JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Medicine Suggestion System from User Review Using NLP Techniques

[1]Ms. Nikila GS, [2]Dr. J Hannah Monisha

[1]Student, [2]Head & Assistant Professor
[1, 2] Department of Computer Science and Engineering,
[1, 2] Vellore Institute of Technology, Indra Gandhi college of Arts and Science

Abstract: Ever since the worldwide spread of the corona virus there has been a lot of issues regarding the non-availability of clinical resources such as health care workers, equipment and medicines. The entire population of medical professionals are in a predicament as to identifying the solution for the ongoing raging crisis. Due to these unfortunate circumstances, individuals resolved to take their own medication without proper knowledge and consultation which has made health conditions even worse and inevitably lead to the death of many people. Presently and in the future most of the applications assess Machine Learning, and there is an increase in innovative work for automation. This paper mainly confers a Medicine Suggestion system that can drastically reduce specialists' heap. In this paper a medicine recommendation system has been built, which takes the patient reviews for sentiment prediction using some vectorization process like Bow, TF-IDF, Word2Vec and Manual Feature Analysis which will help the patients to classify and suggest the top medicine by using different classification algorithm for a particular or a given disease. Precision, recall, flscore, accuracy and AVC Score are used to evaluate the predicted sentiments. The result analysis shows that Classifier Linear SVC using TF-IDF Vectorization is the best among all the other models with 93% accuracy.

Index Terms: *Medicine, Suggestion System, Machine Learning, Bow, TF-IDF, Word2Vec, Sentiment analysis.*

I. Introduction

The entire world is dealing with a shortage of medical professionals or doctors with the increase in corona virus cases especially in rural areas compared to urban areas. Nowadays, clinical disasters are very common. Due to mistakes in the prescription, over 200 thousand in china and 100 thousand people in USA are affected every year. Patients need to choose a top-level medication which is suitable for them to use, which is suggested by a specialist who has knowledge about microscopic organisms and antibacterial medication. New study accompanying drugs which can be accessed by the clinical staffs every day. On the other hand, it has become more complicated and challenging for the doctors to prescribe medications for their patients depending on their symptoms and previous health records.

The population in general buy items depending on their reviews online. Most of the individuals are very much worried about their health and are trying to self-diagnose depending on the information on the internet. According to Pew American Research, almost 60% of the grown-up individuals searched for health-related problems and around 35% of them searched for its diagnosis online. The goal of medication recommender framework is that it can assist specialists and help patients to gain knowledge of specific medicines for particular health conditions. Depending upon the necessity, the recommender framework provides an item to the user. These frameworks do the survey based on sentiments and recommend a solution for their need.

In this medicine suggestion system medicines are provided based on certain conditions, considering patient reviews using sentiment analysis and feature engineering. Sentiment analysis are the strategies, methods and tools that are used to differentiate and extract emotional data like, opinion and attitudes. Feature engineering is used to improve the performance of existing models.

II. RELATED WORKS:

Recommendation system principles are predicted since the mid-1990s, and various other frameworks were created in-order to build large applications. Some of the recommender system unit comprises of e-government area, e-business area, e-commerce/e-shopping area, e-learning area etc. Through online social networking, communication is highly improved, and completely unique information is showcased on the internet at the open pace. Different information must be shared in order to highlight records of potential edges and the availability of utilities bits of knowledge, people practices and items, and so on. The health-related content shared through online feedbacks or surveys contains some assumption designs that emerge from completely different sources in the medical world and provide benefits to the medical industry.

Among this, the online system has become extremely popular for various types of transactions like shopping or different social activities through various websites, and on-line purchasing of medicines.

There are multiple websites that asks clients or users to provide feedback based on their experience in the form of reviews and ratings. These reviews or ratings help websites to improve and helps other clients to choose a particular item.

III. METHODOLOGY:

The dataset used in this research is Medicine Review Dataset (Drugs.com) taken from the UCI ML repository. This dataset contains six attributes, name of drug used (text), review (text) of a patient, condition (text) of a patient, useful count (numerical) which suggest the number of individuals who found the review helpful, date (date) of review entry, and a 10-star patient rating (numerical) determining overall patient contentment. It contains a total of 215063 instances. Fig. 1 shows the proposed model used to build a medicine recommender system. It contains four stages, specifically, Data preparation, classification, evaluation, and Recommendation.

