Analysis of Tweets for Health-Related Information Using Machine Learning

Research Scholar of SKU, Chattarpur, MP, India.
Manthu.p.k@gmail.com

Dr.Monica Tripathi

Professor SKU Chattarpur, MP, India.
Monica.tripathi@skuindia.com

Abstract

Significant growth in health information sharing through Twitter is making it a compelling source for health-related information. Recent health research studies show Twitter data has been used for disease surveillance, health promotion, and sentiment analysis, and perhaps has the potential for clinical decision support. However, identifying health-related tweets in these massive Twitter datasets is challenging. With the increasing global prevalence of diabetes, user-generated health content on Twitter can be useful. Therefore, this preliminary study aims to classify diabetes-related tweets into meaningful health-related categories. Using an ensemble of neural network and stochastic gradient descent classifiers, we classified 13,667 diabetes-related tweets into five clusters. About 25.7% of the tweets were clustered as health-related, where 9.3% were classified as Treatment & Medication, 9.9% as Preventive Measures, and 6.5% as Symptoms & Causes. More than 70% were clustered as Others. Analyzing hashtags of tweets clustered in each of the categories showed significant relevance to health-related information.

Keywords: *Health information sharing, Machine learning, Decision support.*

1. INTRODUCTION

Text sentiment analysis is a method for labelling information based on certain criteria. An ML study contains different types of methods for sentiment classification. Some ML algorithms perform well on certain text datasets but do not give better results on some large datasets [1]. To overcome this problem, instead of using a single classifier, it is better to combine multiple classifiers to improve the classification performance for all the datasets. This work proposes a competitive ensemble classification algorithm as it helps to choose only the best classifier and ensemble those classifiers. This will increase the sentiment classification performance [2].

The accessibility to information via the Internet has allowed stakeholders and consumers of health information to share health-related information online. Information sharing and seeking through social media tools like Facebook and Twitter are fast-growing in significance. The term “information diffusion” has been used to describe the sharing of information amongst Twitter users [3-4]. Such activity contributes to a significant source of data on the Internet, encouraging users to engage in online health-related information searches.

As the Internet becomes accessible and affordable to consumers, healthcare professionals, caregivers, and patients are expected to use social media frequently. Searching for health-related information online provides an avenue for patients to find answers to their medical problems and help influence medical decisions. Healthcare organizations have also started

utilizing social media as part of their organizational strategy to improve patient care, expand medical education and advance medical research [5].

Therefore in this study, we will be classifying diabetes-related tweets into meaningful health-related categories that can be utilized by healthcare stakeholders when making decisions.

