future internet

Article

A Classifier to Detect Informational vs. Non-Informational Heart Attack Tweets

Ola Karajeh [1,*], Dirar Darweesh [2], Omar Darwish [3], Noor Abu-El-Rub [4], Belal Alsinglawi [5] and Nasser Alsaedi [6]

[1] Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[2] Department of Computer Science, Jordan University of Science and Technology, Irbid 3030, Jordan; derardrweesh@gmail.com
[3] Computer Technology and Information Systems, Ferrum College, Ferrum, VA 24088, USA; odarwish@ferrum.edu
[4] Kansas Medical Center, Kansas City, MO 67002, USA; nabuelrub@kumc.edu
[5] School of Computer Data and Mathematical Sciences, Western Sydney University, Rydalmere, NSW 2116, Australia; b.alsinglawi@westernsydney.edu.au
[6] Department of Computer Science, Taibah University, Medina 2003, Saudi Arabia; nsaede@taibahu.edu.sa
* Correspondence: okarajeh@vt.edu

Citation: Karajeh, O.; Darweesh, D.; Darwish, O.; Abu-El-Rub, N.; Alsinglawi, B.; Alsaedi, N. A Classifier to Detect Informational vs. Non-Informational Heart Attack Tweets. *Future Internet* **2021**, *13*, 19. https://doi.org/10.3390/fi13010019

Received: 14 December 2020
Accepted: 14 January 2021
Published: 16 January 2021

Abstract: Social media sites are considered one of the most important sources of data in many fields, such as health, education, and politics. While surveys provide explicit answers to specific questions, posts in social media have the same answers implicitly occurring in the text. This research aims to develop a method for extracting implicit answers from large tweet collections, and to demonstrate this method for an important concern: the problem of heart attacks. The approach is to collect tweets containing "heart attack" and then select from those the ones with useful information. Informational tweets are those which express real heart attack issues, e.g., "Yesterday morning, my grandfather had a heart attack while he was walking around the garden." On the other hand, there are non-informational tweets such as "Dropped my iPhone for the first time and almost had a heart attack." The starting point was to manually classify around 7000 tweets as either informational (11%) or non-informational (89%), thus yielding a labeled dataset to use in devising a machine learning classifier that can be applied to our large collection of over 20 million tweets. Tweets were cleaned and converted to a vector representation, suitable to be fed into different machine-learning algorithms: Deep neural networks, support vector machine (SVM), J48 decision tree and naïve Bayes. Our experimentation aimed to find the best algorithm to use to build a high-quality classifier. This involved splitting the labeled dataset, with 2/3 used to train the classifier and 1/3 used for evaluation besides cross-validation methods. The deep neural network (DNN) classifier obtained the highest accuracy (95.2%). In addition, it obtained the highest F1-scores with (73.6%) and (97.4%) for informational and non-informational classes, respectively.

Keywords: machine learning; classification; support vector machine; deep neural networks; tweets; heart attack; health

1. Introduction

Social media networks like Facebook, Myspace and Twitter, are considered one of the most important methods of communication among people [1]. Nowadays, social media has been developed during the recent decade to form an important tool to gather information and build solutions in several fields such as business, entertainment and crisis management in health care, science and politics [2]. Social media are considered one of the significant sources for extracting information related to health monitoring [3,4]. Therefore, in this research project, we are interested in collecting information related to heart attack problems from social media sites. Many people use social media to share information related to health issues [5]. For example, Twitter is considered one of the largest resources used by people on a daily basis to post their updates and life events.

While there are many social media networks nowadays, the most recent research has focused on Twitter for knowledge discovery data-mining, and tweets semantic analysis [6,7] for the following reasons: First, Twitter has a limitation of 280 characters for each message, which enables quick posting of activities and updates. Consequently, this allows messages to be easily shared and forwarded to other users in the network. This feature causes information and news to rapidly spread over the Internet. The ease of access to people updates is the second reason which makes Twitter a well-known network for crawling and collecting datasets. While in most social media networks, mutual friendship is required to access other profiles, Twitter allows users to follow each other without this restriction [8]. The third reason is Twitter exploits the use of hashtags, which are labels used within social networks to make it easier for users to search for text with specific content or topic. This facilitates the search for tweets with specific subjects using hashtags [9].