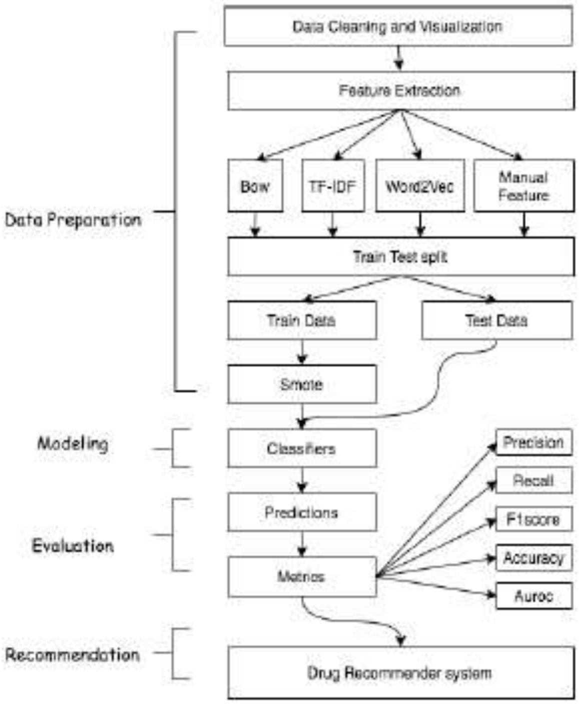


Fig 1 Flowchart of the proposed model

DATA CLEANING AND VISUALIZATION:

The graph (Fig 2) below represents a graphical view of the number of medicines present for a given condition. The total count ranges from approximately a maximum of 200 medicines for pain and a minimum of 35 for obesity. From the given bar graph, it is evident that there are 38 conditions present in the dataset used with a wide variety of medicines for every condition.