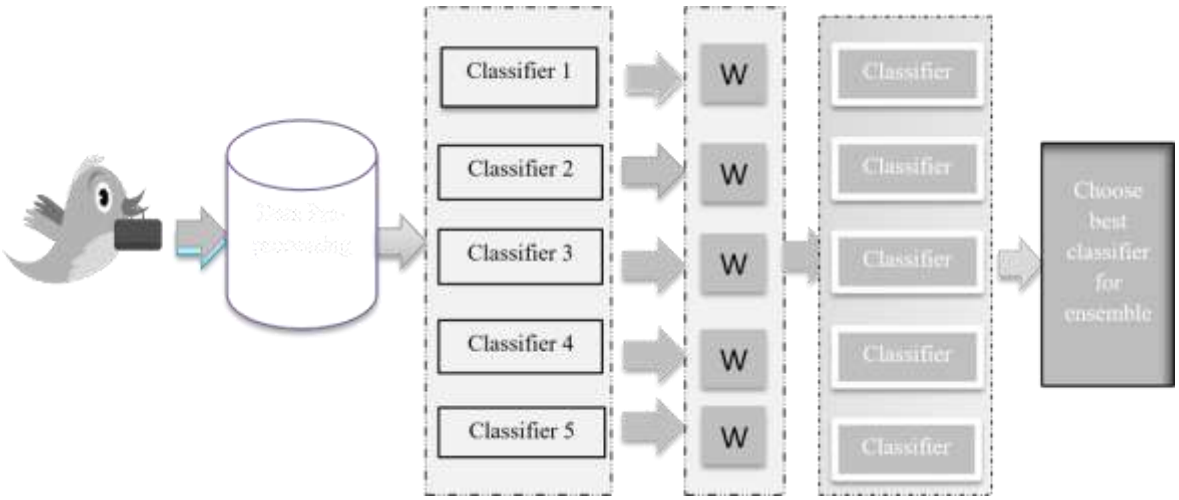

Figure 1: Competitive Ensemble Classifier for Unstructured Data

Figure 1 illustrates the flow of the proposed ensemble classifier for unstructured data. Extracted unstructured tweets are converted into structured data. Structured data is preprocessed to perform classification. Based on the accuracy, weight is assigned to each classifier and the overall weight is calculated. Considering the weight, the best classifier is chosen and it is trained with weight. The classifiers compete for one by one and produce a competitive ensemble classification. The main objective of this chapter is to identify correct sentiments in tweets and select the best classifier for the ensemble method.

2. Related Work

2.1. Health information on Twitter

Information sharing and seeking through social media is fast growing in significance. Currently, Facebook (2414 million) and Twitter (336 million) have a combined total of more than two billion active users worldwide [6]. Daily, about 500 million tweets are generated [7]. Recent studies have shown that Twitter has been useful in health research. Twitter data has been purposefully used in disease surveillance, health promotion, network conversation analysis, sentiment analysis, and medical interventions [8-11]. As a platform, Twitter is used widely for patient engagement and recruitment, and as a biomedical research collaboration tool [5, 11]. For that reason, Twitter is becoming a significant source of health-related data for public health researchers due to its real-time nature of the content and ease of access to publicly available information [11]. [5] also shared that, Twitter users with medical accounts are more likely to share “general information for public health or new research about treatments and technology”, for example, Mayo Clinic.

User-generated health content on Twitter, therefore, has the potential to contain meaningful and practical information. This can supplement healthcare decision-makers with vital patient-related details that can be utilized to improve patient-physician communication and clinical decision-making. With the global diabetes prevalence, understanding how to support and manage the growing chronic population is essential.

2.2. Machine learning topic modeling of tweets

Tweets are frequently analyzed to study the sentiments of their users. The most common approaches adopted to analyze tweets make use of machine learning algorithms [12-14]. In such approaches, machine learning techniques have been used to classify or cluster tweets according to their significance or relevance with associated tweets. [14] used an ensemble of Support Vector Machines (SVM) and Ada-boosted Decision Trees to classify user opinions into good, bad or neutral. The approach provided better accuracy than using machine learning techniques alone, resulting in an accuracy of 84%. [13] used Twitter data to gather customers' feedback regarding their sentiments on US airline services. Using seven different classifiers, namely Decision Tree, Random Forest, SVM, K-Nearest Neighbours, Logistic Regression, Naïve Bayes, and Adaboost, the authors evaluated the accuracies of the classifiers in predicting if the sentiments were positive, negative, or neutral. Using an ensemble of Adaboost and several other algorithms, the accuracy reached 84.5%.

However, other studies adopted the use of machine learning techniques beyond analyzing sentiments alone. In a research study detecting dengue outbreaks through Twitter content, the authors selected the use of supervised classification and unsupervised clustering using topic modeling. [15] trained the classifiers to achieve a prediction accuracy of 80% based on a training dataset of only 1000 tweets. Hence, classification techniques using machine learning algorithms provide an avenue where health-related tweets can be efficiently identified from a massive dataset. This allows for health-related information to be easily extracted and consumed for future decision-making.

3. Machine Learning Techniques

3.1. Nave Bayes (NB)

Multinomial Nave Bayes (NB) can be used for multi-label sentiment classification for tweets. The basic idea is to determine conditional probability among the words to classify multiclass tweets. This method works effectively while performing text classification (Japkowic & Shah, 2011) (Liu, et al., 1998). This classifier does not analyze dependence among words, but it provides a way to calculate posterior probability.

$$\text{PostProb} = \text{Likelihood} * \text{PriorProb}/\ \text{PrdProb}$$

Here, PostProb represents the posterior probability using the Bayesian theorem reversing conditional probability. The quantity of PostProb for different tweet classes is expressed as likelihood. PriorProb represents the prior probability. It is the probability of a class label. PrdProb represents the predicted probability of new tweets that can be classified. This kind of approach is called class conditional independence. Eventually, analysis suggests that it is an optimal classifier for large datasets.