Our research project is focusing on the heart attack field. While surveys provide explicit answers to specific questions, posted texts in social media have the same answers implicitly occurring in the text. Therefore, Twitter is used in our research project for extracting information regarding the heart attack problem. The estimated annual incidence of heart attacks in the United States is 720,000 new attacks [10]; hence different science disciplines should give a hand to those working in the medical field to study the heart attack problem. Computer science is one of them. Therefore, this research aims to construct a dataset of tweets with information related to the heart attack problem to help shed light on this issue. Useful information could be *when* or *where* the heart attack has occurred (Time or place). Having such information helps doctors to decide what is the best time or place for the patient to get his/her medicines. We will not be able to get such useful information unless we filter out the data into informational vs. non-informational tweets.

For this work, to help identify the heart attack problem from Twitter data, we define two types of tweets: informational and non-informational tweets. Informational tweets are those which express real heart attack issues, such as "Yesterday morning, my grandfather had a heart attack while he was walking around the garden". On the other hand, non-informational tweets are tweets that are not related to heart attacks, even if they have words that indicate that, such as, ”Dropped my iPhone for the first time and almost had a heart attack". Machine learning (ML) offers different techniques, algorithms, and tools that can support solving diagnostic and prognostic problems in many medical fields [9]. Our goal in this research resides in comparing the ability of well-known machine-learning algorithms (DNN, J48 decision trees, naïve Bayes, and SVM) in classifying informational tweets vs. non-informational tweets. These classifiers will be compared in terms of accuracy, precision, recall and F1-score measures. Then, we will choose the one with the highest accuracy, precision, recall and F1-score. The main contributions in this paper are summarized as follows:

1. We generate and manually label around 7000 tweets in the heart attack domain into informational and non-informational;
2. We compare the performance of several machine learning models (DNN, J48 decision trees, naïve Bayes, and SVM) in terms of accuracy, precision, recall and F1-score measures on the annotated dataset.

The rest of this paper is structured as follows: Section 2 describes a brief review of the recent related literature, Section 3 provides the proposed method, Section 4 illustrates our obtained results, Section 5 discusses the results and limitations, and finally, Section 5 concludes this paper and provides future research directions.

2. Related Work

In this section, we discuss the literature review related to mining social media and health care. The first section review some of the papers that utilized social media in the health domain, the second section will give an insight about researchers who applied data-mining techniques on social media data for the purpose of classification in the health field.

2.1. Social Media in the Health Domain

In the context of social media analytics, some research proposed frameworks such as in [11–13] illustrating the growing interest of people in social media in the domain of healthcare. Although they addressed the ethical and confidential problems related to the collected health data from social media, they emphasized the importance of this huge amount of data in the health domain. The authors pointed out the fact that such social media data have the feasibility to be analyzed to obtain valuable rules, insights and patterns which will help future health research. However, a heart attack is a critical health condition that has not been addressed specifically in the aforementioned studies, albeit, public health domain, in general, was the main concern. Furthermore, the previous attempts were limited to analyzing the state of the social media analytics in healthcare and providing general guidelines for researchers in the fields; it did not put in the consideration a practical framework to attest the usefulness of tweets analytics in the health domain; In particular, addressing the heart attack condition based on social media analysis.

2.2. Mining Social Media for Health Care and Diseases

On the subject of mining social media data for health care and disease analytics, some research attempts such as in [14–17] utilized data-mining on social media to help to monitor people's health and for marketing purposes. However, most of the attempts did not focus on analyzing and extracting knowledge from informational tweets for specific diseases or health conditions.

On the other hand, other attempts by [18,19] used social media tweets to predict rates of heart attack with the use of sentiment analysis methods where tweets were analyzed based on negative and positive emotions. Nonetheless, the attempts were lack of insights and knowledge into mining tweets of informational and non-informational based on heart attack according to the proposed approach in our research.