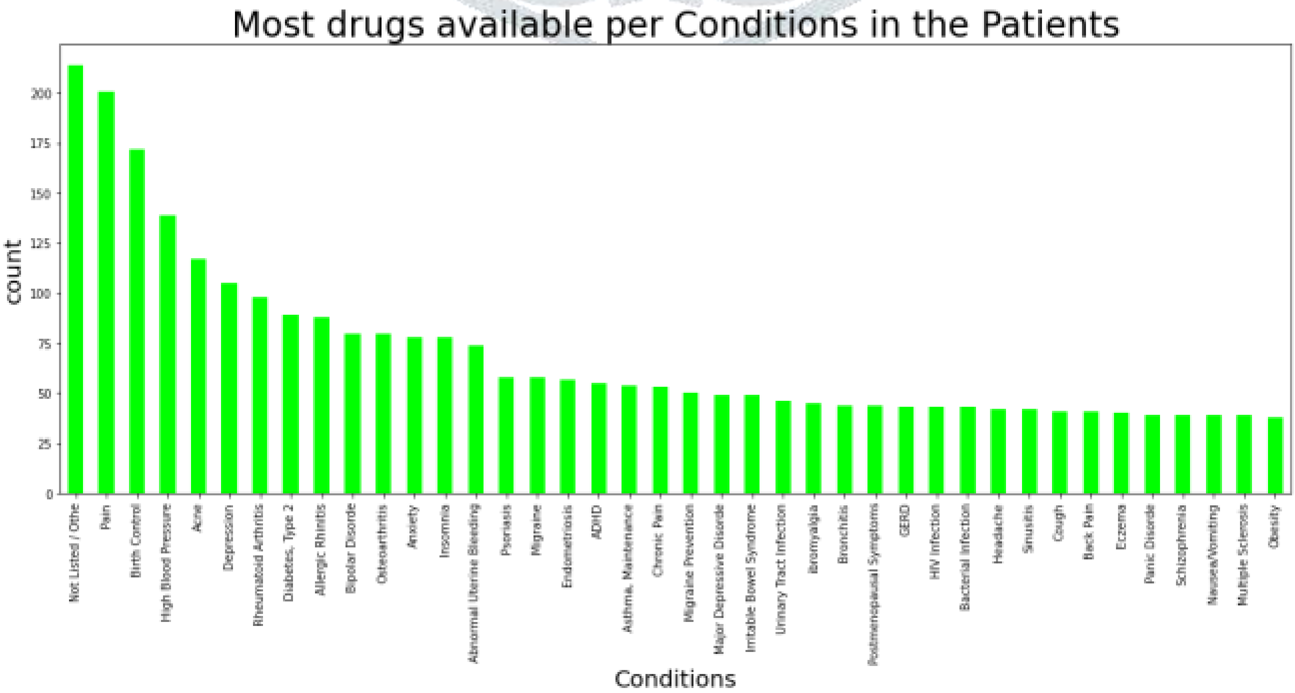


Fig 2 shows the top 40 conditions that have a maximum number of medicines available.

The graph (Fig 3) below represents the total number of ratings for every range (1-10). From the visualization it can be witnessed that majority of the ratings are on the positive side in comparison to the negative end. From this, it can be inferred that the reviews in the dataset are predominantly positive.

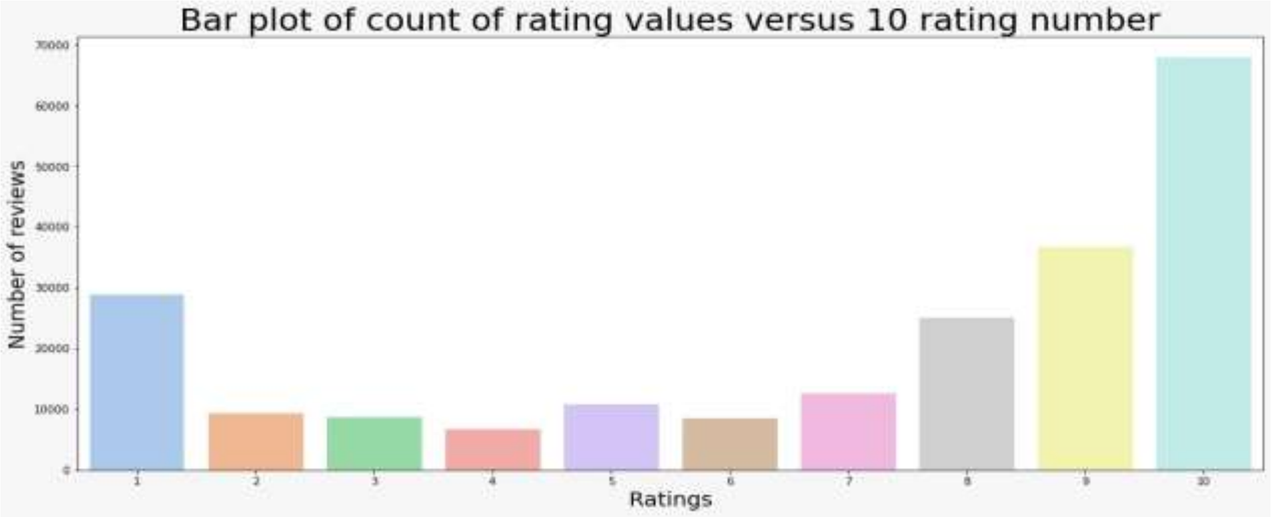


Fig 3 shows the number of patients who gave reviews on medicines.

This graph (Fig 4) depicts the sentiment analysis of the reviewers over the years. From the graph we can notice that, there is a growing trend of negative reviews over the years, with the year 2017 having the maximum number of negative reviews while 2008 has the minimum number.