3.2. LR

LR is a statistical algorithm mainly used for classification purposes. LR is generally used in conditional probability for text classification for the Twitter dataset. This classifier does not have previous knowledge; it expects a uniform distribution for the outcome occurrence. The training occurrence considers a condition that helps the classifier find Maximum Entropy.

3.3. RF

This classifier works on the concept of combining multiple decision trees on a subset of features. A decision tree is formed by breaking the dataset into small subsets to form branches with roots and leaves. It reduces the possibility of overfitting data, so it works better for unstructured data. The irrelevant trees are removed to improve the text classification performance.

3.4. K-Nearest-Neighbor

This classification technique works on the concept of identifying similarities between new occurrences and existing occurrences. The similarities of each occurrence are chosen by marking a point in dimensional space. The new occurrence is measured by the nearest point to the already existing occurrence. Meanwhile, the KNN classifies unknown occurrences based on similarity to the nearest training instances in a space. It is used for tweet sentiment classification in this thesis to analyze its performance.

3.5. SVC

This Support Vector Classifier (SVC) classifier works on the concept of separating hyperplanes. One of the main advantages of SVM is that the high-dimensional space vector problem is solved . For a given trained data, the algorithm produces an output in a hyperplane that is optimal between two classification classes. Multiclass SVM is used to classify targets into more than two labels based on a set of elements. SVM considers only instances that are very near to the boundary and discards instances that are far away from the boundary

4. Methodology

4.1. Data Collection

We used a corpus of 13,667 publicly available tweets, which were collected from 4 February to 14 February 2019. Due to the limitation set by the Twitter API, only two weeks' worth of tweets can be extracted at any point in time. All the tweets were retrieved using tweepy, a Twitter Streaming API. To search for the corpus, we used '#diabetes' as the search keyword. The collected tweet contained the full tweet details such as the full tweeted text, user information such as user details and geolocation, as well as the retweeted status.

4.2. Identifying meaningful health-related tweets

It is challenging to identify health-related tweets from massive Twitter datasets. Classifying tweets or themes by singling out tweets that may contain meaningful health-related information is essential. It is also vital that tweets that contain medical terms but are otherwise not meaningfully health-related, be classified accordingly. We identified five categories of tweets that define our understanding of what is a health and non-health-related tweet. Three categories, i.e. Preventive Measures (PM), Symptoms & Causes (SC), and Treatment & Medication (TM), are closely associated with health-related information that is routinely communicated during a patient visit. The remaining two, News (NW) and Others (OT), are used to categorize tweets that are not meaningful in our context. A semi-supervised clustering technique is then adopted to cluster all tweets into relevant categories.

4.3. Manual Classification of tweets as a labeled dataset

A semi-supervised clustering approach utilizes both labeled and unlabeled data to classify datasets. Using Python's Scikit-learn library, we randomly generated a sample of 1048 tweets (7%) from the main corpus. The process of labeling is then done manually by five coders according to the predefined coding description, as shown in Table 1. An agreement index of at least 0.6 (3 out of 5 with the same label) was agreed upon to achieve label consensus, inter-coder reliability, and a high degree of consistency across coders. The result of the manual coding is as follows:

Table 1. Classification of tweets description

Categories	Description	Tweets	%
Preventive Measures (PM)	Tweets containing measures that hinder or act as obstacles to diabetes	144	13.74%
Symptoms & Causes (SC)	Tweets that contain features, both physical and mental that indicates signs of diabetes	107	10.21%
Treatment & Medication (TM)	Tweets containing medical care rendered to patients, words associated with possible treatments (such as medication, therapy methods, etc)	145	13.74%
News (NW)	Tweets containing news or regarding news shared on diabetes	63	6.01%
Others (OT)	Tweets that cannot be classified in all other categories (e.g. motivational, awareness, complaints)	589	56.20%