3. Methodology

The methodology of our research project consists of seven stages. These phases are building up the corpus, cleaning, remove redundancy, word embedding, building term-document matrix (TDM), dimensionality reduction and classification using different data-mining techniques (DNN, J48, naïve Bayes, and SVM). The flow chart of the proposed methodology for these stages is shown in Figure 1. These stages in Figure 1 are explained in detail in the following sections.

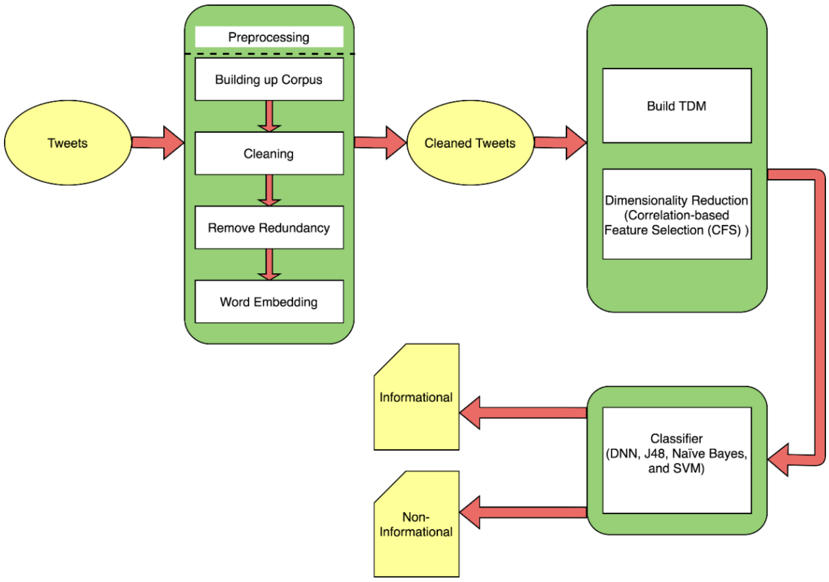

Figure 1. Proposed approach flowchart.

3.1. Building Up the Corpus

This is the first stage in our research project in which a dataset of tweets that include heart attack keyword was crawled. The dataset consists of 7000 tweets. The tweets are categorized into informational and non-informational classes. Informational tweets are those which are related to real heart attack issues, while non-informational are tweets that do not indicate a real heart attack problem. The percentages of informational and non-informational tweets in our dataset are 11% and 89%, respectively.

3.2. Cleaning

The second stage in our approach is data cleaning. Preprocessing and cleaning the text is an important step before creating the feature sets and building predictive models. This stage includes the following steps:

- Removal of punctuations;
- Removal of retweet (RT);
- Removal of hashtags;
- Removal of stop-words;
- Removal of URLs;
- Stemming (Porter stemmer);
- Convert all letters to lower case.

We demonstrate the cleaning process using the following example.

- Original Tweet: "My aunt had a heart attack in church today they thought she just passed out in the spirit Keep my family in your prayers tweeps."
- After cleaning: "aunt church today thought pass spirit keep family prayer tweep"

3.3. Redundancy Removal

The training time has a strong relation to the size of the dataset used for classification, so it is necessary to remove redundancy from the training dataset in order to enhance the speed of prediction and decrease the time required to build the model used for prediction. By removing redundancy, the number of tuples that existed in the training dataset is minimized, so we will achieve less training time with competitive accuracy. In this phase, we have removed the redundant tweets in the dataset before proceeding to the next step.

3.4. Word Embedding

Word embedding is considered one of the most used techniques in natural language processing. The objective of word embedding is to reduce the size of the dimensional representation of a word in a text corpus. Word embedding presents a more efficient and expressive representation for words by creating a small vector representation. Word2Vec is a well-known technique used for creating word embedding. It is a word embedding model stated by Mikolov; this model predicts words according to their context using CBOW or Skip-Gram-neural models [20]. In our research project, we implement word embedding using the Word2Vec model.