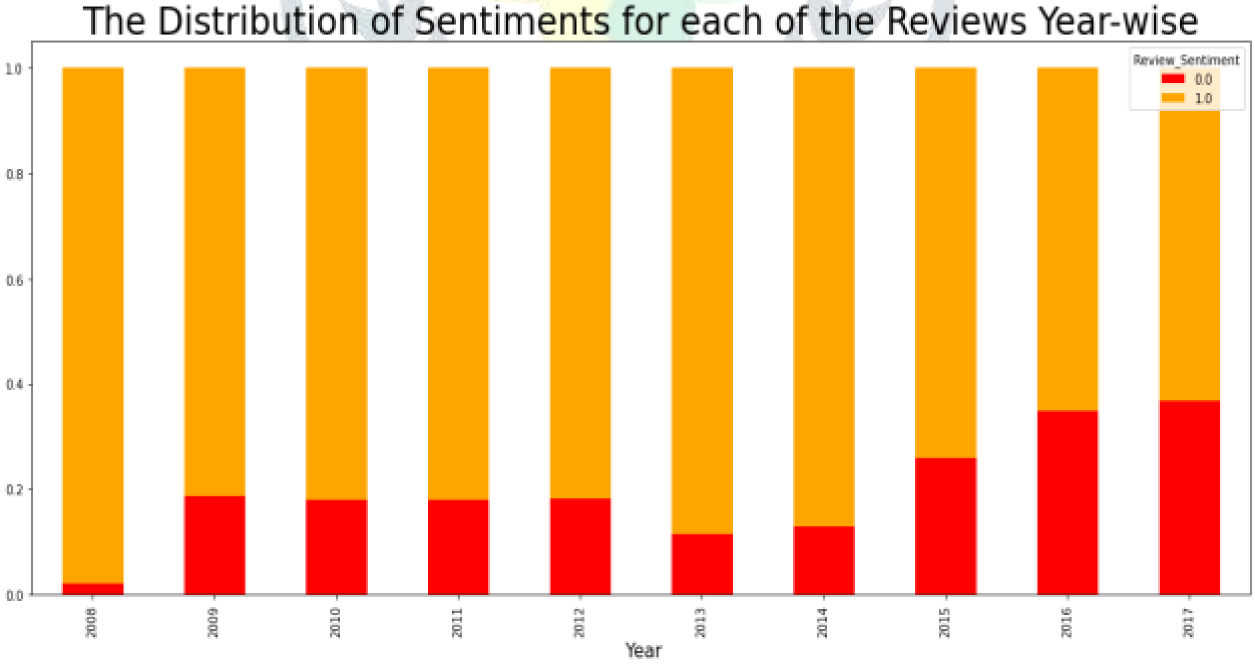


Fig 4 shows year wise reviews on medicines.

IV. PROPOSED METHOD

1) Machine Learning

Machine Learning is a self-learning algorithm which improves rapidly with the help of experience and data. It is considered to be a part of Artificial Intelligence. Machine learning algorithms make predictions based on a “training data” which is built by the model using sample data. Machine Learning algorithms finds its application in a vast range like, email filtering, speech recognition, and computer vision, where it is impossible or complicated to develop usual algorithms to perform various tasks. Computational statistics is a closely related subset of machine learning which mainly focuses on making predictions using computers, but not all machine learning is statistical learning. Mathematical optimization plays a huge role in the field of machine learning by providing methods, theory and application domains. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Working of a biological brain can be mimicked through implementations of machine learning data and neural networks. In the application of machine learning to business problems, it is known as predictive analysis.

Train Test Split:

Four different datasets were created using Bow, TF-ID, Word2Vec and manual features. These four datasets were split in a ratio of 3:1 of training and testing data respectively. An equal random state was set to make sure that all four generated datasets have the same set of random numbers generated for train test split.

2)Neural Networks

Computing systems that are inspired by the biological neural networks that constitute the brain are called Artificial Neural Networks (ANNs) or simply Neural Networks (NN). An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. A signal to the other neurons can be transmitted by a signal like the synapses in a biological brain. An artificial neuron receives a signal after which it processes it and further transmits it to signal neurons that are connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. As the learning proceeds edges and neurons that adjusts as the learning continues. The weight of a signal determines the strength of the connection. A threshold might be associated with a neuron such that a signal is sent only if the aggregate signal crosses that threshold. In general, neurons are aggregated into layers in which different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

2) Natural Language Processing

Natural Language Processing (NLP) is typically used to process and analyze huge amounts of natural data as it is mainly concerned with the interactions between computers and human language. NLP is considered to be a subset of computer science, linguistics and artificial intelligence. The main aim of using NLP is to produce a program which is capable of comprehending the contents of the documents, including the textual nuances of the languages within them. By using this program the technology will be able to accurately extract information and insights present in the documents as well as organize and categorize the documents automatically.