4.4. Data Preprocessing

The data preprocessing phase is required to clean the tweets before any machine learning algorithms can use them. Using Python's NLTK library (a Natural Language Processing package), the tweet corpus went through the process of cleansing. The corpus was split into a 30:70 ratios and underwent the data preprocessing phase where stop words were removed. All texts were converted to lowercase and tokenized by stripping all the handles and stemming all the words. TF-IDF or term frequency inverse document frequency was then

used to calculate the statistical analysis of how a particular word or term is relevant in a given document [14].

4.5. Machine learning approach to clustering tweets

We narrowed down the following machine learning algorithms for our multi-class text classification; they are Logistic Regression (LR), Multinomial Naïve Bayes (NB), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and Backward Propagation Neural Networks (NN). The labeled data was split into training (30%) and testing datasets (70%) and used to evaluate the performance of each of the five machine-learning algorithms using a cross-validation approach. The evaluation process was done using Python's Scikit-Learn library. The results showed that the NN algorithm performed the best amongst the rest at 73.8% accuracy with SGD coming closest at 72.4% accuracy. This was followed by SVM at 71.9%, LR at 64.5%, and NB at 60.7%. The parameters for SGD and NN were then tuned further to identify the best model fitting that can result in further improved performance. A voting ensemble method was also implemented. With the tuned parameters, both SGD and NN algorithms were tested for model fitting. SGD had an accuracy of 71.7% and NN at 70.4%. While an ensemble method, which is expected to improve the accuracy further, scored a 72.7% accuracy rate. The model fit result is illustrated in Table 2. With that, the voting ensemble method was chosen to cluster the tweet corpus.

Table 2. Tuned parameters and ensemble method

Classifiers	precision	Recall	f1-score	Accuracy (%)
Stochastic Gradient Descent	*0.71*	*0.72*	*0.70*	71.7%
Backward Propagation Neural Network	*0.7*	*0.7*	*0.69*	70.4%
Voting Ensemble Method	*0.73*	*0.73*	*0.71*	72.7%

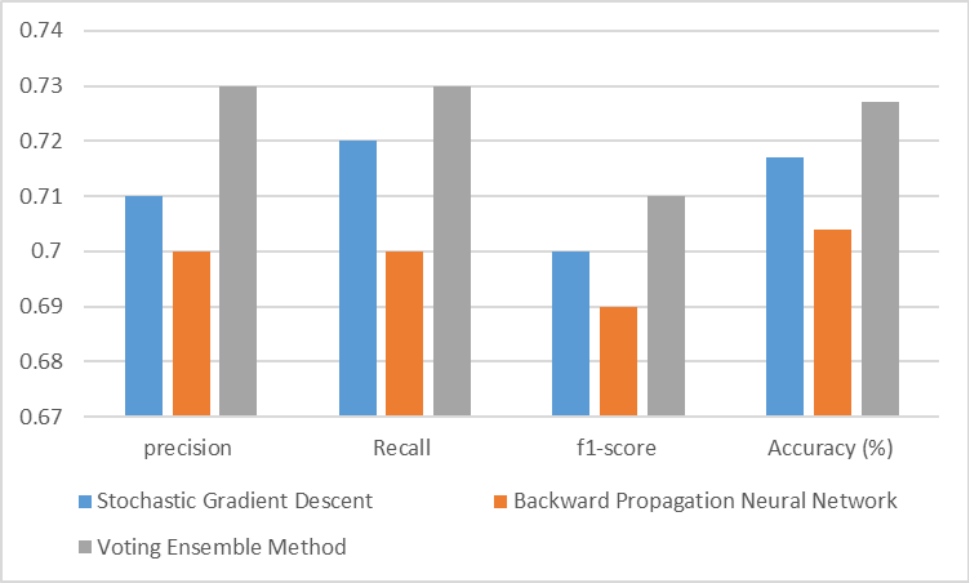

Figure 2: The Clustering Tweets Data

There were 3,512 tweets with a total of 5,802 hashtags in the health-related clusters. PM had 2,954 hashtags, SC with 1,188 hashtags, and TM with 1,660 hashtags. A word cloud helped to visualize the frequent tags used for each cluster. However, it did not rank the frequency of words used.