3.5. Building TDM

A term document matrix (TDM) is defined as a mathematical matrix that demonstrates the frequency of terms that occurs in documents. It consists of columns and rows where rows are associated with documents and columns are represented by terms. The values for this matrix are determined using several schemes [21]. After applying word embedding to our dataset, tweets are converted to a TDM, suitable to be fed into different machine-learning algorithms. We used the R programming language as a tool for generating the TDM, which takes either 0 or 1 as values for its entries.

In this paper, each tweet is considered as a single document in the process of building the TDM. However, we utilized word embedding (Word2Vec) to find words that have a similar meaning in the corpus. Words with similar meanings are replaced by just one of them. Based on that, the number of words in the corpus is decreased, and TDM dimensionality even reduced before using correlation-based feature selection (CFS) [22]. For example, word embedding shows that the following words are similar in meaning "Mom", "Mama", "Mother", and "Mammy". Then, we replace all of these words in the corpus with one of them, which can be the word "Mom". Consequently, three attributes are eliminated from the TDM. Tweets are obtained from the Digital Library Research Laboratory (DLRL) [23] at Virginia Tech University, and the full TDM dataset is publicly available in the following link https://github.com/okarajeh/Heart-Attack-Tweets-TDM.

3.6. Dimensionality Reduction

Dimensionality reduction is a leading feature transformation technique in machine learning. It facilitates the classification, compression, and visualization tasks for large dimensional datasets. It is done by eliminating non-informative features in high dimensional spaces. The CFS is a metric to evaluate the efficacy of feature subsets. The core idea of CFS is that a good feature subset has features that are highly correlated to the output class and independent from each other. Thus, the goal is to reduce feature-to-feature correlation (*r_ff*) and increase feature-to-class correlation (*r_fc*). The CFS metric is defined using Pearson's coefficient, which is calculated based on the ratio: *r_fc*/*r_ff*, where a higher ratio indicates a better subset of features according to Equation (1).

$$CFS = \max_{S_k}\left[\ \frac{r_{cf1}+r_{cf2}+\ldots+r_{cfk}}{\sqrt{k+2\left(r_{f1f2}+\ldots+r_{fifj}+\ldots+r_{fkf1}\right)}}\ \right] \tag{1}$$

3.7. Classification

This is the last step in our research project, where we are interested in discovering patterns and rules using machine-learning methods. The well-known tool (Weka) is used for the classifiers SVM, J48 and naïve Bayes. In contrast, the Sklearn library from Python is used for the DNN classifier. Thus, all of these machine-learning algorithms are applied to our labeled dataset. Features learning in DNN differs from the traditional classifiers used in this study. While features engineering is a manual process in the traditional classifiers (shallow learning), the DNN consists of layers, and each layer has a number of neurons that mimic the functionality of the neurons in the human brain. Hence, DNN has the ability to automatically learning features [24]. Noting that we have used three layers, with 10 neurons in each layer. The well-known algorithm, which is called the C4.5 decision tree, is implemented in java to form a J48 classifier. J48 algorithm basically uses the gain ratio measure for selecting its attributes. In this classifier, features that have high gain ratio values have a higher power in the tree and acquires their top levels. SVM is considered as one of the famous binary techniques used in classification. It relies on locating the best hyper-plane, which can split various classes from each other. naïve Bayes is a probabilistic technique that depends on Bayes theorem. This algorithm supposes that all attributes values are independent. Naïve Bayes has competitive performance in classification regardless of naive supposition [25].

We divided the data into 66% for training and 34% for testing. We use K-cross validation to train the model and ensure a robust model among different samples. We used different k values: 2, 5 and 10 and reported the results for all the experiments in the next section.

