	in English	priority past i am cancer be new learn really hard accept life cling live tight good morale morale morale we power women

bilistic, which includes simple, random, stratified, cluster and multi-stage sampling types, and non-probabilistic, which includes purpose, convenience, snowball and quota sampling types [41].

Considering the prepared data set, all data with 25 different labels should be included in all training and test sets proportionally. Otherwise, results that are lower than in some folds and higher than in some folds may be encountered. For this reason, it was evaluated that the stratified random sampling model was suitable for the data set to be used in the research. In this model, the data of all labels are divided into subgroups called strata and then randomly distributed to the training and test sets to be equally represented [41].

Scikit-Learn library in Python programming language was used to split data set as training and test datasets with stratified sampling for 10 fold cross validation [42].

3.6. fastText

Word2vec is one of the most popular applications of word embedding. Word embedding is a method of showing similar words in close proximity to each other and was developed by Tomas Mikolov in 2013 [43]. CBOW and Skip-gram models, which are effective methods for learning vector representations of words from large amount of unstructured text data with high efficiency, are proposed with Word2Vec model. The architectures of the proposed models [43] are presented in Figure 4.

In CBOW model, surrounding words are used to find the central word, while in Skip-gram model, the central word is used to predict the surrounding words in the sentence or document.

fastText was developed by the Facebook research team in 2016 as an extension to the aforementioned word2Vec model [44], [45]. Similar to Word2vec,

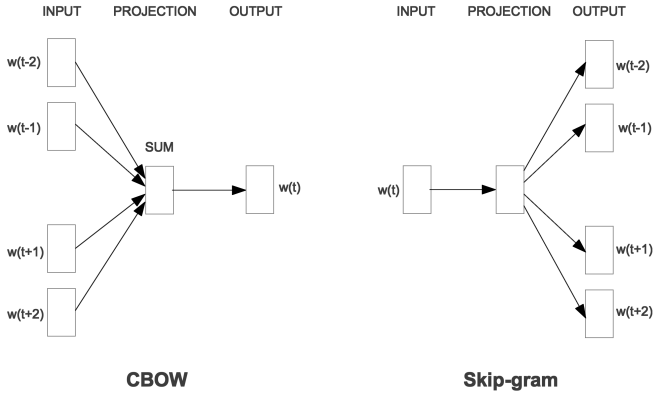

Figure 4. CBOW and Skip-gram model architectures [43].

CBOW and Skip-gram models are used to compute word representations in fastText [45]. The main contribution to the Word2vec model has been made by taking into account subword information [46]. In this method, words are learned as character n-grams and shown as the sum of character n-grams. Subwords are given instead of words as input to the artificial neural network. Architecture of fastText is similar to CBOW model in word2vec. But, to capture the order of words fastText also adds n-gram features [47]. An example showing differences between two models presented in Figure 5.

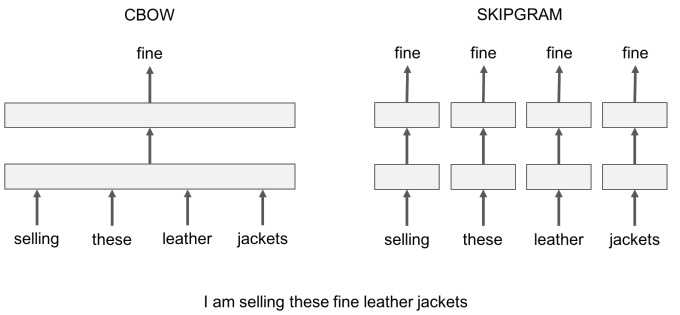

Figure 5. Examples of CBOW and Skip-gram models [45].

The fastText model consists of three layers: input layer, hidden layer and output layer. Word count and n-gram features are used in input layer, the hidden layer calculates the maximum probability and a Huffman tree is build based on the weights and model parameters as an output [48]. Softmax function is used to compute the probability distribution over the predefined classes [47].

When the studies in the literature and model architectures are evaluated, it is seen that fastText provides many advantages in text classification processes. Main advantages presented below.

- As well as, high performance, fast computation and low resource consumption, achieving results comparable to deep learning methods [47], [49].
- Using subwords that increases performance in morphologically rich languages such as Turkish [46], [50].
- Recognizing the words that are not included in the data set by using the character n-gram method and providing high-performance language models in the classification studies carried out with Turkish texts in the literature [50], [51], [52], [53].