TF-IDF

A weighing strategy in which words are offered weight and not count is TF-IDF. The goal is to use TF-IDF to estimate relevance, not accuracy by giving low importance to the terms that frequently occur on the dataset. The likelihood of locating a word in the document is called Team Frequency (TF).

$$tf\ (t,d) = log(1 + freq(t, d))$$

Inverse document frequency (IDF) is the opposite of the number of times a specific term showed up in the whole corpus. It catches how a specific term is document specific.

$$idf(t,d) = log(\frac{N}{count(d\epsilon D : t\epsilon d)})$$

TF-IDF is the multiplication of TF with IDF, suggesting how vital and relevant a word is in the document.

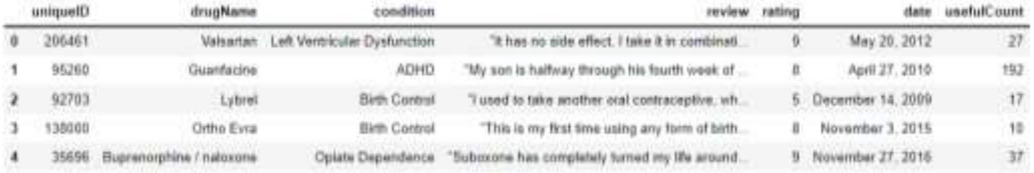
$$tfidf(t,d,D) = tf\ (t,\ d).idf(t,D)$$

Word2Vec

Semantic and syntactic likeliness between words are disregarded by TF and TF-IDF despite them being the most popular methods used in different natural language preparing tasks. To exemplify, both lovely and delightful are called two unique words in both TF and TF-IDF vectorization techniques even though they are almost synonyms. Word2Vec is a model used to produce word embedding. Gargantuan corpora using multiple deep learning models reproduce word embeddings. Word2Vec takes an enormous corpus of text as an input and outputs a vector space, generally composed of hundred dimensions. The basic ideology was to take the semantic meaning of the words and arrange vectors of words in vector space such that the words that are similar in meaning are found close to one another in vector space.

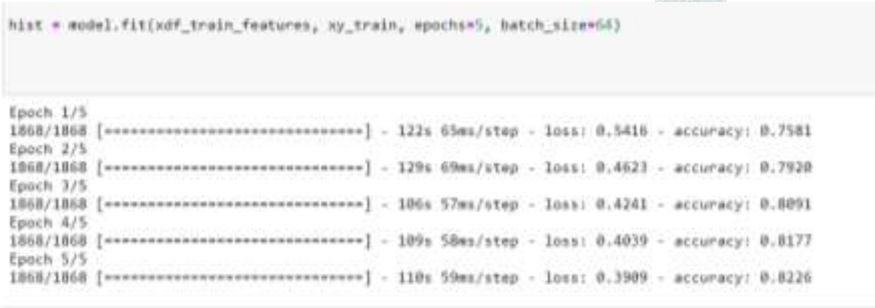
V. Result analysis:

Given below is the format of the dataset which is being used for the analysis. The dataset of 2 million records have been obtained from online. Based on the data set, NLP algorithms and Artificial Neural Network (ANN) algorithms have been implemented to obtain the desired result.



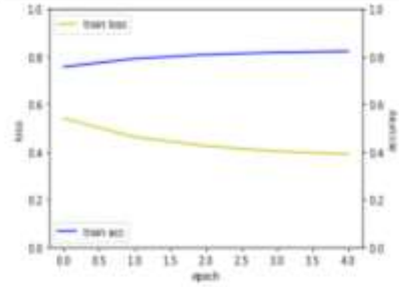
	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect. I take it in combinati...	9	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	November 27, 2016	37

The given below figure shows the accuracy obtained after fitting the model. We have considered a total of 5 epochs, it can be visibly noticed that for every epoch the train loss keeps reducing while the accuracy increases. A maximum of 82.26% accuracy has been obtained from the model.