4. Results

Figure 2 represents the performance for the machine learning techniques (DNN, SVM, naïve Bayes and J48 decision tree) models in terms of accuracy before and after applying the CFS reduction method. We notice that DNN has the highest accuracy for all testing types. It obtains 95.2% and 94% before and after applying CFS reduction, respectively. J48 has the lowest accuracy with approximately 91% for all testing types and for both cases before and after applying CFS reduction. Figure 3 demonstrates the results of the machine-learning algorithms (DNN, SVM, naïve Bayes and J48 decision tree) in terms of the

F1-score for informational class. As we notice in Figure 3, DNN has the highest F1-score for informational class with (10-fold) cross-validation where it obtains (73.6%) before applying CFS reduction and with (66%) split it obtains (67%) after applying CFS reduction. J48 shows the lowest F1-score with 2-fold before applying CFS reduction (49.4%) and with a 66% split after applying CFS reduction (33.9%). On the other hand, for the non-informational class, Figure 4 demonstrates that DNN has the highest F1-score with 10-fold cross-validation (97.4%) before applying CFS reduction. DNN also obtains the highest F1-score (97%) with (66%) split after applying CFS reduction. Moreover, J48 has the lowest F1 score with 2-fold cross-validation (95.1) before applying CFS reduction. Finally, naïve Bayes has the lowest F1-score with 5 and 10-fold cross-validation (95%) after applying CFS reduction. More information about precision, and recall are available in Appendix A (Tables A1–A4).

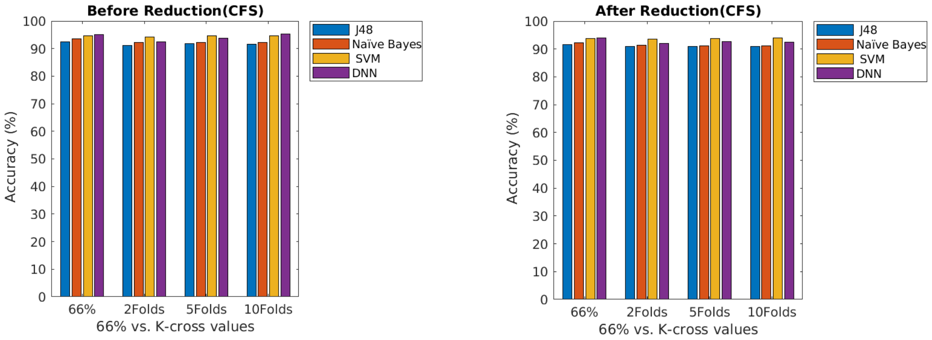

Figure 2. Accuracy based on deep neural network (DNN), support vector machine (SVM), naïve Bayes and J48 models (before reduction and after reduction using (correlation-based feature selection (CFS)) method).

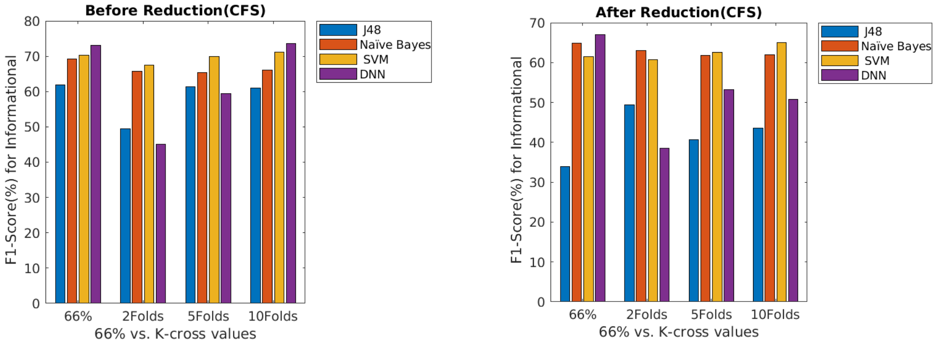

Figure 3. F1-score for informational class based on DNN, SVM, naïve Bayes and J48 models (before reduction and after reduction using (CFS) method).