Since it is considered that these advantages will contribute to the solution of the problem of detecting personal health data disclosures, the fastText model was preferred in the classification process to be carried out within the scope of the study.

4. Experiment and Results

Within the scope of the study, it is aimed to make a binary classification in order to detect tweets with and without personal health data disclosure. In this context, the labels in the data set created from 25 separate labels have been changed to be suitable for binary classification as ”kvkk (personal)” total

number of 2126 and ”nonkvkk (nonpersonal)” total number of 2376.

To train a model with fasText with personal health data corpus prepared in this study, two methods were implemented. Firstly, hyperparameter values presented in Table 3 used for training with different combinations to get best results.

Table 3.
Hyperparameters used for model training.

Hyperparameters	Value/Range	Default
dim (Dimension)	50-300	100
lr (Learning rate)	0.1-0.6	0.1
Epoch	5-75	5
ws (Window size)	3-5	5
wordNgrams	3-6	1

Secondly, fastText's autotune feature which tries to find best hyperparameters automatically in a given time period was used. The models were tested with test data previously prepared for 10 fold cross validation with stratified sampling.

Commonly used metrics to evaluate the performance of classification models are accuracy, precision, recall, and F1 measure. While calculating these metrics 4 parameters are required; True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

- TP: Correctly predicted positive class.
- TN: Correctly predicted negative class.
- FP: Incorrectly predicted positive class.
- FN: Incorrectly predicted negative class.

The calculation formulas for performance metrics used in this study to evalute model performance are presented in equations (1), (2), (3), (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Presicion = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Measure = \frac{2 * TP}{2 * TP + FP + FN} \quad (4)$$

fastText does not calculate metrics such as precision, recall, accuracy and F1 measure automatically for binary classification. Therefore, Scikit-Learn library functions was used to calculate performance metrics. Training and test processes were carried out for 10 folds. The best performance results obtained in line with the tests carried out are presented in Table 4.

The data collected from Twitter is called as short text. The classification of short texts has many disadvantages. In addition to the low number of words and low word frequencies, the rapid change of subjects in the text reduces the performance of classification processes [54]. While the best results in terms of high classification performance are obtained with supervised learning methods, it is seen that the correct classification rates of the methods based on unsupervised or semi-supervised learning are lower [55]. Among the reasons for this situation are the limited number of labelled data, which can also cause over-fitting problems [56], and the sequencing of the samples [54]. Due to sequencing of samples some similar examples may be very intense in the data set or some examples are not included at all.

In addition to the disadvantages mentioned, when the special expressions, abbreviations, erroneous grammatical spellings, sarcastic and indirect expressions, and losses in natural language processing [4] are taken into account, the %88 accuracy rate obtained within the scope of the study is considered to represent a high performance.

Table 4.
Model performance for detecting personal health data disclosures.

Fold Number	Word Count	Avg.Loss	Accuracy	Label	Precision	Recall	F-1 Measure	Support
Fold 1	5538	0,026	0,874	Kvkk	0,882	0,845	0,863	213
				Nonkvkk	0,866	0,899	0,882	238
Fold 2	5572	0,025	0,889	Kvkk	0,890	0,873	0,882	213
				Nonkvkk	0,888	0,903	0,896	238
Fold 3	5570	0,026	0,889	Kvkk	0,879	0,887	0,883	212
				Nonkvkk	0,898	0,891	0,895	238
Fold 4	5499	0,043	0,891	Kvkk	0,876	0,896	0,886	212
				Nonkvkk	0,906	0,887	0,896	238
Fold 5	5543	0,029	0,889	Kvkk	0,909	0,849	0,878	212
				Nonkvkk	0,873	0,924	0,898	238
Fold 6	5500	0,036	0,896	Kvkk	0,899	0,878	0,888	213
				Nonkvkk	0,893	0,911	0,902	237
Fold 7	5502	0,033	0,849	Kvkk	0,840	0,840	0,840	213
				Nonkvkk	0,857	0,857	0,857	237
Fold 8	5574	0,038	0,862	Kvkk	0,862	0,854	0,858	212
				Nonkvkk	0,871	0,878	0,874	238
Fold 9	5508	0,039	0,913	Kvkk	0,903	0,915	0,909	213
				Nonkvkk	0,923	0,911	0,917	237
Fold 10	5497	0,038	0,865	Kvkk	0,873	0,869	0,876	213
				Nonkvkk	0,885	0,878	0,881	237
Average	5530	0,033	0,88	Kvkk	0,87	0,88	0,88	213
				Nonkvkk	0,89	0,89	0,89	238

5. Conclusion

In this study, which was carried out to detect personal health data disclosures in social networks, a corpus was formed by preprocessing (data cleaning, normalization, stemming, data labelling, deleting stop words) the data gathered from Twitter in line with the search keywords chosen for health data. Created corpus was splitted by stratified sampling method to 10 folds in terms of equal representation of each label for reliability of the tests. Within the scope of the study, a language model has been developed and classification was carried out with an average of %88 accuracy by using the fastText library.