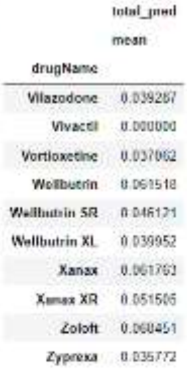


```
hist = model.fit(xdf_train_features, xy_train, epochs=5, batch_size=64)
```

```
Epoch 1/5
1868/1868 [==============================] - 122s 65ms/step - loss: 0.5416 - accuracy: 0.7581
Epoch 2/5
1868/1868 [==============================] - 129s 69ms/step - loss: 0.4623 - accuracy: 0.7920
Epoch 3/5
1868/1868 [==============================] - 106s 57ms/step - loss: 0.4241 - accuracy: 0.8091
Epoch 4/5
1868/1868 [==============================] - 109s 58ms/step - loss: 0.4039 - accuracy: 0.8177
Epoch 5/5
1868/1868 [==============================] - 110s 59ms/step - loss: 0.3909 - accuracy: 0.8226
```



The below figure depicts the result obtained from the model for a given condition. The condition which is input for this result is "depression", for which the top 10 drugs are recommended by the system.



	total_pred
	mean
drugName	
Vilazodone	0.039267
Vivactil	0.000000
Vortioxetine	0.037062
Wellbutrin	0.061518
Wellbutrin SR	0.046171
Wellbutrin XL	0.039952
Xanax	0.061763
Xanax XR	0.051505
Zoloft	0.060451
Zyprexa	0.035772

VI. CONCLUSION

The system suggests medicine for a particular medical condition based on the user's input. The system deploys Natural language Processing (NLP) and Artificial Neural Network(ANN) methodologies to fit and train the model based on the review ratings, review sentiment and the useful count. This system can be used majorly by people in need of a medicine for a particular condition immediately. The system ensures safe and reliable reviews from a huge range of dataset. This will prove to be beneficial during the times of emergency and shortage of trained medical practitioners. Since the system recommends multiple drugs based on rating, it will the users can also use this to look up for the next best medicine for the condition in case of shortage of vaccines and medicines.

REFERENCES

[1]H. Arshad, A. Jantan, and E. Omolara, "Evidence collection and forensics on social networks: Research challenges and directions," Digit. Invest., vol. 28, pp. 126–138, Mar. 2019.

[2] S.Seo et al., "Partially generative neural networks for gang crime classification with partial information," in Proc. AAAI/ACM Conf. AI, Ethics, Soc., New York, NY, USA, 2018, pp. 257–263, doi: 10.1145/3278721.3278758.

[3] V D. Ramalingam, V. Chinnaiah, and A. Jeyagobi, “Privacy preserving schemes for secure interactions in online social networks,” in Proc. Int. Conf. Soft Comput. Syst., vol. 837, 2018, pp. 548–557.

[4] S. Jiang, M. Duan, and L. Wang, “Toward privacy-preserving symp- toms matching in SDN-based mobile healthcare social networks,” IEEE Internet Things J., vol. 5, no. 3, pp. 1379–1388, Jun. 2018, doi: 10.1109/JIOT.2018.2799209.

[5] K. Zhang, J. Ni, K. Yang, X. Liang, J. Ren, and X. S. Shen, “Security and privacy in smart city applications: Challenges and solu- tions,” IEEE Commun. Mag., vol. 55, no. 1, pp. 122–129, Jan. 2017, doi: 10.1109/MCOM.2017.1600267CM.

[6] R. Yu, J. Kang, X. Huang, S. Xie, Y. Zhang, and S. Gjessing, “MixGroup: Accumulative pseudonym exchanging for location privacy enhancement in vehicular social networks,” IEEE Trans. Depend. Secure Comput., vol. 13, no. 1, pp. 93–105, Jan./Feb. 2016.

[7] B. Desmet and V. Hoste, “Online suicide prevention through optimised text classification,” Inf. Sci., vol. 439, pp. 61–78, May 2018.

[8] K. Zhang, X. Liang, J. Ni, K. Yang, and X. Shen, “Exploiting social network to enhance human-to-human infection analysis without pri- vacy leakage,” IEEE Trans. Depend. Sec. Comput., vol. 15, no. 4, pp. 607–620, Jul./Aug. 2018, doi: 10.1109/TDSC.2016.2626288.