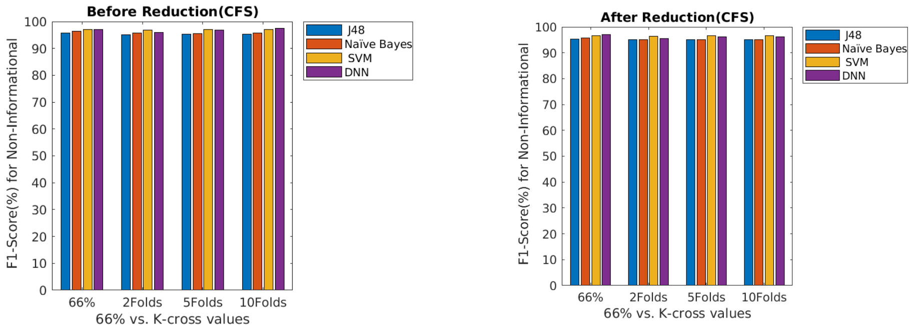

Figure 4. F1-score for non-informational class based on DNN, SVM, naïve Bayes and J48 models (before reduction and after reduction using (CFS) method).

Reduction with CFS shows a significant effect on a number of attributes, and that reflects on the time to build the model. The number of attributes before applying reduction with CFS was 9385 attributes, where it dropped to 242 attributes after the reduction process. Table 1 shows a significant difference in time to build the model before and after applying the CFS reduction technique.

Table 1. The effect of applying reduction with CFS on the time to build the model using the J48 algorithm.

Testing Method	Before CFS Reduction	After CFS Reduction
66%	491.37 s	2.96 s
2-fold	490.76 s	2.47 s
5-fold	490.76 s	2.44 s
10-fold	492.28 s	2.45 s

5. Discussion and Study Limitation

The purpose of this paper was to generate a model that has the ability to extract useful information from a large set of tweets. Our focus in this project was extracting valuable information from tweets that are related to the heart attack problem. We started our research by collecting 7000 tweets. 11% of these tweets were informational, while the rest was non-informational. This sampled tweet collection was classified manually and used as ground truth for the machine learning techniques. The models that were generated from this process will help us in classifying larger datasets of tweets automatically.

The experimental results using the machine learning techniques (DNN, SVM, naïve Bayes and J48 decision tree) indicated that DNN outperformed the other models and achieved the highest accuracy and F1-score. In addition, although the reduction of attributes using CFS slightly decreased the accuracy, it showed a significantly dropping in time to build the model. Moreover, the decrease in accuracy values was really small compared to the significant difference in the time to build the model. The reason behind the small differences in accuracy values before and after reduction was that the CFS method selects the attributes with a high correlation to the output class. However, still, this research has the following remarkable limitations:

1. Twitter was the only platform used in our project, while there are many other platforms in social media such as Facebook and LinkedIn that can be utilized;
2. The unbalanced dataset was a challenge where the percentage of non-informational tweets was higher than the informational tweets. This limitation can be solved by using adaptive learning techniques that avoid the dependency on manual labeling of tweets

However, such limitations will be considered in our future work.

6. Conclusions and Future Work

The goal behind this research is to classify tweets into two categories informational and non-informational. Informational tweets are those which express real heart attack issues, while non-informational ones are not related to heart attacks. For this work, we used different natural language processing techniques to convert tweets to a vector representation (TDM), suitable to be fed into different machine-learning algorithms: DNN, SVM, naïve Bayes and J48 decision tree. The highest accuracy (95.2%) was obtained using DNN, and the highest F1-scores (73.6%) and (97.4%) were also obtained by DNN for informational and non-informational classes, respectively. Although applying the CFS reduction method slightly decreased the accuracy, it showed a significant effect in decreasing the time to build the model. The CFS selects attributes with high correlation to the output class. In future work, we are interested in determining the time of heart attack (morning, noon, or evening) by extracting time-based features from the tweets. In addition, we are interested in the geographic location to determine where the heart attack had occurred (e.g., work, home, or gym). These time-based and geographic features will help to better monitor and predict heart attack incidents using social media. A noteworthy point, we are aiming to extend our work in a future study to evaluate our research framework on a broader selection of machine learning models and architectures.