Although there is no study in the literature that can be taken as a reference for detecting personal health data in Turkish, a higher performance has been achieved than the values obtained in the classification studies carried out for Turkish health data. In the studies on personal data carried out for English, the same result as the best result was obtained with our study. In this sense, the language model created in this study is considered to be successful.

It is considered that this study will contribute to the literature in terms of being the first study to be carried out in this field in Turkish language, the classification model created, the results and findings presented, and the preparation of the Turkish Personal Health Data Corpus that can be used as a basis for academic studies.

Conducting academic studies to detect personal data expropriated in social networks increase privacy awareness of users, with methods to be developed in line with the findings obtained in these studies personal data disclosures can be reduced and used as a proactive measure to prevent many security incidents such as identity theft, digital fraud, cyber attack, etc., on information assets.

In the future, it is aimed to increase the performance by using the prepared corpus with different language models and deep learning applications, and to make it available to users with real-time applications.

Acknowledgments

The authors would like to thank to volunteer labellers who manually labelled more than 5000 personal health data.

References

[1] S. Kemp. (2021) We Are Social. Digital 2021 Global Overview Report. "https://wearesocial-cn.s3.cn-north-1.amazonaws.com.cn/common/digital2021/digital-2021-global.pdf. Accessed: 01.05.2022.

[2] M. Timothy, F. Theodore, and S. Allison. Customer data: Designing for transparency and trust. https://www.hipaajournal.com/december-2021-healthcare-data-breach-report/. Accessed: 01.05.2022.

[3] C. Zhang, J. Sun, X. Zhu, and Y. Fang, "Privacy and security for online social networks: Challenges and opportunities," *IEEE Network*, vol. 24, no. 4, pp. 13–18, 2010.

[4] S. E. Erol and S. Sagiroglu, "Privacy awareness in social networks," in *2021 International Conference on Information Security and Cryptology (ISCTURKEY)*, 2021, pp. 57–62.

[5] C. of a Data Breach Report. (2021) Ibm security. https://www.ibm.com/downloads/cas/OJDVQGRY. Accessed: 01.05.2022.

[6] S. Alder. December 2021 healthcare data breach report. https://www.hipaajournal.com/december-2021-healthcare-data-breach-report/. Accessed: 01.05.2022.

[7] I. Lella, M. Theocharidou, E. Tsekmezoglou, and A. Malatras. (2021) Enisa threat lanscape 2021. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021. Accessed: 01.05.2022.

[8] Kamuoyu duyurusu (Veri ihlali bildirimi) – Yonca Sağlık Hizmetleri Ltd. Şti. https://www.kvkk.gov.tr/Icerik/7199/Kamuoyu-Duyurusu-Veri-Ihlali-Bildirimi-Yonca-Saglik-Hizmetleri-Ltd-Sti-. Accessed: 01.05.2022.

[9] Verizon. (2021) Data breach investigations report. https://www.verizon.com/business/resources/reports/2021/2021-data-breach-investigations-report.pdf. Accessed: 01.05.2022.

[10] O. Karajeh, D. Darweesh, O. Darwish, N. Abu-El-Rub, B. Alsinglawi, and N. Alsaedi, "A classifier to detect informational vs. non-informational heart attack tweets," *Future Internet*, vol. 13, p. 19, 2021.

[11] S. Malla and A. P.J.A., "COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets," *Applied Soft Computing*, vol. 107, p. 107495, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156849462100418X

[12] S. Ohashi, T. Kajiwara, C. Chu, N. Takemura, Y. Nakashima, and H. Nagahara, "IDSOU at WNUT-2020 task 2: Identification of informative COVID-19 English tweets," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. Online: Association for Computational Linguistics, 2020, pp. 428–433. [Online]. Available: https://aclanthology.org/2020.wnut-1.62

[13] E. Kucuk, K. Yapar, D. Kucuk, and D. Kucuk, "Ontology-based automatic identification of public health-related Turkish tweets," *Computers in Biology and Medicine*, vol. 83, 2017.