Figure 3: Preventive Measures, Symptoms, Causes, and Treatment clusters

A bar diagram was then used to illustrate the ranking of hashtags used in each cluster. The top 25 hashtags from each cluster were analyzed and ranked according to their frequency of use. We defined PM as tweets that contain measures that hinder or act as obstacles to diabetes. As such, hashtags like "diet, weight loss, obesity, sugar, exercise, blood sugar, fitness" seemed to suggest the actions that perhaps could help in the prevention of diabetes or reduce the impact of diabetes. Diet and food intake also appeared to be the general theme where hashtags like "nutrition, wholegrain, dietary fiber, lower-carb, alcohol" could be implied as the need to look after food consumption. "Insulin" was a frequently used hashtag for TM, which may not come as a surprise since the most common form of treatment for both type 1 and type 2 diabetes mellitus is insulin therapy. However, this may not necessarily be true as, according to [16], insulin therapy has the potential to increase the risk of cardiovascular risk and mortality in patients with type 2 diabetes. It was also interesting to notice that "estrogen" was also frequently used as a hashtag associated with diabetes. Coincidentally, there is a recent research study by [17] that investigates the protective role of estrogen and how its protective nature can be conferred to patients with diabetes mellitus. As we did not analyze the content of the tweet, we were not able to substantiate if indeed, the hashtags were about similar studies or findings [17]. From the list of top hashtags, we found four other tags, such as "stem cells", "circadian", "immunotherapy" and "afrezza" that could probably be linked to current and future studies on the treatment or medication for the management of diabetes. According to the definition by the National Institute of General Medical Science, circadian rhythms "are physical, mental and behavioral changes that follow a daily cycle, which primarily responds to light and darkness in the environment". According to the study by [18], circadian rhythms may impact sleep disturbance, which is linked to abnormal glucose metabolism and increased diabetes risk. "Afrezza" is a powdered insulin that is administered via a breath-powered inhaler for patients with diabetes that require prandial insulin (rapid-acting), which is usually in injection form.