Author Contributions: Conceptualization, O.K. and O.D.; methodology, O.K. and D.D.; software, O.K. and D.D.; validation, O.D., N.A.-E.-R. and B.A.; formal analysis, O.K.; investigation, O.K. and D.D.; resources, O.D. and N.A.; data curation, O.D. and D.D.; writing—original draft preparation, O.K., D.D.; writing—review and editing, N.A.-E.-R., O.D., and B.A.; visualization, O.K.; supervision, N.A.-E.-R.; project administration, O.D. and N.A.; funding acquisition, O.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Precision and recall values for informational class before attributes reduction.

	Precision				**Recall**			
	66%	**2-fold**	**5-fold**	**10-fold**	**66%**	**2-fold**	**5-fold**	**10-fold**
J48	63.3	62.6	61.1	61	60.5	40.8	61.6	61
Naïve Bayes	66.9	62.7	62	62.8	71.8	69	69.3	69.6
SVM	83.6	85.3	85.5	85.2	60.5	55.8	59.2	61.2
DNN	85	93.5	70.2	86.7	64	30.5	51.6	64.6

Table A2. Precision and recall values for non-informational class before attributes reduction.

	Precision				**Recall**			
	66%	**2-fold**	**5-fold**	**10-fold**	**66%**	**2-fold**	**5-fold**	**10-fold**
J48	95.5	93.2	95.4	95.3	95.8	95.1	95.3	95.3
Naïve Bayes	96.8	96.2	96.3	96.3	96.4	95.7	95.6	95.7
SVM	95.6	94.9	95.3	95.5	97.1	96.8	97	97.1
DNN	96	92.5	94.4	95.9	97	96	96.8	97.4

Table A3. Precision and recall values for informational class after attributes reduction.

	Precision				**Recall**			
	66%	**2-fold**	**5-fold**	**10-fold**	**66%**	**2-fold**	**5-fold**	**10-fold**
J48	82.5	62.1	69.4	65.6	21.4	41.1	28.8	32.6
Naïve Bayes	60.7	58.2	57.9	57.8	69.5	68.5	66.2	66.9
SVM	83.6	84.7	86.7	86.3	48.6	47.4	49	52.1
DNN	90	89	83.2	76.9	53	24.5	41	39.2

Table A4. Precision and recall values for non-informational class after attributes reduction.

	Precision				**Recall**			
	66%	**2-fold**	**5-fold**	**10-fold**	**66%**	**2-fold**	**5-fold**	**10-fold**
J48	91.7	93.2	92	92.4	99.5	97	98.5	97.9
Naïve Bayes	96.5	96.1	95.9	96	94.9	94.1	94.2	94.1
SVM	94.4	94	94.2	94.5	98.9	99	99.1	99
DNN	95	92	93.2	93	99	100	98.6	98.9