[14] G. Raju, K. Subbaraj, and P. Kumaraguru, "Tweet-Scan-Post: A system for analysis of sensitive private data disclosure in online social media," *Knowledge and Information Systems*, vol. 63, 2021.

[15] Z. Yin, D. Fabbri, S. T. Rosenbloom, and B. A. Malin, "A scalable framework to detect personal health mentions on Twitter," *Journal of Medical Internet Research*, vol. 17, 2015.

[16] R. Geetha, S. Karthika, N. Pavithra, and V. Preethi, "Tweedle: Sensitivity check in health-related social short texts based on regret theory," *Procedia Computer Science*, vol. 165, pp. 663–675, 2019, 2nd International Conference on Recent Trends in Advanced Computing ICRTAC - DISRUP - TIV INNOVATION , 2019 November 11-12, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920300703

[17] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on Twitter," pp. 789–795, 2013. [Online]. Available: https://aclanthology.org/N13-1097

[18] P. Karisani and E. Agichtein, "Did you really just have a heart attack? Towards robust detection of personal health mentions in social media," *Proceedings of the 2018 World Wide Web Conference*, 2018.

[19] R. Saniei and V. Doncel, "PHDD: Corpus of physical health data disclosure on Twitter during COVID-19 pandemic," *SN Computer Science*, vol. 3, pp. 1–10, 2022.

[20] K. Jiang, S. Feng, Q. Song, R. Calix, M. Gupta, and G. Bernard,

"Identifying tweets of personal health experience through word embedding and LSTM neural network," *BMC Bioinformatics*, vol. 19, 2018.

[21] W. B. Tesfay, J. Serna, and K. Rannenberg, "Privacybot: Detecting privacy sensitive information in unstructured texts," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2019, pp. 53–60.

[22] J. Lee, M. Decamp, M. Dredze, M. Chisolm, and Z. Berger, "What are health-related users tweeting? a qualitative content analysis of health-related users and their messages on Twitter," *Journal of Medical Internet Research*, vol. 16, 2014.

[23] W. Ahmed, R. Jagsi, T. Gutheil, and M. Katz, "Public disclosure on social media of identifiable patient information by health professionals: Content analysis of Twitter data," *Journal of Medical Internet Research*, vol. 22, 2020.

[24] M. Bozuyla and A. Özçift, "Developing a fake news identification model with advanced deep languagetransformers for Turkish COVID-19 misinformation data," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 3, 2022. [Online]. Available: https://doi.org/10.55730/1300-0632.3818

[25] J. Eichstaedt, R. Smith, R. Merchant, L. Ungar, P. Crutchley, D. Preotiuc-Pietro, D. Asch, and H. Schwartz, "Facebook language predicts depression in medical records," *Proceedings of the National Academy of Sciences*, vol. 115, p. 201802331, 2018.

[26] R. Thorstad and P. Wolff, "Predicting future mental illness from social media: A big-data approach," *Behavior Research Methods*, pp. 1–15, 2019.

[27] I. Syarif, N. Ningtias, and T. Badriyah, "Study on mental disorder detection via social media mining," in *2019 4th International Conference on Computing, Communications and Security (ICCCS)*, 2019, pp. 1–6.

[28] B. Alkouz, Z. A. Aghbari, and J. H. Abawajy, "Tweetluenza: Predicting flu trends from Twitter data," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 273–287, 2019.

[29] A. Khatua, A. Khatua, and E. Cambria, "A tale of two epidemics: Contextual Word2vec for classifying Twitter streams during outbreaks," *Information Processing & Management*, vol. 56, pp. 247–257, 2019.

[30] H. Kucukali, O. Atac, A. S. Palteki, A. Z. Tokac, and O. Hayran, "Vaccine hesitancy and anti-vaccination attitudes during the start of COVID-19 vaccination program: A content analysis on Twitter data," *Vaccines*, vol. 10, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2076-393X/10/2/161

[31] U. Tankut, M. Esen, and G. Balaban, "Analysis of tweets regarding psychological disorders before and during the COVID-19 pandemic: The case of Turkey," *Digital Scholarship in the Humanities*, 2021.