[9] B. Desmet and V. Hoste, “Online suicide prevention through optimised text classification,” Inf. Sci., vols. 439–440, pp. 61–78, May 2018, doi: 10.1016/j.ins.2018.02.014.

[10] Z. Yu, F. Yi, Q.Lv, and B. Guo, “Identifying on-site users for social events: Mobility, content, and social relationship,” IEEE Trans. Mobile Comput., vol. 17, no. 9, pp. 2055–2068, Sep. 2018, doi: 10.1109/TMC.2018.2794981.

[11] A. Tundis, A. Jain, G. Bhatia, and M. Muhlhauser, “Similarity analysis of criminals on social networks: An example on Twitter,” in Proc. 28th Int. Conf. Comput. Commun. Netw. (ICCCN), Valencia, Spain, Jul./Aug. 2019, pp. 1–9, doi: 10.1109/ICCCN.2019.8847028.

[12] G. Rigopoulos and N. V. Karadimas, “Military student assignment using NexClass decision support system,” in Proc. 3rd Int. Conf. Math. Comput. Sci. Ind. (MCSI), Chania, Greece, Aug. 2016, pp. 213–218, doi: 10.1109/MCSI.2016.047.

[13] V. Ingilevich and S. Ivanov, “Crime rate prediction in the urban environ- ment using social factors,” Procedia Comput. Sci., vol. 136, pp. 472–478, Jan. 2018.

[14 B. R. Prathap and K. Ramesha, “Twitter sentiment for analysing different types of crimes,” in Proc. Int. Conf. Commun., Com- put. Internet Things, Chennai, India, Feb. 2018, pp. 483–488, doi: 10.1109/IC3IoT.2018.8668140.

[15] X. Sun, P. Zhang, J. K. Liu, J.Yu, and W. Xie, “Private machine learning classification based on fully homomorphic encryption,” IEEE Trans. Emerg. Topics Comput., to be published, doi: 10.1109/TETC.2018.2794611.

[16]M. Abadi et al., “Deep learning with differential privacy,” in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., New York, NY, USA, 2016, pp. 308–318, doi: 10.1145/2976749.2978318.

[17] O. Ohrimenko et al., “Oblivious multi-party machine learning on trusted processors,” in Proc. USENIX Secur., vol. 16, 2016, pp. 619–636.

[18] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, “Machine learning classification over encrypted data,” in Proc. NDSS, 2015.

[19] H. Hassani, X. Huang, M. Ghodsi, and E. S. Silva, “A review of data mining applications in crime,” Stat. Anal. Data Mining, ASA Data Sci. J., vol. 9, no. 3, pp. 139–154, Apr. 2016, doi: 10.1002/sam.11312.

[20] D. J. Wu, T. Feng, M. Naehrig, and K. Lauter, “Privately evaluating decision trees and random forests,” in Proc. Privacy Enhancing Technol., vol. 4, pp. 335–355, 2016.

[21] R. K. H. Tai, J. P. K. Ma, Y. J. Zhao, and S. S. M. Chow, “Privacy- preserving decision trees evaluation via linear functions,” in Proc. Eur. Symp. Res. Comput. Secur. (Lecture Notes in Computer Science), vol. 10493. Berlin, Germany: Springer, 2017, pp. 494–512.

[22] M. Joye and F. Salehi, “Private yet efficient decision tree evaluation,” in Data and Applications Security and Privacy XXXII. Berlin, Germany: Springer, 2018, pp. 243–259, doi: 10.1007/978-3-319-95729-6_16.

[23] T. Veugen, “Improving the DGK comparison protocol,” in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), Tenerife, Spain, Dec. 2012, pp. 49–54, doi: 10.1109/WIFS.2012.6412624.

[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. New York, NY, USA: Chapman And Hall, 1993.

[25] P. Paillier and D. Pointcheval, “Efficient public-key cryptosystems provably secure against active adversaries,” in Proc. Int. Conf. Theory Appl. Cryptol. Inf. Secur., vol. 1999, pp. 165– 179.