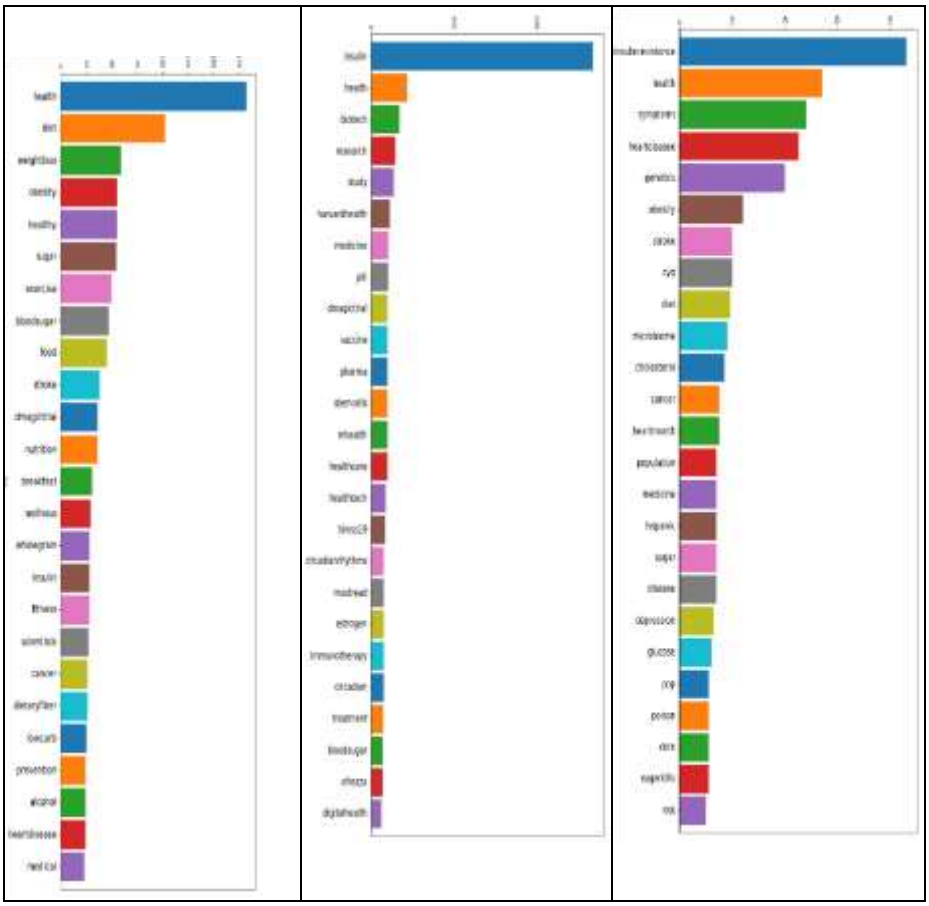

Figure 4: Frequency of hashtags in PM, TM and SC clusters

"Insulin resistance" was a highly used hashtag in SC cluster. Perhaps, the study by [16] could also explain the frequency of the word used. However, a study by [19], could better explain the link. In this study, the authors described the relationship between insulin resistance and metabolic disease that is associated with obesity, liver, pancreas, and skeletal muscle. We also found a recent study that investigates how stem cells can be used to restore insulin production and cure diabetes [20].

5. Conclusion

We implemented a machine learning approach that facilitated the classification of tweets into health-related categories which could be meaningful to decision-makers such as healthcare professionals, caregivers, and even patients themselves. From a collection of tweets related to diabetes, our approach clustered about 25% of the tweets into meaningful categories of Preventive Measures, Symptoms & Causes, and Treatment & Medication. While these clusters contain a small number of tweets, with an average number of tweets per day is 500 million tweets, the potential of curating information useful for decision support is intriguing. We believed that Twitter shows excellent promise in facilitating health information sharing, and thus, is a significant source for purposeful information. Based on our findings, a higher concentration of health-related tweets was clustered in PM and TM, which implied that Twitter users were more interested in sharing or finding out ways to prevent the onset of diabetes or treatments for the condition. Perhaps, it also had to do with diabetes mellitus being one of the four major chronic diseases prevalent globally; thus, such information is

more rampant. From the analysis of hashtags, we found significant keywords that were relevant to the three categories. In PM cluster, activities such as exercise and weight loss were evident as frequently used hashtags, suggesting actions that could be undertaken to prevent the onset of diabetes. Diet and food intake-related hashtags could suggest ways in which food consumption can have an impact on diabetes. Similarly, in clusters of SC and TM, hashtags for probable causes of diabetes mellitus, and current or latest treatments for diabetes management, recorded high recurrences.

Clinical decision-making of any sort is a complex cognitive task that requires immense intellectual competencies. While software tools help to reduce the workload, they cannot be achieved without the availability of information to support decisions. In improving the usability and interaction between humans and computer systems, our approach has identified a viable avenue where practical information can be curated for use in supporting clinical decisions. However, there are several limitations to the study. We only managed to collect a small number of tweets that spans two weeks. While our machine learning classifier was able to cluster tweets with reasonably high accuracy, more effort is required to optimize our algorithms further. As a preliminary study, we only managed to analyse the hashtags to evaluate the relevance of tweets according to the categories the tweets were clustered in.

In our future works, besides using a significant number of tweets that span over one year or more, a deep learning model is proposed to analyse the full tweet text or information and its corresponding hyperlinked contents. Also, an improved data pre-processing approach to eliminate fake and malicious content will be implemented. A natural language approach can be adopted to find meaning in tweets so that more meaningful information can be extracted for use in a decision support system.

References

[1]. Hu, F. B., Satija, A., and Manson, J. E. Curbing the Diabetes Pandemic: The Need for Global Policy Solutions. JAMA. 313(23): p. 2319-2320 (2015).

[2]. Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., and Ogurtsova, K. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice. 157: p. 107843 (2019).