[32] J. Phua, S. Jin, and J. Kim, "Gratifications of using Facebook, Twitter, Instagram, or Snapchat to follow brands: The moderating effect of social comparison, trust, tie strength, and network homophily on brand identification, brand engagement, brand commitment, and membership intention," *Telematics and Informatics*, vol. 34, 2016.

[33] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: A set theoretical approach to big data analytics," *IEEE Access*, vol. 4, pp. 2542–2571, 2016.

[34] M. Makita, A. Mas-Bleda, G. Morris, and M. Thelwall, "Mental health discourses on Twitter during mental health awareness week," *Issues in Mental Health Nursing*, vol. 42, 2020.

[35] Kişisel Verilerin Korunması Kanunu. https://www.mevzuat.gov.tr/mevzuatmetin/1.5.6698.pdf. Accessed: 01.05.2022.

[36] L. Wissler, M. Almashraee, D. Monett, and A. Paschke, "The Gold Standard in corpus annotation," 2014.

[37] K. Tomanek, J. Wermter, and U. Hahn, "An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data," 2007, pp. 486–495.

[38] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. USA: Association for Computational Linguistics, 2008, p. 254–263.

[39] S. Alqaraleh and M. Işık, "Efficient Turkish tweet classification system for crisis response," *Turkish Journal of Electrical Engineering & Computer Sciences*, 2020.

[40] A. A. Akın and M. D. Akın. Zemberek-nlp. https://github.com/ahmetaa/zemberek-nlp. Accessed: 01.05.2022.

[41] P. Bhardwaj, "Types of sampling in research," *Journal of the Practice of Cardiovascular Sciences*, vol. 5, p. 157, 2019.

[42] Scikit-learn machine learning in Python. https://scikit-learn.org/stable/. Accessed: 01.05.2022.

[43] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013) Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781. Accessed:01:05:2022.

[44] R. Velioğlu, T. Yıldız, and S. Yıldırım, "Sentiment analysis using learning approaches over emojis for Turkish tweets," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 2018, pp. 303–307.

[45] "Fasttext library for efficient text classification and representation learning," https://fasttext.cc/, Accessed:01:05:2022.

[46] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. https://arxiv.org/abs/1607.04606. Accessed:01:05:2022.

[47] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. (2016) Bag of tricks for efficient text classification. https://arxiv.org/abs/1607.01759. Accessed:01:05:2022.

[48] T. Zhou, Y. Wang, and X. Zheng, "Chinese text classification method using FastText and term frequency-inverse document frequency optimization," *Journal of Physics: Conference Series*, vol. 1693, no. 1, p. 012121, 2020. [Online]. Available: https://doi.org/10.1088/1742-6596/1693/1/012121

[49] Z. Wang and S. Ji, "Learning convolutional text representations for visual question answering," pp. 594–602, 2017.

[50] G. Nergız, Y. Safali, E. Avaroğlu, and S. Erdoğan, "Classification of Turkish news content by deep learning based LSTM using Fasttext model," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–6.

[51] B. Kuyumcu, C. Aksakalli, and S. Delil, "An automated new approach in fast text classification (FastText): A case study for Turkish text classification without pre-processing," in *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, ser. NLPIR 2019. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–4. [Online]. Available: https://doi.org/10.1145/3342827.3342828

[52] O. Celik and B. Koc, "Classification of Turkish news text by TF-IDF, Word2vec and Fasttext vector model methods," *Deu Muhendislik Fakultesi Fen ve Muhendislik*, vol. 23, pp. 121–127, 2021.

[53] H. Yagiz Erdinc and A. Guran, "Semi-supervised Turkish text categorization with Word2vec, Doc2vec and Fasttext algorithms," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, 2019, pp. 1–4.

[54] A. Tommasel and D. Godoy, "Short-text learning in social media: a review," *The Knowledge Engineering Review*, vol. 34, p. e7, 2019.

[55] A. Onan and S. Korukoğlu, "A review of literature on the use of machine learning methods for opinion mining," *Pamukkale Univ Muh Bilim Derg*, vol. 22, no. 2, pp. 111–122, 2016, doi: 10.5505/pajes.2015.90018. [Online]. Available: https://dx.doi.org/10.5505/pajes.2015.90018

[56] S. Kaddoura, G. Chandrasekaran, D. Popescu, and J. Duraisamy, "A systematic literature review on spam content detection and classification," *PeerJ Computer Science*, vol. 8, p. e830, 2022.