References

1. Shahare, F.F. Sentiment analysis for the news data based on the social media. In Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 15–16 June 2017; pp. 1365–1370.
2. Stieglitz, S.; Mirbabaie, M.; Ross, B.; Neuberger, C. Social media analytics—Challenges in topic discovery, data collection, and data preparation. *Int. J. Inf. Manag.* **2018**, *39*, 156–168. [CrossRef]
3. Zhao, Y.; Zhang, J. Consumer health information seeking in social media: A literature review. *Health Inf. Libr. J.* **2017**, *34*, 268–283. [CrossRef] [PubMed]
4. Sidana, S.; Mishra, S.; Amer-Yahia, S.; Clausel, M.; Amini, M.-R. Health Monitoring on Social Media over Time. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 17–21 July 2016; pp. 849–852.
5. Sarker, A.; Gonzalez, G. Data, tools and resources for mining social media drug chatter. In Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), Osaka, Japan, 12 December 2016; pp. 99–107.
6. Pershad, Y.; Hangge, P.; Albadawi, H.; Oklu, R. So-cial medicine: Twitter in healthcare. *J. Clin. Med.* **2018**, *7*, 121. [CrossRef] [PubMed]
7. Kale, S.; Padmadas, V. Sentiment Analysis of Tweets Using Semantic Analysis. In Proceedings of the 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 17–18 August 2017; pp. 1–3. [CrossRef]
8. Harish, B.N.; Reena, K.; Kumar, S.; Zhong, J. How much do you care? Mining and Analysis of Tweets Pertaining to Health Issues. In *SoutheastCon 2018*; IEEE: New York, NY, USA, 2018; pp. 1–8. [CrossRef]
9. Bouazizi, M.; Ohtsuki, T. A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter. *IEEE Access* **2017**, *5*, 20617–20639. [CrossRef]
10. Heart Disease: Facts, Statistics, and You. Available online: https://www.healthline.com/health/heart-disease/statisticsn#1 (accessed on 31 July 2019).
11. Sridevi, M.; ArunKumar, B.R. Role of social media in health-care domain: An integrated review. *Int. J. Eng. Res. Appl.* **2017**, *7*, 49–54.
12. Tripathi, M.; Singh, S.; Ghimire, S.; Shukla, S.; Kumar, S. Effect of social media on human health. *Virol. Immunol. J.* **2018**, *2*, 1–3.
13. Perkins, J.M.; Subramanian, S.; Christakis, N.A. Social networks and health: A systematic review of sociocentric net-work studies in low-and middle-income countries. *Soc. Sci. Med.* **2015**, *125*, 60–78. [CrossRef] [PubMed]
14. Paul, M.J.; Sarker, A.; Brownstein, J.S.; Nikfarjam, A.; Scotch, M.; Smith, K.L.; Gonzalez, G. Social media mining for public health monitoring and surveillance. In Proceedings of the Pacific Symposium on Biocomputing, Kohala Coast, HI, USA, 4–8 January 2016; pp. 468–479.
15. Sutar, S.G. Intelligent data mining technique of social media for improving health care. In Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 15–16 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1356–1360.
16. Kaushik, S.; Choudhury, A.; Mallik, K.; Moid, A.; Dutt, V.; Perner, P. Applying Data Mining to Healthcare: A Study of Social Network of Physicians and Patient Journeys. In *Computer Vision*; Perner, P., Ed.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9729, pp. 599–613.
17. Lu, Y.; Wu, Y.; Liu, J.; Li, J.W.; Zhang, P. Un-derstanding health care social media use from different stakeholder perspectives: A content analysis of an on-line health community. *J. Med. Internet Res.* **2017**, *19*, e109. [CrossRef] [PubMed]
18. Twitter Can Predict Rates of Coronary Heart Disease | Authentic Happiness. Available online: https://www.authentichappiness.sas.upenn.edu/news/twitter-can-predict-rates-coronary-heart-disease (accessed on 12 December 2020).
19. Brown, N.J.L.; Coyne, J.C. Does Twitter language reliably predict heart disease? A commentary on Eichstaedt et al. (2015a). *PeerJ* **2018**, *6*, e5656. [CrossRef] [PubMed]
20. Naili, M.; Chaibi, A.H.; Ghezala, H.H.B. Com-parative study of word embedding methods in topic segmentation. *Procedia Comput. Sci.* **2017**, *112*, 340–349. [CrossRef]
21. Document-Term Matrix. Available online: https://en.wikipedia.org/wiki/Document-term-matrix (accessed on 1 November 2020).
22. Van der Maaten, L. An introduction to dimensionality reduction using matlab. *Report* **2007**, *1201*, 62.
23. DLRL. Available online: https://dlib.vt.edu/index.html (accessed on 1 November 2020).

24. Alsinglawi, B.; Alnajjar, F.; Mubin, O.; Novoa, M.; Alorjani, M.; Karajeh, O.; Darwish, O. Predicting Length of Stay for Cardiovascular Hospitalizations in the Intensive Care Unit: Machine Learning Approach. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 5442–5445. [CrossRef]
25. Darwish, O.; Al-Fuqaha, A.; Brahim, G.B.; Jenhani, I.; Anan, M. Towards a streaming approach to the mitigation of covert timing channels. In Proceedings of the 2018 14th International Wireless Communications Mobile Computing Conference (IWCMC), Limassol, Cyprus, 25–29 June 2018; pp. 255–260.