[3]. Data.gov.sg. Prevalence of Hypertension, Diabetes, High Total Cholesterol, Obesity and Daily Smoking. [cited 2020; Available from: https://data.gov.sg/dataset/prevalence-ofhypertension-diabetes-high-total-cholesterol-obesity-and-dailysmoking?view_id=36a54ebf-3db6-48c8-84c8-c15e48ed5c0a&resource_id=c5f26f19-b6aa4f4f-ae5b-ee62d840f8e7 (2020).

[4]. Jung, A.-K., Mirbabaie, M., Ross, B., Stieglitz, S., Neuberger, C., and Kapidzic, S. Information Diffusion between Twitter and Online Media. (2018).

[5]. Pershad, Y., Hangge, P. T., Albadawi, H., and Oklu, R. Social Medicine: Twitter in Healthcare. Journal of Clinical Medicine. 7(6): p. 121 (2018).

[6]. Statista. Most popular social networks as of January 2020, ranked by number of active users. Available from: https://www.statista.com/statistics/272014/global-social-networks-rankedby-number-of-users/ (2020).

[7]. Mention. Twitter Engagement Report 2018. 2018.

[8]. Finfgeld-Connett, D. Twitter and Health Science Research. Western Journal of Nursing Research. 37(10): p. 1269-1283 (2015).

[9]. Gabarron, E., Dorronzoro, E., Rivera-Romero, O., and Wynn, R. Diabetes on Twitter: A Sentiment Analysis. Journal of Diabetes Science and Technology. 13(3): p. 439-444 (2019). 10. Sedrak, M. S., Salgia, M. M., Bergerot, C. D., Ashing-Giwa, K., Cotta, B. N., Adashek, J. J., Dizman, N., Wong, A. R., Pal, S. K., and Bergerot, P. G. Examining Public Communication About Kidney Cancer on Twitter. JCO Clinical Cancer Informatics.(3): p. 1-6 (2019).

[10]. Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., and Merchant, R. M. Twitter as a Tool for Health Research: A Systematic Review. American Journal of Public Health. 107(1): p. e1-e8 (2017).

[11]. Joyce, B. and Deng, J. Sentiment analysis of tweets for the 2016 US presidential election. in 2017 IEEE MIT Undergraduate Research Technology Conference (URTC). (2017).

[12]. Rane, A. and Kumar, A. Sentiment Classification System of Twitter Data for US Airline Service Analysis. in 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). (2018).

[13]. Rathi, M., Malik, A., Varshney, D., Sharma, R., and Mendiratta, S. Sentiment Analysis of Tweets Using Machine Learning Approach. in 2018 Eleventh International Conference on Contemporary Computing (IC3). (2018).

[14]. Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., Cedrim, D., and da Silva Sousa, L. Tracking Dengue Epidemics Using Twitter Content Classification and Topic Modelling. (2016). Cham: Springer International Publishing.

[15]. Herman, M. E., O'Keefe, J. H., Bell, D. S. H., and Schwartz, S. S. Insulin Therapy Increases Cardiovascular Risk in Type 2 Diabetes. Progress in Cardiovascular Diseases. 60(3): p. 422434 (2017).

[16]. De Paoli, M. and Werstuck, G. H. Role of Estrogen in Type 1 and Type 2 Diabetes Mellitus: A Review of Clinical and Preclinical Data. Canadian Journal of Diabetes. (2020).

[17]. Reutrakul, S. and Van Cauter, E. Interactions between sleep, circadian function, and glucose metabolism: implications for risk and severity of diabetes. Annals of the New York Academy of Sciences. 1311(1): p. 151-173 (2014).

[18]. Czech, M. P. Insulin action and resistance in obesity and type 2 diabetes. Nature Medicine. 23(7): p. 804-814 (2017).

[19]. Sordi, V., Pellegrini, S., Krampera, M., Marchetti, P., Pessina, A., Ciardelli, G., Fadini, G., Pintus, C., Pantè, G., and Piemonti, L. Stem cells to restore insulin production and cure diabetes. Nutrition, Metabolism and Cardiovascular Diseases. 27(7): p. 583-600 (2017).